Theoretic Foundation of Predictive Data Analytics

By Jun Huan

July, 2015

Contents

To my family

Preface

There are many excellent books on data science, statistical learning, machine learning, and data mining. Examples include Pattern Recognition and Machine Learning by Bishop, Elements of Statistical Learning, Data Mining: Concepts and Techniques by Hastie *et al.* among others.

Why do I want to write another book? I want to share my own experience of working on data science and data analytics. When I studied machine learning, data mining, and statistical learning as a graduate student at UNC-Chapel Hill many years ago, I had many many questions. Vast majority of these questions were answered by those great textbooks while some of them not. Later I started teaching data science, machine learning, and data mining myself. In order to provide a clear references I started to trace back from which book that my questions were answered. I noticed that each book answered only a subset of my questions. So I started to develop my notes to document my progresses on understanding the rich body of which I now call theoretic foundation of predictive data analytics where we aim to develop a model to predict future events. I started to share the notes with my students and colleagues and asked for feedbacks. Eventually I decide to publish my notes and share them with a large body of audience.

So far I have outline the primary motivation of this book as to present the theory of predictive data analytics in a unified way. A few sample questions that I personally am very interested to discuss are listed below; just to give the readers more insights regarding the organization of the book. For example in learning support vector machines, we get a message that agnostic learning (which basically means learning without knowing the distribution of the data) is important for real-world applications. One immediate question is why we still use maximum likelihood estimator (where we use a likelihood function to describe how data are generated) or Bayesian estimator (where we use a likelihood function AND a prior distribution to describe the joint distribution of data and model parameters). When we switch to the Bayesian treatment of predictive analytics, we may adopt the subjective interpretation of probability. With that interpretation should we start to question whether non-Bayesian approaches are "irrational"? Please note by no means I want to settle those questions (since many of them are still debatable). It is my intention to present the related materials well so that the reader of the text could make an informed decision.

The second motivation of this book is primarily from pedagogy on how we balance mathematic rigorousness and intuitive description. Of central to the discussion is the concept of random variable. I tried to develop the book based on the intuitive description of random variable as a function mapping from a probability space to real numbers without emphasizing that such functions should be "measurable". All the materials are arranged in such way that the readers can capture most of the discussions in this book without knowing the concept of measure and measurable space. However a number of "pointer" are offered to guide interested readers to perform in-depth study if interested. A separate chapter on the strong law of large numbers is provided at the end of the book to demonstrate what we could obtain if we utilize the advanced concepts.

The third motivation of this book is the intensive discussion of big data at present time. We have a number of questions in dealing with big data. Does large amount of independent and identically

distributed data make statistics easier with law of large number? Does statistical learning theory provide guidelines regarding the number of samples that you need to collect in order to obtain a classifier that is close to the optimal one? To better plan the future for big data it is postulated by the author of this book that we should revisit the theoretic principles that are behind many successful data analytic tools and see how those algorithms are connected.

Last but not least the book is primarily for Computer Science and Computer Engineering graduate students and upper-level undergraduate students who are interested in data science and want to gain a reasonably deep insight of the subject. In Computer Science training, students have learn a great deal of computational thinking. I believe we should help them to start to think about uncertainty, inference, and decision making, using the language such as consistent, loss function, risk, which usually require a rather different set of skills and thinking. I hope this book will be helpful to those students.

This book started as the class notes that I used in teaching two courses at the University of Kansas. These two courses are Machine Learning (EECS 738) and Theoretic Foundation of Data Analytics (EECS 940). I taught these two courses in 2011 - 2016. I want to thank all the students and the TAs of the class. Those people gave me valuable feedbacks regarding the notes. In 2016 I took a position at National Science Foundation as a Program Director in the Information and Intelligence Division of the Computer and Information Science and Engineering Directorate. Thanks to the generous support of NSF as enabled by the Individual Research Development program, I could spent non-trivial effort in continuing my own research, including finishing this book.

Part I

Basics

Chapter 1

Introduction

1.1 A Big Picture of Predictive Data Analytics

With the fast development of data collection, data transition, and data storage techniques researchers and practitioners often
Let's consider a few examples. Computer vision Speech recognition Bio informatics

1.2 Data, Model, and Prediction

It is hard to define what is data. Here we adopt a rather narrow view and we define data as any digitalized description of the world. Here the key point is "digitalized". Even with this definition we find data come in with vast different forms. For example we may have images, audios, free text such as twitter messages, pdf files, records in relational databases. We call these data as "raw data".
To provide a general framework of data analytics, for the time being we modify our working definition of data. Unless specified otherwise, data in this book is organized in a matrix format $\mathbf{X} = (x_{i,j})_{i=1,j=1}^{n,p}$. Rows in the matrix are samples and columns in the matrix are features. The entry at the ith row and jth column is the measure of the jth feature at the sample i.

1.3 Uncertainty, Consistency, Loss, and Risk

Consistency is a concept that goes back to Fisher.
Is consistency sufficient or desired? First in reality we seldom have sufficient large number of samples. Second if we have limited number of samples, our goal ought to be finding the best way to perform data analysis rather than finding an approach that works best if we have ample samples.
Is consistency necessary? Different people have different opinions.
Loss is the function that we use to measure the difference between predicted value and the true value.
Risk is the expected loss. Using probability,

1.4 Overfitting and Its Prevention

In the construction of predictive models we often encounter an interesting phenomenon of "overfitting" where we have high quality model after training. However when we apply the model to test

data, the quality of the model drop significantly. For example we may have perfect separation in the training data but the prediction on the test data may be close to random prediction.
How do we deal with overfitting?
I believe that the uniform laws of large numbers provide a common ground for studying. The theorem is needed to prove the consistency of commonly used estimation methods such as maximum likelihood estimation. The same theorem is also well connected to VC-dimension and Support Vector Machines. With this reason we plan to spend fair amount of time to explain the theorem and its connections to different chapters of this book.

1.5 Bayesian vs. Frequentist Interpretation

An almost unavoidable question in writing a theory oriented book regarding data analytics is the discussion of the difference of frequentist's view and the Bayesian statistician's view of data analytics. From pedagogy's perspective I feel it is hard to explain the Baye's theorem to students from the beginning. We often run into questions regarding topics such as how we select prior distribution. What is the interpretation of probability in the Bayesian setting? For most students I feel the concept of long term frequency, or physical probability, is well explained. It is difficult to present alternative views, for example evidential probability, at the beginning. So I decide to start with the classical maximum likelihood estimation. I then present penalized maximum likelihood estimation with the discussion centered around model selection. I then discuss agnostic learning, using support vector machines as an example, to show a view point that learning can be distribution free. Finally I present the Bayesian learning, centerer around its subjective interpretation, and talk about connections of Bayesian learning to the previous discussed topics.

1.6 Methematical Treatment of Random Variable

Random variables and their asymptotic behavior are of critical role in the theoretical discussion of data analytics. The mathematic description of random variable requires rather sophisticated treatment such as measure and σ-algebra. Such mathematica rigorousness is absolutely necessary if we want to have a fair understanding of useful concepts such as the strong law of large number. However I believe that requiring such knowledge up front is intimating and unfair to many students. It is my goal to not introduce a concept until it is absolutely necessary. I set a test to myself to see how much that I can accomplish in presenting the materials without requiring mathematical tools other than those commonly covered in college level analysis and linear algebra courses.

1.7 Model Interpretation vs. Model Evaluation

There is a perspective called the "operational view" of data analytics. In this view, the data analytics is described as a on-line machinary where data are coming one by one (the exact details are not relevant now). We use the data to update our model and to continuously evaluate the model. The objective of the research is not to come up with the "true" model (since people are arguing that there is NO such notation of true model). The objective is to get the prediction as good as possible. If there is an argument, the only way to dispute the argument is for each side of the argument to put down their predictions and using time and the continuously arriving data to tell who is more "accurate". Following this view, there is not much room for model interpretation. Arguably in industry applications people adopt the operational view quite frequently. This perspective does lead to a natural following up question regarding what is the metric(s) to evaluate the model. In

this book we will talk about a closely related concepts called "loss function" in learning. A loss function measures quantitatively the difference between a target value and your estimation of the value. Usually the larger the difference is, the higher the "loss" is. Loss function's primary role is in deriving machine learning theory and algorithms. Please notice that in common practice with a few exceptions loss functions are not directly used to evaluate models.

Not everyone is happy with the operational view. Scientists often complain that many predictive data analytics algorithms are blackbox. For example in bio sciences, if we predict that a gene is associated with a disease, biologists usually will not carry the results away happily and go from there. They usually ask how do you reach the conclusion? What are the features that you used? And so on so forth. Based on my limited interdisciplinary research experience (e.g. about 20 years), the discussion could quickly goes to the underlying biological machinary (or a social science theory or a health problem etc) where data scientists can get lost frequently. Our approach towards model interpretation in this book is quite different. We primarily centered around "asymptotic" behavior of statistical learning. In the discussion we set up a thought experiments, imaging that we have very large number (or even infinite number) of samples.

Clearly we omitted several important topics. Examples include scalability of predictive data analytics, security and privacy of predictive data analytics, social impacts of predictive data analytics, and causality inference. Covering everything delays the delivery of the book infinitely. Fortunately there are many excellent books (or papers) on those topics. We refer to the bibliographic note sections of this book for interested readers.

I believe that to better understand the content of a book, a few exercises are healthy practice. I tried to add a few exercises at the end of each chapter. People can find the answers to most of the exercises easily with the help of on-line search engines and hence I do not provide answering keys to those problems.

Chapter 2

Probability Theory and Laws of Large Numbers

Probability plays a central role in data science. Below we present a classical view of probability to start our conversation. A more rigorous definition, based on σ algebra and measure theory is provided in a later chapter in Appendix D.

2.1 Axioms of Probability

Definition 2.1 (Probability)**.** Given a set Ω and a set $\mathcal{A}$ of subsets of Ω containing Ω and the empty set $\emptyset$, $\mathbb{P} : \mathcal{A} \to \mathbb{R}$ is a *probability function* if it satisfies the following three axioms:

Axiom 1: $\mathbb{P}(A) \geq 0$ for all $A \subseteq \mathcal{A}$

Axiom 2: $\mathbb{P}(\Omega) = 1$

Axiom 3: If $A_1, A_2, \ldots$ are pairwise disjoint,

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

We call Ω the *sample space*, each element $\omega \in \Omega$ an *outcome* in the sample space, and $\mathcal{A}$ the set of *events*. The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is a *probability space*. In other words, a probability space is composed of a sample space, a set of events, and a probability function defined with the sample space and the set of events.

Example 2.1. With the axioms we have

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) \text{ for all } A \in \mathcal{A}, \tag{2.1.1}$$

$$\mathbb{P}(\emptyset) = 0. \tag{2.1.2}$$

$A^c = \Omega - A$ is the set complement of A. (2.1.1) is a consequence of the probability Axiom 3. (2.1.2) follows (2.1.1) and the probability Axiom 2.

Making the association of probability to real world applications is not an easy task. In literature we see that the interpretation of probability can be grouped in (at least) the following two categories.

- Physical probability (e.g. the long-term behavior of an object)

- Evidential probability (e.g. belief, or the subjective judgement of how likely an event may happen)

Before we start to explore different interpretations, we use a few examples to refresh our concept of probability that is natural to either interpretations.

Example 2.2. Let us start with a trivial case. Let's consider a situation where we have a group of N balls. All the balls are identical except that they are numbered as $1, 2, \ldots, N$. If we randomly pick up one, what is the probability that we pick a ball that is number as "1"? The answer $1/N$ is very straightforward. Since all the balls are identical, we believe that they have the same chance of being picked up. Let us denote the action that we pick up the ball numbered as i as an event ($1 \leq i \leq N$). Clearly the sum of the probabilities for all the events should be 1. Putting all together we obtain the answer $1/N$.

Through this simple example we want to emphasize two points. First, the probability calculation has a belief component where we assume all balls are identical AND labels do not have any influences on how we pick up a ball. Second, our assumption or "belief" could easily be verified by simply repeating an experiment again and again and recording the outcome of the experiments. In the later part of this chapter we show a case (central limit theorem) where we use mathematics to replace running experiments repetitively and show that with minimal belief we are able to assign probability to evens. However we also want to point out that (at least theoretically at this moment) the coupling of our straightforward assumption and the availability of easy-running experiments is leisure. We may well run into situations where it is hard, if not impossible, to validate our belief by physically running an experiment repetitively.
Continuing the discussion we use a few more examples to refresh our knowledge of combinatorics. We will see the same example again when we introduce evidential probability.

Example 2.3. Given a group of N balls, $m \leq N$ are identical green balls, the rest are identical red balls. Green balls and red balls are otherwise identical despite the difference in their colors. If we randomly pick up one ball from the group without replacement, what is the probability that we pick up a green ball?
Given that all balls have equal chance to be picked up, straightforward calculation shows the probability of picking up a green ball is

$$\frac{m}{N} \quad \text{or} \quad \frac{m(N-1)!}{N!}.$$

The first solution is quite self-explaining. The second solution however looks quite tortuous at the first glance. We shall see its utility soon (in Chapter 16 for example). The solution is derived by a simple counting as following. Let us imagine that we label the balls as $1, 2, \ldots, N$. For N balls there are a total of $N!$ permutations. Among those permutations let us count how many of the permutations have a green ball as the first one. For the first green ball we have a total of $\binom{m}{1} = m$ choices. Once the first one is fixed for the rest $N-1$ balls we have a total of $(N-1)!$ permutations. Since all balls are identical except their colors, all the permutations should have equal chance to occur and the probability of picking up a green ball is hence $\frac{m(N-1)!}{N!}$.

Example 2.4. With the same set up as (2.3) what is the probability that the first two randomly selected balls (without replacement) are both green? The answer is

$$\frac{m}{N}\frac{m-1}{N-1} \quad \text{or} \quad \frac{\binom{m}{2}2!(N-2)!}{N!}.$$

The first solution is again self-explaining. For the second solution we again count the permutations that have two green balls as the first two. To do so for the first two green balls we have a total of $\binom{m}{2}$ choices. These two balls have 2! permutations. Once the first two are fixed for the rest $N-2$ balls we have a total of $(N-2)!$ permutations. Combining the information and following the same calculation we did in (2.3) we have the results.

Example 2.5. With the same set up as (2.3) what is the probability that in the first two randomly selected balls, one is green, one is red? The answer is

$$2\frac{m}{N}\frac{N-m}{N-1} \quad \text{or} \quad \frac{\binom{m}{1}\binom{N-m}{1}2!(N-2)!}{N!}.$$

In counting those permutations that have one green ball and one red ball as the first two balls, we notice that for the green ball we have $\binom{m}{1}$ choices. For the red ball we have $\binom{N-m}{1}$ choices. We then obtain the result.

Example 2.6. With the same set up as (2.3) we solve a general problem of calculation the probability that in the first $n \leq N$ randomly selected balls (without replacement) there are exactly l green balls ($l \leq \min(n, m)$). The answer to the problem is

$$\frac{\binom{m}{l}\binom{N-m}{n-l}n!(N-n)!}{N!}.$$

2.2 Random Variables

Definition 2.2 (Random variables)**.** Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, a *random variable* X is a function that maps the sample space Ω to real numbers $\mathbb{R}$ [1].

Definition 2.3 (Cumulative Distribution Function)**.** Given a random variable X, the *cumulative distribution function* $F(X) : \mathbb{R} \to [0, 1]$ of X is the probability that X takes the value that is no more than x, or

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(\{\omega | X(\omega) \leq x\}). \end{aligned} \tag{2.2.1}$$

For the two notations $\mathbb{P}(X \leq x)$, $\mathbb{P}(\{\omega | X(\omega) \leq x\})$, the first one is simpler while the second one is more informative. By using the set notation we explicitly point out that the random variable X is a function and the probability is measured for a subset of the domain of X. A random variable is *continuous* if its range is infinite and uncountable and otherwise it is *discrete*.

Definition 2.4 (Probability Density Function)**.** Given a random variable X, the *probability density function* $f_X(x)$ of X, or its PDF if such a function exists, is the derivative of the cumulative distribution function:

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{2.2.2}$$

Definition 2.5 (Expectation)**.** The *expectation* $\mathbb{E}[X]$ of the random variable X is the weighted mean of X,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, dF_X(x). \tag{2.2.3}$$

[1] Any random variable must be measurable. See Appendix (D) for further discussions.

The *variance* $\mathbb{V}[X]$ of the random variable X is the expectation of the random variable $(X-\mathbb{E}[X])^2$ or $\mathbb{V}[X] = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]$.
Given a random variable X, if the expectation $\mathbb{E}[X]$ exists we have

$$\int_{-\infty}^{\infty} x\, dF_X(x) = \int_{\Omega} X(\omega)\, d\mathbb{P}. \tag{2.2.4}$$

The left part of (2.2.4) is what we are familiar with but it is quite unusual in terms of the integral definition. The idea is that we perform integration on the co-domain of the sample space Ω. In this approach we avoid the discussion of what is the sample space Ω. It can be a finite set (for discrete random variables), a finite-dimensional Euclidian space, or more exotic ones such as a set of functions (as we will experience in Bayesian nonparametrics). This type of integration is called Lebesgue integration, which may not be familiar to some readers.
Alternatively we could define expectation directly on the sample space as we did on the right part of (2.2.4). This Riemann-Stieltjesin integral approach is straightforward in theory and offers intuitive connection to the statement that an expectation of a random variable is the "weighted mean" of the function. We want to pint out that (2.2.4) needs to be proved but we do not do so here. We would rather state a general statement that for a function if its Lebesgue and Riemann-Stieltjes integrals both exist, they must agree.
For a continuous random variable where its PDF exists, we have

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x\, dF_X(x) \\ &= \int_{-\infty}^{\infty} x f_X(x)\, dx. \end{aligned} \tag{2.2.5}$$

For a discrete random variable the integration degenerates to a summation

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x\, dF_X(x) \\ &= \sum x\mathbb{P}(X=x). \end{aligned} \tag{2.2.6}$$

We have seen a few examples of discrete random variables in Example 2.3 to 2.6. Normal (or Gaussian) random variables are commonly used continuous random variables. The central limit theorem (which we will introduce shortly) states that if we have many random variables, the mean of these random variable will be approximately normal, regardless of the original distributions the random variables take. To better understand important results such as the central limit theorem and alike, we provide a formal definition of normal random variable below.

Definition 2.6 (Normal Random Variable)**.** A continuous random variable Z is said to be a *normal random variable* (or a Gaussian random variable), shown as $Z \sim \mathcal{N}(\mu, \sigma)$, if its PDF takes the form of

$$f_Z(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}). \tag{2.2.7}$$

A normal random variable Z is in its standard form if $\mu = 0$ and $\sigma = 1$, or $Z \sim \mathcal{N}(0,1)$.
As an example we briefly show how we calculate expectation and variance of a normal random variable.

To calculate the expectation of a normal random variable we have

$$\begin{aligned}\mathbb{E}[Z] &= \int_{-\infty}^{\infty} x f_Z(x)\, dx \\ &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})\, dx.\end{aligned}$$

Clearly the PDF function of Z is symmetric with the line $x = \mu$ and hence $\mathbb{E}[Z] = \mu$.
To calculate the variance of a normal random variable we have

$$\begin{aligned}\mathbb{V}[Z] &= \int_{-\infty}^{\infty} (x-\mu)^2\ dF_Z(x) \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})\, dx\end{aligned}$$

Let $t = (x-\mu)/\sigma$ and we have

$$\begin{aligned}\mathbb{V}[Z] &= \int_{-\infty}^{\infty} t^2\sigma^2 \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{t^2}{2})\, d(t\sigma) \\ &= \sigma^2 \int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{t^2}{2})\, dt.\end{aligned}$$

Straightforward calculation shows that $\int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{t^2}{2\sigma^2})\, dt$ is one and hence we have $\mathbb{V}[Z] = \sigma^2$.
In dealing with different distributions we may run into cases where different parameters may lead to the same distribution (or the same probability density function). If such cases happen, we say that the distribution is non-identifiable. A distribution is *identifiable* with a parameter θ_0 if there does not exists a $\theta \neq \theta_0$ such as $f(x;\theta) = f(x;\theta_0)$ for all x. For discrete variables we use probability mass function rather than PDF. See an example of non-identifiable distributions in Exercise 2.2.

2.3 Transformation of Random Variables

In this section we consider transformations of random variables. Such transformations are useful in a number of places include deriving new distributions (e.g. χ-squared distributions).
Consider a (measurable) function $g(:)$ that maps real numbers to real numbers. Given a random variable X, $Y = g(X)$ is another random variable defined as $Y = g \circ X$. $\circ$ is the composition of two functions. The CDF function of $F_Y(:)$ is easily calculated as $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(\{\omega | g(X(\omega)) \leq y\})$.
Let $g^{-1}\mathcal{A} = \{x | gx \in \mathcal{A}$ be the pre-image of a set $\mathcal{A}$. It is straightforward to show that
In the following discussion we assume the transformation function g is one-to-one and is continuous at a point x. The requirement looks quite restrictive at the beginning but it is actually quite flexible as demonstrated in a few exercises. With the one-to-one requirement, g^{-1} is defined for every singletons
We want to establish a straightforward but very general and useful relationship between the PDF of X and its transformation $Y = g(X)$.

$$\begin{aligned}f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|\end{aligned} \tag{2.3.1}$$

Example 2.7. Given a standard Gaussian variable $X \sim \mathcal{N}(0,1)$, let $g(x) = \sigma x + \mu$ be a linear transformation where $\sigma > 0$ and μ are two constants. Clearly $g^{-1}(y) = (y-\mu)/\sigma$ and $\left|\frac{dg^{-1}(y)}{dy}\right| = \left|\frac{d(y-\mu)/\sigma}{dy}\right| = \frac{1}{\sigma}$. Now we have $Y = g(X)$ as a random variable obtained by transforming the standard Gaussian X with g. By (2.3.1) we have $f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| = \frac{1}{\sqrt{2\pi}}\exp(-\frac{(y-\mu)^2}{2\sigma^2})\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\mu)^2}{2\sigma^2})$.

2.4 Inequality of Random Variables

Inequalities are widely used to study random variables. Below we introduce a few simple examples and we show how to use them to obtain very useful results in the next section.

Theorem 2.1 (Markov's Inequality)**.** *For a non-negative random variable $X \geq 0$ we have*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}. \tag{2.4.1}$$

Proof. For any $t > 0$,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x \; dF_X(x) \\
&= \int_0^t x \; dF_X(x) + \int_t^\infty x \; dF_X(x) \\
&\geq t \int_t^\infty dF_X(x) \\
&= t \cdot (1 - F_X(t)) \\
&= t \cdot \mathbb{P}(X > t).
\end{aligned}$$

□

Theorem 2.2 (Johnson's Inequality)**.** *For a random variable X and any $t > 0$ we have*

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\mathbb{V}[X]}{t^2}. \tag{2.4.2}$$

Proof. Let $Y = (X-\mu)^2$. Notice that $Y \geq 0$, $\mathbb{E}[Y] = \mathbb{V}[X]$, and $\mathbb{P}(Y > t^2) = \mathbb{P}(|X-\mu| > t)$. Applying Markov's Inequality (2.1) we have Johnson's Inequality. □

2.5 Laws of Large Numbers

Law of large numbers played a central role in the development of statistics. In discussing law of large numbers here we want to focus on two questions: (1) what we may obtain from data and (2) how do we interpret probability using law of large numbers. The discussion is very important in the context of large-scale data analytics and big data.

Although random variables is a familiar concept that is usually covered in an introductory probability courses, the full treatment of the concept here posts significant challenges. Part of the challenge is that we plan to discuss the convergence of random variables. If we adopt the view that a random variable is a function (it is!) we may bring in a rich body of literature regarding function

convergence. Related topics include convergence point-wise, uniform convergence, convergence in measure, and convergence almost surely. We do not plan to do so at this moment.
Fortunately in the context of data science, typically we are dealing with a rather special situation where a random variable converges to a constant (or a *constant function* that assigns a constant, a.k.a. the true parameter, to any data). With this special situation we believe we could present a big picture without requiring too much in-depth mathematical knowledge. We touch the strong law of large number and give details that require some understanding of measure theory in the Appendix for interested readers.

Definition 2.7 (Converges in Probability)**.** A sequence of random variables X_n converges in probability to a constant μ if for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mu| > \epsilon) = 0. \tag{2.5.1}$$

We use the notation $X_n \xrightarrow{\text{P}} \mu$ to denote that the sequence of random variable X_n converges in probability to μ.

Theorem 2.3 (Weak Law of Large Number)**.** *Given n i.i.d. random variables $X_1, X_2, \ldots, X_n$ where we have $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \sigma^2$, μ and σ are finite, $1 \le i \le n$. Let $\overline{X}_n = \frac{1}{n}\sum_i X_i$ and for any $\epsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}(\left|\overline{X}_n - \mu\right| > \epsilon) = 0 \text{ or } \overline{X}_n \xrightarrow{\text{P}} \mu.$$

Proof. Let us first compute the mean and variance of sample mean. We have

$$\begin{aligned}
&\mathbb{E}\left[\overline{X}_n\right] = \mathbb{E}\left[\frac{\sum_i x_i}{n}\right] = \mu, \\
&\mathbb{E}\left[(\overline{X}_n - \mu)^2\right] = \mathbb{E}\left[(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right])^2\right] = \mathbb{V}\left[\overline{X}_n\right], \\
&\mathbb{V}\left[\overline{X}_n\right] = \mathbb{V}\left[\frac{\sum_i x_i}{n} - \mu\right] = \mathbb{V}\left[\frac{\sum_i (x_i - \mu)}{n}\right] = \frac{n\cdot\sigma^2}{n^2} = \frac{\sigma^2}{n}.
\end{aligned}$$

With these quantities calculated we have

$$\begin{aligned}
\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) &= \mathbb{P}((\overline{X}_n - \mu)^2 > \epsilon^2) \\
&\le \frac{\mathbb{E}\left[(\overline{X}_n - \mu)^2\right]}{\epsilon^2} \quad \text{(By Markov's Inequality Theorem (2.1))} \\
&= \frac{\mathbb{V}\left[\overline{X}_n\right]}{\epsilon^2} \\
&= \frac{\sigma^2}{n\epsilon^2}.
\end{aligned} \tag{2.5.2}$$

Once we bound $\mathbb{P}(|\overline{X}_n - \mu| > \epsilon)$ by $\frac{\sigma^2}{n\epsilon^2}$, we let $n \to \infty$, take limit on both side, and obtain the conclusion. □

Definition 2.8 (Converges Almost Surely)**.** A sequence of random variables X_n converges almost surely to a constant μ if

$$\mathbb{P}(\lim_{n\to\infty} X_n \neq \mu) = 0.$$

Convergence almost surely (a.s.) looks like a natural extension of convergence in probability but it is actually much harder to prove. We do not intend to provide a proof here but we just want to use the following two examples to show that such convergence could be quite hard to obtain.

Example 2.8. With n *i.i.d.* Bernoulli random variables from a distribution indexed by the parameter θ, and assume θ_0 is an irrational number. We have $\mathbb{P}(\overline{X}_n = \theta_0) = 0$ for any n.

Example 2.9. With n *i.i.d.* Bernoulli random variables from a distribution with the parameter $\theta_0 = \frac{1}{2}$. We have $\mathbb{P}(\overline{X}_n = \theta) = 0$ for any odd n.

2.6 Moment Generating Function and Converges in Distribution

Definition 2.9 (Converges in Distribution)**.** A sequence of random variables X_n converges in distribution to a random variable X if the CDF of X_n converges to that of X point-wise or

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

for any x in $\mathbb{R}$ where $F_X(x)$ is continuous.

We use the notation $X_n \xrightarrow{\text{d}} X$ for convergence in distribution.
We show that the distribution of the sample mean $\overline{X}_n$ converges in distribution to a Gaussian distribution:

$$\overline{X}_n \xrightarrow{\text{d}} \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}).$$

To prove this we introduce an important function of random variables: the moment generating function.

Definition 2.10 (Moment Generating Function)**.** For a random variable X, its moment generating function m is

$$m_X(t) = \mathbb{E}\left[\exp(tX)\right].$$

Example 2.10. For standard Gaussian $X \sim \mathcal{N}(0, 1)$,

$$\begin{aligned} m_X(t) &= \mathbb{E}\left[\exp(tX)\right] \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot e^{tx}\, dx \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} \cdot e^{\frac{t^2}{2}}\, dx \\ &= e^{\frac{t^2}{2}} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}}\, dx \\ &= e^{\frac{t^2}{2}}. \end{aligned}$$

With standard Gaussian we see that $m'_X(0) = \frac{d\exp(\frac{t^2}{2})}{dt}|_{t=0} = e^{\frac{t^2}{2}} \cdot t|_{t=0} = 0 = \mathbb{E}[X]$, and $m''_X(0) = \frac{d^2\exp(\frac{t^2}{2})}{dt^2}|_{t=0} = e^{\frac{t^2}{2}} + e^{\frac{t^2}{2}} \cdot t|_{t=0} = 1 = \mathbb{E}\left[X^2\right]$. The following Theorem shows that the relationship between the expectation and derivatives of the moment generating function extends beyond standard Gaussian and holds for all random variables.

Theorem 2.4. *For any random variable X we have:*

$$\begin{aligned} &m_X(0) = 1, \\ &m'_X(0) = \mathbb{E}[X], \\ &m''_X(0) = \mathbb{E}\left[X^2\right]. \end{aligned} \tag{2.6.1}$$

Proof. We show how these conclusions are obtained in (2.6.1). Plug in the definition of moment generating function $m_X(t) = \mathbb{E}[\exp(tX)]$ we have

$$\begin{aligned} m_X(0) &= \int \exp(0x) f_X(x)\, dx \\ &= \int f_X(x)\, dx \\ &= 1. \end{aligned}$$

Similarly we have

$$\begin{aligned} m'_X(0) &= \frac{d\int \exp(tx) f_X(x)\, dx}{dt}|_{t=0} \\ &= \int \frac{d\exp(tx)}{dt} f_X(x)\, dx|_{t=0} \\ &= \int x f_X(x)\, dx \\ &= \mathbb{E}[X]. \end{aligned}$$

We also have

$$\begin{aligned} m''_X(0) &= \frac{d^2\int \exp(tx) f_X(x)\, dx}{dt^2}|_{t=0} \\ &= \int \frac{d^2\exp(tx)}{dt^2} f_X(x)\, dx|_{t=0} \\ &= \int x^2 f_X(x)\, dx \\ &= \mathbb{E}\left[X^2\right]. \end{aligned}$$

□

Theorem 2.5. *Let X, Y are two independent random variables, $Z_1 = X + Y$, $Z_2 = c \cdot X + \mu$ where c and μ are two constants. We have*

$$\begin{aligned} &m_{Z_1} = m_X \cdot m_Y \\ &m_{Z_2} = e^{t\mu} \cdot m_X(ct). \end{aligned} \tag{2.6.2}$$

Proof.

$$\begin{aligned}
m_{Z_1}(t) &= \mathbb{E}\left[\exp(tZ_1)\right] \\
&= \mathbb{E}\left[\exp(t(X+Y))\right] \\
&= \int \exp(t\cdot(x+y)) f_X(x) f_Y(y)\, dxdy \\
&= m_X(t)\cdot m_Y(t). \\
m_{Z_2}(t) &= \mathbb{E}\left[\exp(tZ_2)\right] \\
&= \mathbb{E}\left[\exp(t(cX+\mu))\right] \\
&= \int \exp(t\cdot(cx+\mu)) f_X(x)\, dx \\
&= e^{t\mu}\int \exp((tc)x) f_X(x)\, dx \\
&= e^{t\mu}\cdot m_X(ct).
\end{aligned}$$

□

Theorem 2.6 (Central Limit Theorem)**.** *Given n i.i.d. random variables $X_1, X_2, \ldots, X_n$ where we have $\mathbb{E}[X_i] = \mu$, $\mathbb{V}[X_i] = \sigma^2$, μ and σ are finite, $1 \le i \le n$. Let $\overline{X}_n = \frac{1}{n}\sum_i X_i$ and $Y_n = \sqrt{n}\cdot\frac{(\overline{X}_n - \mu)}{\sigma}$, we have*

$$\begin{aligned}
&\lim_{n\to\infty} m_{Y_n} = e^{\frac{t^2}{2}}, \\
&Y_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1), \\
&\overline{X}_n \xrightarrow{\mathrm{d}} \mathcal{N}(\mu, \frac{\sigma^2}{n}).
\end{aligned} \tag{2.6.3}$$

Proof.

$$\begin{aligned}
Y_n &= \sqrt{n}\cdot\frac{(\overline{X}_n - \mu)}{\sigma} \\
&= \sqrt{n}\cdot\frac{\sum_i (X_i - \mu)}{n\sigma} \\
&= \frac{\sum_i (X_i - \mu)}{\sqrt{n}\sigma} \quad (\text{Let } Z_i = \frac{X_i - \mu}{\sigma}) \\
&= \sum_i \frac{Z_i}{\sqrt{n}}.
\end{aligned}$$

By Theorem 2.5 we have

$$\begin{aligned}
m_{Y_n}(t) &= [m_{\frac{Z_i}{\sqrt{n}}}(t)]^n, \\
m_{\frac{Z_i}{\sqrt{n}}}(t) &= m_{Z_i}(\frac{t}{\sqrt{n}}).
\end{aligned}$$

In our study n approaches infinity and we could study the moment generating function $m_{Z_i}(\frac{t}{\sqrt{n}})$ at the neighborhood of 0 with Taylor series. That is $m_{Z_i}(t) = m_{Z_i}(0) + m'_{Z_i}(0)\cdot t + \frac{m''_{Z_i}(0)\cdot t^2}{2} + \frac{m'''_{Z_i}(t_0)\cdot t^3}{6}$, $0 \le |t_0| \le |t|$. In addition we have $m_{Z_i}(0) = 1$, $m'_{Z_i}(0) = \mathbb{E}[Z_i] = 0$, and $m''_{Z_i}(0) = \mathbb{E}\left[Z_i^2\right] = 1$.

Putting everything together we have

$$m_{Z_i}(\frac{t}{\sqrt{n}}) = 1 + 0 \cdot \frac{t}{\sqrt{n}} + \frac{t^2}{2n} + \frac{m'''_{Z_i}(t_0) \cdot t^3}{6n\sqrt{n}},$$
$$[m_{Z_i}(\frac{t}{\sqrt{n}})]^n = (1 + \frac{t^2}{2n} + \frac{m'''_{Z_i}(t_0) \cdot t^3}{6n\sqrt{n}})^n,$$

Assuming $m'''_{Z_i}(:)$ is finite at a neighborhood of 0, we have

$$\lim_{n\to\infty}[m_{Z_i}(\frac{t}{\sqrt{n}})]^n = e^{\frac{t^2}{2}} \quad t \in (-\mathbb{R}, \mathbb{R}).$$

□

2.7 Bibliographic Notes and Further Reading

With the clear justification of normal distribution, there are many distributions that can be derived from normal distribution. For example, the k-degree chi-squared distribution, or χ^2-distribution, is the distribution of the sum of the squares of k independent standard normal random variables. Given a two-dimensional random vector $\mathbf{z} = (x, y)^T$, if x and y are independent and normally distributed with equal variance, and zero mean, then the distribution of the magnitude of $\mathbf{z}$ (the length of the vector $\mathbf{z}$) is the Rayleigh distribution.

2.8 Exercises

2.1[More Examples of Distributions] Derive the expectation and variance for the following commonly used distributions. Here we use a notation $f_X(x) = f(x|\theta)$ for parametric models where the distributions are specified ("indexed") by a finite vector. The notation $f(x|\theta)$ (or $\mathbb{P}(X = k|\theta)$)simply means the distribute has a parameter θ.

- Poisson distribution $\mathbb{P}(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Binomial distribution $\mathbb{P}(X = k|n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$
- Rademarcher distribution $\mathbb{P}(X = 1|p) = p$ and $\mathbb{P}(X = -1|p) = 1 - p$
- Geometric distribution $\mathbb{P}(X = k|\theta) = (1-\theta)^k\theta$
- Hyper-geometric distribution $\mathbb{P}(X = k|M, N, K) = \frac{\binom{M}{k}\binom{N-M}{K-k}}{\binom{N}{K}}$
- Negative geometric distribution $\mathbb{P}(X = k|\theta, n) = \binom{k-1}{n-1}\theta^n(1-\theta)^{k-n}$
- Normal distribution (one dimensional Gaussian) $X \sim \mathcal{N}(\mu, \sigma^2)$ where $f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\ \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$
- Exponential distribution $f(x|\lambda) = \lambda e^{-\lambda x} \cdot I_{(0,\infty)}(x)$
- Gamma distribution $\Gamma(x)$ with $f(x|k, \theta) = \frac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-\frac{x}{\theta}} \cdot I_{(0,\infty)}(x), k > 0, \theta > 0$
- Beta distribution Beta(α, β) with $f(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \cdot I_{(0,1)}(x), \alpha > 0, \beta > 0$

- Cauchy distribution $f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$
- Lognormal $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp(-\frac{(\ln x-\mu)^2}{2\sigma^2})$
- Double exponential distribution $f(x|\mu,\sigma^2) = \frac{1}{2\sigma}\exp(-\frac{|x-\mu|}{\sigma})$

$I_A(x)$ is an indicator function where $I_A(x) = 1$ if $x \in A$ and 0 otherwise. $\Gamma(k)$ is the gamma function where $\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}\,dt$. $B(\alpha,\beta)$ is the beta function and $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

2.2[Non-identifiable Distribution] A modified Poisson $\mathbb{P}(X = k|\lambda) = \frac{\lambda^{2k}e^{-\lambda^2}}{k!}$. Show that $\mathbb{P}(X|\lambda) = \mathbb{P}(X|-\lambda)$ for all λ and hence the distribution is not identifiable.

2.3[Transformations of Random Variables] Let X be a random variable with $\mathbb{E}[X] = \mu$, $\mathbb{V}[X] = \sigma^2$, and the moment generating function $m_X(t)$. Calculate the expectation, variance, and moment generating function of the following random variables.

- $Y = X + 1$
- $Y = 2X$
- $Y = kX + b$

2.4[Additional Convergence Modes of Random Variables] Prove the following statement

- convergence quadratically implies convergence in probability,
- convergence in probability implies convergence in distribution.

Part II

Learning with Maximum Likelihood Estimation

Chapter 3

Maximum Likelihood Estimation

In this chapter we consider a problem where we have samples that are generated by a model and our mission is to use the samples to estimate the data generating model. Such type of "reverse engineering" is a critical step in predictive modeling. To simplify our mission we concentrate on parametric models where the data generating model, or the probability density function, is indexed by a parameter which may be a scalar, a vector, or a matrix with a finite dimensionality. One approach to estimate the true parameter is that we identify the "best" parameter that maximizes the chance that the data is generated by the parameter. This principle is intuitively appealing and is commonly known as the maximum likelihood estimation method as we should concentrate in this chapter.

3.1 Maximum Likelihood Estimation

Definition 3.1 (Likelihood function)**.** Given *i.i.d.* random variables $X = (X_1, \ldots, X_n)$ from a distribution with the PDF function $f_X(x) = f(x|\boldsymbol{\theta}_0)$ where $\theta_0 \in \Theta$. $\Theta \subseteq \mathbb{R}^n$ is a set of finite dimensional vectors. For each possible sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ the *likelihood function* $\mathcal{L}_n(\theta)$ is a real-valued function defined on Θ as

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \tag{3.1.1}$$

and the *log-likelihood function* is given by

$$\begin{aligned} \ell_n(\theta) &= \ln \mathcal{L}_n(\theta) \\ &= \sum_{i=1}^{n} \ln f(x_i|\theta) \\ &= \sum_{i=1}^{n} \ell(x_i|\theta). \end{aligned} \tag{3.1.2}$$

Here $\boldsymbol{\theta} \in \Theta$. We use the notation $\ell(x|\theta) = \ln f(x|\theta)$ to denote the log PDF at a outcome x in the sample $\mathbf{x}$.

Notice that the likelihood function (and hence the log-likelihood function) is a function of the sample $\mathbf{x}$ and the variable θ. By convention we usually only specify θ as the parameter for the likelihood function and using the subscript n to hint that the function is indexed by a sample of n random variables.

Definition 3.2 (Maximum Likelihood Estimation)**.** Given a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ for n *i.i.d.* random variables X from a distribution with the PDF function $f(x|\boldsymbol{\theta}_0)$ where $\theta_0 \in \Theta$, suppose a statistics $S(\mathbf{x})$ maximizes the likelihood function:

$$S(\mathbf{x}) = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta). \tag{3.1.3}$$

We call such statistics $S(X)$ the maximum likelihood estimation (MLE) of θ_0, denoted by $\widehat{\theta}_{\text{MLE}}$. Since logarithm is a monotonically increasing transformation, the above definition is equivalent to defining $\widehat{\theta}_{\text{MLE}}$ as a statistics to maximize the log-likelihood function as

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \ell_n(\theta). \tag{3.1.4}$$

Example 3.1 (MLE of Bernoulli distribution)**.** With n *i.i.d.* random variables from a Bernoulli distribution with the true parameter $\theta_0 \in \Theta = [0, 1]$, given a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, the log-likelihood function is hence

$$\begin{aligned}
\ell_n(\theta) &= \sum_{i=1}^{n} \ln \mathbb{P}(x_i|\theta) \\
&= \sum_{i=1}^{n} \ln(\theta^{x_i}(1-\theta)^{1-x_i}) \\
&= \sum_{i=1}^{n} (x_i \ln(\theta) + (1-x_i)\ln(1-\theta)),
\end{aligned}$$

which has derivative

$$\frac{d\ell_n(\theta)}{d\theta} = \sum_{i=1}^{n} (\frac{x_i}{\theta} - \frac{1-x_i}{1-\theta}).$$

Setting the derivative to 0, solving for θ, and noticing that $\ell_n(\theta)$ is concave we conclude that

$$S(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

maximizes the log-likelihood function.
Hence,

$$\begin{aligned}
\widehat{\theta}_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
&= \overline{X}
\end{aligned}$$

Example 3.2 (MLE of Exponential distribution)**.** With n *i.i.d.* random variables from an exponential distribution with the true parameter λ_0, given a sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, the log-likelihood function is hence

$$\begin{aligned}
\ell_n(\theta) &= \sum_{i=1}^{n} \ln f(x_i|\lambda) \\
&= \sum_{i=1}^{n} \ln(\lambda \exp(-\lambda x_i)) \\
&= n\ln\lambda - \lambda\sum_{i=1}^{n} x_i,
\end{aligned}$$

which has derivative

$$\frac{d\ell_n(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n}(x_i).$$

Setting the derivative to 0, solving for λ, and noticing that $\ell_n(\lambda)$ is concave we obtain that

$$S(\mathbf{x}) = \frac{n}{\sum_{i=1}^{n} x_i}$$

maximizes the log-likelihood function.
Hence,

$$\begin{aligned}\widehat{\lambda}_{\text{MLE}} &= \frac{n}{\sum_{i=1}^{n} X_i} \\ &= \frac{1}{\overline{X}}\end{aligned}$$

3.2 Global Behavior of MLE and Consistency

Below we provide a very strong theoretic property of MLE: under mild conditions, MLE is consistent. The proof is not straightforward but it is based on a rather simple observation which we call the global behavior of the log-likelihood function. Before we proceed it makes sense for us to study the critical points and extreme values of the log-likelihood function, as what follows.
Let $M_n(\theta)$ be the sample mean of the log-likelihood function of θ or

$$\begin{aligned}M_n(\theta) &= \frac{1}{n}\ell_n(\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\ell(x_i|\theta), \\ &= \frac{1}{n}\sum_{i=1}^{n}\ln f(x_i|\theta)\end{aligned}$$

and let $M(\theta)$ be the expectation of the log likelihood function

$$M(\theta) = \int \ell(x|\theta) f(x|\theta_0)\, dx.$$

Theorem 3.1. *Let θ_0 be the true parameter we have $M(\theta) \leq M(\theta_0)$ for all $\theta \in \Theta$.*

Proof.

$$\begin{aligned}M(\theta) - M(\theta_0) &= \int \ln \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)\, dx \\ &\leq \int \left(\frac{f(x|\theta)}{f(x|\theta_0)} - 1\right) f(x|\theta_0)\, dx \qquad (\text{since } \ln t \leq t-1 \text{ for all } t > 0.) \\ &= \int f(x|\theta)\, dx - \int f(x|\theta_0)\, dx \\ &= 1 - 1 \\ &= 0.\end{aligned} \tag{3.2.1}$$

□

In order to show the consistency of MLE, we first would like to establish the connection of the parameter that maximizes the likelihood function and the true parameter. For identifiable distributions where different parameters lead to different density functions, we conclude that $M(\theta) < M(\theta_0)$ if $\theta \neq \theta_0$. To reach the conclusion we notice that $\ln t = t - 1$ only if $t = 1$ (see the comment in (3.2.1)). In other words, θ_0 is the only global maximum of $M(\theta)$. We call the fact established by the Theorem (3.1) as the global behavior of the MLE. However we have not established the relationship between MLE and the true parameter yet, which is what we are about to do.
The quantity $M(\theta) - M(\theta_0) = \int \ln \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0)\, dx$ is the *Kullback-Leibler divergence* of the PDF $f(x|\theta)$ from $f(x|\theta_0)$. KL divergence is an important concepts in predictive analytics and we introduce the general definition below.

Definition 3.3 (Kullback-Leibler divergence)**.** The KL divergent of a PDF g from f is

$$\mathrm{D}_{\mathrm{KL}}(f||g) = \int f(x) \ln \frac{f(x)}{g(x)} dx. \tag{3.2.2}$$

Readers might already notice that the KL divergence is asymtric in that we usually have $\mathrm{D}_{\mathrm{KL}}(f||g) \neq \mathrm{D}_{\mathrm{KL}}(g||f)$. We further illustrate the concept of KL divergence using the following example.

Example 3.3. With an exponential distribution family where the PDF takes the form of $f(x|\theta) = \lambda \exp(-\lambda x)$. The KL divergence of a PDF indexed by $\lambda > 0$ from that indexed by $\lambda_0 > 0$ is

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\lambda_0||\lambda) &= \int f(x|\lambda_0) \ln \frac{f(x|\lambda_0)}{f(x|\lambda)} dx \\
&= \int \lambda_0 \exp(-\lambda_0 x) \ln \frac{\lambda_0 \exp(-\lambda_0 x)}{\lambda \exp(-\lambda x)} dx \\
&= \int \lambda_0 \exp(-\lambda_0 x)(\ln \frac{\lambda_0}{\lambda} + (\lambda - \lambda_0)x) dx \\
&= \ln \frac{\lambda_0}{\lambda} \cdot \int \lambda_0 \exp(-\lambda_0 x) dx + (\lambda - \lambda_0) \cdot \int \lambda_0 x \exp(-\lambda_0 x) dx \\
&= \ln \frac{\lambda_0}{\lambda} + \frac{\lambda - \lambda_0}{\lambda_0} \\
&= \frac{\lambda}{\lambda_0} - 1 - \ln \frac{\lambda}{\lambda_0} \\
&\geq 0 \quad (\text{since } t - 1 \geq \ln t \text{ for all } t > 0.)
\end{aligned}$$

The KL divergence takes the value 0 if and only if $\lambda = \lambda_0$. Clearly $\mathrm{D}_{\mathrm{KL}}(\lambda||\lambda_0) = \frac{\lambda_0}{\lambda} - 1 - \ln \frac{\lambda_0}{\lambda} \neq \mathrm{D}_{\mathrm{KL}}(\lambda_0||\lambda)$.
With the KL divergence we are about to show a very strong theoretic results of MLE.

Theorem 3.2. *For identifiable distributions the maximum likelihood estimation is consistent, i.e.* $\widehat{\theta}_{\mathrm{MLE}} \xrightarrow{\mathrm{P}} \theta_0$*, where* θ_0 *is the true parameter. In addition with mild assumptions* $\widehat{\theta}_{\mathrm{MLE}}$ *converges almost surely to* θ_0 *or* $\widehat{\theta}_{\mathrm{MLE}} \xrightarrow{\mathrm{a.s.}} \theta_0$*.*

Proof. There are two key components that we should consider in proving the consistency of MLE. One is the characteristics of the parameter set Θ and the other is the characteristics of $M(\theta)$. Here we consider a rather simplex case where Θ is a finite set. In addition we use the strong law of large number to prove that $\widehat{\theta}_{\mathrm{MLE}}$ converges almost surely. Our approach has two advantages. First we do

not need to make assumptions of $M(\theta)$ and second convergence almost surely implies convergence in probability.
Now consider an event where for all $\theta \in \Theta$ we have $\lim_{n\to\infty} M_n(\theta) = M(\theta)$. We notice that this event implies that $\widehat{\theta}_{\text{MLE}} = \theta_0$ by (3.2.1) since θ_0 is the only global maximum among all $M(\theta)$. By the strong law of large numbers we have $\mathbb{P}(\lim_{n\to\infty} M_n(\theta) \neq M(\theta)) = 0$ for each θ. If Θ is finite by the union bound of probability we have $\mathbb{P}(\exists\theta \in \Theta, \lim_{n\to\infty} M_n(\theta) \neq M(\theta)) = 0$. Hence we conclude that $\widehat{\theta}_{\text{MLE}} \xrightarrow{\text{a.s.}} \theta_0$.
Although the proof is rigorous, the constraint that Θ is a finite set is much less desirable. We are not going to try to extend the proof to general cases where Θ is infinite. We will do so in a later chapter. □

3.3 Local Behavior of MLE and Fisher Information

We start this section by first studying a simple example of exponential distribution. In Example (3.3) we calculate the KL divergence of $\text{D}_{\text{KL}}(\theta_0||\theta)$ as $\text{D}_{\text{KL}}(\theta_0||\theta) = \frac{\lambda}{\lambda_0} - 1 - \ln\frac{\lambda}{\lambda_0}$.
In the previous section we show that θ_0 is a critical point of $M(\theta)$ if θ_0 is the true parameter of a parametric distribution family. Since every global maximum must be a local maximum as well, we study the local behavior of MLE by studying the derivatives of $M(\theta)$.
In order to derive those derivatives we have to make a few assumptions. First with some regularity we always assume we are provisioned with the interchangeability of a derivative and an integral for $M(\theta)$. In addition $M(\theta)$ is twice differentiable at a neighborhood of θ_0.
We first show that $M(\theta)$ has a critical point θ_0 as

$$\begin{aligned}
\left.\frac{\partial M(\theta)}{\partial\theta}\right|_{\theta=\theta_0} &= \frac{\partial\int_{-\infty}^{\infty}\ell(x|\theta)f(x|\theta_0)\,dx}{\partial\theta} && (\theta=\theta_0)\\
&= \frac{\partial\int_{-\infty}^{\infty}\ln f(x|\theta)f(x|\theta_0)\,dx}{\partial\theta}\\
&= \int_{-\infty}^{\infty}\frac{\partial\ln f(x|\theta)}{\partial\theta}f(x|\theta_0)\,dx\\
&= \int_{-\infty}^{\infty}\frac{f(x|\theta_0)}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}\,dx\\
&= \int_{-\infty}^{\infty}\frac{\partial f(x|\theta)}{\partial\theta}\,dx\\
&= \frac{\partial\int_{-\infty}^{\infty}f(x|\theta)\,dx}{\partial\theta}\\
&= 0 && (\int_{-\infty}^{\infty}f(x|\theta)\,dx = 1 \text{ for all } \theta).
\end{aligned} \tag{3.3.1}$$

We use notations $\ell'(x|\theta) = \frac{\partial\ell(x|\theta)}{\partial\theta}$ and $\ell''(x|\theta) = \frac{\partial^2\ln f(x;\theta)}{\partial\theta^2}$. With the rotations (3.3.1) can be rewritten as

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\ell'(x|\theta)\right] &= \int_{-\infty}^{\infty}\frac{\partial\ln f(x|\theta)}{\partial\theta}f(x|\theta_0)\,dx\\
&= 0.
\end{aligned} \tag{3.3.2}$$

We then compute the second derivative $M(\theta)$ at θ_0 as

$$\begin{aligned}
\left.\frac{\partial^2 M(\theta)}{\partial\theta^2}\right|_{\theta=\theta_0} &= \frac{\partial\left(\frac{\partial\int_{-\infty}^{\infty}\ell(x|\theta)f(x|\theta_0)\,dx}{\partial\theta}\right)}{\partial\theta} && (\theta=\theta_0)\\
&= \frac{\partial\int_{-\infty}^{\infty} f(x|\theta_0)\ell'(x|\theta)\,dx}{\partial\theta}\\
&= \frac{\partial\int_{-\infty}^{\infty}\frac{f(x|\theta_0)}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}\,dx}{\partial\theta}\\
&= \int_{-\infty}^{\infty}\frac{f(x|\theta_0)}{f(x|\theta)f(x|\theta)}\left(\frac{\partial^2 f(x|\theta)}{\partial\theta^2}f(x|\theta)-\frac{\partial f(x|\theta)}{\partial\theta}\frac{\partial f(x|\theta)}{\partial\theta}\right)dx && (3.3.3)\\
&= \int_{-\infty}^{\infty}\frac{\partial^2 f(x|\theta)}{\partial\theta^2}\,dx-\int_{-\infty}^{\infty}\frac{f(x|\theta_0)}{f(x|\theta)f(x|\theta)}\left(\frac{\partial f(x|\theta)}{\partial\theta}\right)^2 dx\\
&= 0-\int_{-\infty}^{\infty} f(x|\theta_0)\left(\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}\right)^2 dx\\
&= -\int_{-\infty}^{\infty} f(x|\theta_0)\left(\ell'(x|\theta)\right)^2 dx\\
&= -\mathbb{E}_{\theta_0}\left[(\ell'(x|\theta))^2\right]\\
&\le 0.
\end{aligned}$$

The derivation used a fact that $\int_{-\infty}^{\infty}\frac{\partial^2 f(x|\theta)}{\partial\theta^2}\,dx=0$. This is because $\frac{\partial\int_{-\infty}^{\infty}f(x|\theta)\,dx}{\partial\theta}=0$ for all θ as shown in (3.3.1). (3.3.3) shows that θ_0 might be a local maximum of the function $M(\theta)$. It turns out we could prove a much stronger result regarding $M(\theta)$ *i.e.* $M(\theta)$ reaches its global maximum at θ_0. Before we do that we first further explore the statistical implications of (3.3.1) and (3.3.3) with connections to Fisher information as defined below:

Definition 3.4 (Fisher information)**.** Let X be a random variable with PDF $f_X(x)=f(x|\theta_0)$. The *Fisher information* is defined as

$$\begin{aligned}
I(\theta) &= \mathbb{E}_{\theta_0}\left[\left(\frac{\partial\ln f(x|\theta)}{\partial\theta}\right)^2\right]\\
&= \mathbb{E}_{\theta_0}\left[(\ell'(x|\theta))^2\right].
\end{aligned} \tag{3.3.4}$$

With the definition we have

$$\begin{aligned}
I(\theta) &= \mathbb{E}_{\theta_0}\left[(\ell'(x|\theta))^2\right]\\
&= \mathbb{E}_{\theta_0}\left[(\ell'(x|\theta)-0)^2\right]\\
&= \mathbb{E}_{\theta_0}\left[\left(\ell'(x|\theta)-\mathbb{E}_{\theta_0}\left[\ell'(x|\theta)\right]\right)^2\right] \qquad (\ \mathbb{E}_{\theta_0}\left[\ell'(x|\theta)\right]=0\ )\\
&= \mathbb{V}_{\theta_0}\left[\ell'(x|\theta)\right].
\end{aligned} \tag{3.3.5}$$

So Fisher information is the variance of the partial derivative of the log-likelihood function evaluated at θ_0. We define $J(\theta)$ as the expected value of the second derivative of $\ell(\theta)$ as evaluated at θ_0 or

$$\begin{aligned}
J(\theta) &:= -\mathbb{E}_{\theta_0}\left[\frac{\partial^2\ln f(x|\theta)}{\partial\theta^2}\right]\\
&= -\mathbb{E}_{\theta_0}\left[\ell''(x|\theta)\right].
\end{aligned} \tag{3.3.6}$$

With this definition (3.3.3) is conveniently rewritten as

$$I(\theta) = -J(\theta). \tag{3.3.7}$$

So the Fisher information is the negative of the expected second derivative of the log-likelihood function evaluated at the true parameter θ_0. The notation of $I(\theta)$ and $J(\theta)$ are confusing at the first glance but all we are doing here is to compute two quantities (rather than defining two functions). Fisher information is always non-negative. A high value of Fisher information indicates a sharp curve of the log-likelihood function $\ell(\theta)$ and hence estimation is easy and a low value of Fisher information indicates a flat curve of the log-likelihood function $\ell(\theta)$ and hence accurate estimation is difficult. In this sense Fisher information provides very useful information of the local behavior of $\ell(\theta)$ around the true parameter θ_0.

3.4 Asymptotic Normality of MLE

We have demonstrated two properties of MLE: it is consistent and it converges to the true parameter almost surely. We now come to the third property of the distribution of MLE: it is asymptotically normal.

Theorem 3.3. *The maximum likelihood estimation has an asymptotic normal distribution, more specifically*

$$\sqrt{n}\left(\widehat{\theta}_{\mathrm{MLE}} - \theta_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \frac{1}{I(\theta_0)}).$$

Proof. Let $X_1, \ldots, X_n$ be i.i.d. with PDF $f(x|\theta)$. Consider the derivative of the log-likelihood function normalized by n,

$$\ell_n'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln f(X_i|\theta)}{\partial \theta}.$$

By the Mean Value Theorem, there exists θ_n^* between $\widehat{\theta}_{\mathrm{MLE}}$ and θ_0 such that

$$\ell_n'(\widehat{\theta}_{\mathrm{MLE}}) - \ell_n'(\theta_0) = \ell_n''(\theta_n^*)(\widehat{\theta}_{\mathrm{MLE}} - \theta_0).$$

Since $\ell_n'(\widehat{\theta}_{\mathrm{MLE}}) = 0$,

$$\widehat{\theta}_{\mathrm{MLE}} - \theta_0 = -\frac{\ell_n'(\theta_0)}{\ell_n''(\theta_n^*)}.$$

By equation (3.3.2),

$$\mathbb{E}\left[\ell_n'(\theta_0)\right] = 0$$

and by the definition of Fisher information (3.3.4),

$$\mathbb{V}\left[\ell_n'(\theta_0)\right] = \frac{I(\theta_0)}{n}.$$

Therefore, by the Central Limit Theorem,

$$\sqrt{n}\,\ell_n'(\theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, I(\theta_0)).$$

By Theorem (3.2) $\theta_n^* \xrightarrow{\mathrm{P}} \theta_0$. By the Law of Large Numbers we have

$$\begin{aligned} -\ell_n''(\theta_n^*) &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i|\theta)}{\partial\theta^2}|_{\theta=\theta_n^*} \\ &\xrightarrow{\mathrm{P}} -\mathbb{E}\left[\frac{\partial^2 \ln f(x|\theta)}{\partial\theta^2}\right]|_{\theta=\theta_n^*} \\ &= I(\theta_0). \end{aligned}$$

The last step follows from (3.3.7).
Let $Y \sim \mathcal{N}(0, I(\theta_0))$, then

$$\frac{Y}{I(\theta_0)} \sim \mathcal{N}(0, \frac{1}{I(\theta_0)})$$

Hence,

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_{\mathrm{MLE}} - \theta_0) &= \frac{\sqrt{n}\,\ell_n'(\theta_0)}{-\ell_n''(\theta_n^*)} \\ &\xrightarrow{d} \frac{Y}{I(\theta_0)} \\ &\sim \mathcal{N}(0, \frac{1}{I(\theta_0)}). \end{aligned}$$

□

3.5 Fisher Information Matrix(Need some re-edit on this section)

Let us consider a multivariate case for Fisher Information.

Theorem 3.4.

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right]$$

Proof.
By theorem **??**,

$$0 = \int_{-\infty}^{\infty} \frac{\partial \ln(f(x;\theta))}{\partial\theta} f(x;\theta)\,dx.$$

Taking the derivative of both sides with respect to θ we get

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} \frac{\partial \ln(f(x;\theta))}{\partial\theta} f(x;\theta)\,dx \\ &= \int_{-\infty}^{\infty} \frac{\partial^2 \ln(f(x;\theta))}{\partial\theta^2} f(x;\theta) + \frac{\partial \ln(f(x;\theta))}{\partial\theta}\frac{\partial \ln(f(x;\theta))}{\partial\theta} f(x;\theta)\,dx \\ &= \int_{-\infty}^{\infty} \frac{\partial^2 \ln(f(x;\theta))}{\partial\theta^2} f(x;\theta)\,dx + \int_{-\infty}^{\infty} \left(\frac{\partial \ln(f(x;\theta))}{\partial\theta}\right)^2 f(x;\theta)\,dx \\ &= \mathbb{E}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right] \\ &= \mathbb{E}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right] + I(\theta). \end{aligned}$$

□

Let us work on an example of Fisher information below.

Example 3.4. Given a Gaussian distribution $\mathcal{N}(x|\mu_0, 1)$.

$$\begin{aligned}
\ell(x;\mu) &= \ln \mathcal{N}(x|\mu, 1) \\
&= -\frac{1}{2}(x-\mu)^2 - \ln\sqrt{2\pi}, \\
\ell'(x;\mu) &= \frac{\partial \ell(x|\mu)}{\partial \mu} \qquad\qquad \ell''(x|\mu) = \frac{\partial^2 \ell(x|\mu)}{\partial \mu^2} \\
&= x - \mu, \qquad\qquad = -1, \\
\mathbb{E}_{\mu_0}\left[\ell'(x;\mu)\right] &= \mu_0 - \mu \qquad\qquad \mathbb{E}_{\mu_0}\left[\ell''(x|\mu)\right] = -1, \\
&= 0, \\
\mathbb{V}_{\mu_0}\left[\ell'(x;\mu)\right] &= \mathbb{E}_{\mu_0}\left[(x-\mu)^2\right] \\
&= \mathbb{E}_{\mu_0}\left[(x-\mu_0+\mu_0-\mu)^2\right] \\
&= 1 + (\mu_0 - \mu)^2 \\
&= 1.
\end{aligned}$$

With the results we calculate $I(\mu)$ and $J(\mu)$ as

$$\begin{aligned}
I(\mu) &= \mathbb{E}_{\mu_0}\left[(\ell'(x|\mu))^2\right] \qquad\qquad J(\mu) = \mathbb{E}_{\mu_0}\left[\ell''(x|\mu)\right] \\
&= 1, \qquad\qquad = -1.
\end{aligned}$$

Clear we have

$$\begin{aligned}
I(\mu) &= \mathbb{V}_{\mu_0}\left[\ell'(x;\mu)\right], \\
I(\mu) &= -\mathbb{E}_{\mu_0}\left[\ell''(x;\mu)\right], \\
I(\mu) &= -J(\mu).
\end{aligned}$$

3.6 Bibliographic Notes and Further Reading

[10] (Chapter 9) provides alternative way to prove the consistency of MLE by assuming the peak of $M(\theta)$ is "well isolated". For a detailed discussion on the challenges of proving the consistency of MLE see [5] (Chapter 5). Critiques of MLE may be found in many places include [7] (Chapter 6).

3.7 Exercises

3.1[More Examples of MLE] Suppose that we have n *i.i.d.* samples from the distributions. Derive the maximum likelihood estimation as instructed.

- Poisson distribution $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, derive $\widehat{\lambda}_{\text{MLE}}$.
- Binomial distribution $\mathbb{P}(X = k) = \binom{n}{k}\theta^k(1-\theta)^{(n-k)}$, derive $\widehat{\theta}_{\text{MLE}}$.

- Rademarcher distribution $\mathbb{P}(X = 1|p) = p$ and $\mathbb{P}(X = -1|p) = 1 - p$, derive $\widehat{p}_{\text{MLE}}$.
- Geometric distribution $\mathbb{P}(X = k) = (1 - \theta)^k\theta$, derive $\widehat{\theta}_{\text{MLE}}$.
- One-dimensional Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$, derive $\widehat{\mu}_{\text{MLE}}$, assuming σ is fixed.
- Multi-dimensional Gaussian distribution $\mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, derive $\widehat{\boldsymbol{\mu}}_{\text{MLE}}$, assuming $\boldsymbol{\Sigma}$ is fixed.
- Exponential distribution $f(x) = \lambda e^{-\lambda x}$ if $x \geq 0$ and $f(x) = 0$ otherwise. Derive $\widehat{\lambda}_{\text{MLE}}$.

3.2[Fisher Information] For the distributions listed above, set $\ell(\theta) = \ln f(x; \theta)$ and θ_0 is the true parameter. Compute $\ell'(\theta)$ and $\ell''(\theta)$, and the Fisher information $I(\theta)$. Show that $\mathbb{E}_{\theta_0}[\ell'(\theta)] = 0$, $\mathbb{E}_{\theta_0}[\ell''(\theta)] = -I(\theta)$.

3.3[Cauchy distribution] Cauchy distribution with the pdf function as $f(X; \mu, \gamma) = \frac{1}{\pi}\frac{\gamma}{(X-\mu)^2+\gamma^2}$. Notice that Cauchy distribution is always identifiable. Assuming that γ is fixed. Show that identifying $\widehat{\mu}_{\text{MLE}}$ requires to solve a polynomial with degree $2n$-1 where n is the total number of samples.

3.4[Degerate Cases of MLE] Let $X_1, X_2, \ldots, X_n$ be n *i.i.d.* samples uniformly distributed in $(0, \theta_0)$, or $f(X; \theta_0) = 1/\theta_0$. $\theta_0 > 0$ is the model parameter. Show that the likelihood function is $\mathcal{L}(\theta) = \theta^{-n} I(\theta < \max(X_1, X_2, \ldots, X_n))$. Show that $\widehat{\theta}_{\text{MLE}}$ does not exist.

Chapter 4

Linear Regression

Linear regression is one of the most studied problem in data analytics. The general idea of linear regression is to fit a line in order to map independent variables to dependent variables. In this chapter we study a general form of linear regression with least square fitting in finite dimensional spaces. Linear regression is quite special in data analytics in that we are able to obtain many analytic solutions for the learning problem. We should take advantage of this special case and gain insights regarding linear regression by studying the generalization error of our parameter estimation and our prediction estimation. We shall see that some of the insights are applicable to a wide range of problems in subsequent chapters. We begin our study by working on a special case of linear regression in a one-dimensional space.

4.1 Linear Regression in One Dimensional Space

Suppose that we have the training data set $\{(x_i, y_i)\}$ with n data points ($i \in \{1, \ldots, n\}$). x_i and y_i are scalars and we call x_i the independent variable and y_i the dependent variable. The goal of linear regression here is to find a linear function in the form of $f(x_i) = \beta_1 x_i + \beta_2$ that maps x_i to y_i with minimal loss.
We choose to use the squared loss for regression. Hence the lost function $L(\beta_1, \beta_2)$ is written as

$$L(\beta_1, \beta_2) = \sum_{i=1}^{n} [y_i - (\beta_1 x_i + \beta_2)]^2.$$

In order to minimize the loss function, we take the partial derivatives and set them to zero. That is

$$\frac{\partial L(\beta_1, \beta_2)}{\partial \beta_1} = -\sum_{i=1}^{n} 2[y_i - (\beta_1 x_i + \beta_2)] \cdot x_i = 0,$$
$$\frac{\partial L(\beta_1, \beta_2)}{\partial \beta_2} = \sum_{i=1}^{n} 2[y_i - (\beta_1 x_i + \beta_2)] = 0.$$

From (4.1) we obtain part of the solution as

$$\bar{y} = \beta_1 \bar{x} + \beta_2,$$

where $\bar{x} = \frac{\sum_i x_i}{n}$ and $\bar{y} = \frac{\sum_i y_i}{n}$.

Substitute it back to (4.1), we obtain the final result:

$$\beta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}\bar{x}},$$
$$\beta_2 = \frac{\overline{xy}\bar{x} - \bar{y}\overline{x^2}}{\bar{x}\bar{x} - \overline{x^2}},$$

where $\overline{xy} = \frac{\sum_i x_i * y_i}{n}$ and $\overline{x^2} = \frac{\sum_i x_i^2}{n}$

4.2 Linear Regression in High Dimensional Space

In general we have several independent variables. If that is the case the sample $\mathbf{x_i} \in \mathbb{R}^p$ is a p-dimensional real vector rather than a scalar. Our training data set is hence $\{(\mathbf{x_i}, y_i)\}$ where $i \in \{1, \ldots, n\}$. The dependent variable y_i is still a scalar. The goal of linear regression is to find a linear function in the format of $f(x_i) = \boldsymbol{\beta}^T \mathbf{x_i}$ that minimize the squared loss function. Notice that in this format, we do not have an explicit constant for the linear function as we did in the previous section. To include the constant, we could simply add a constant 1 to each vector $\mathbf{x}_i$, or $\mathbf{x} = [x_1, x_2, \ldots, x_p, 1]^T$. Since we include the constant factor, we usually assume the expectation of an object is zero, or $\mathbb{E}[\mathbf{x}] = 0$. In the following discussion we always assume there is a constant 1 as a feature included in each object $\mathbf{x}$.
Following convention we use an n by p matrix $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}]^T \in \mathbb{R}^{n \times p}$ to organize all the samples that we have. Such matrix is called an object-feature matrix where each row is an object and each column is a feature. Other names are used in literature. Notice that even with this notation we still use a column vector to represent a sample. $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$ is an n-dimensional vector of labels for $\mathbf{X}$.
To wrap up, in multivariate linear regression we have

$$\mathbf{X} = \left[\ \mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\ \right]^T_{n \times p}, \qquad \mathbf{x_i} \in \mathbb{R}^p, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}, \qquad f(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}.$$

For each object (matrix or vector), we explicitly label the dimensionality of the object to avoid confusion.
We formulate the problem of $\boldsymbol{\beta}$ estimation as an optimization problem to find the optimal coefficients $\widehat{\beta}_{\text{LR}}$ as

$$\widehat{\beta}_{\text{LR}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X} \cdot \boldsymbol{\beta}||_2^2 .$$

To solve the optimization problem, we define $L(\boldsymbol{\beta})$ as

$$\begin{aligned} L(\boldsymbol{\beta}) =& ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \\ =& (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ =& \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ =& \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Taking the partial derivative, we have

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}.$$

Setting the partial derivatives to zero we have $\hat{\boldsymbol{\beta}}_{\text{LR}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

4.3 Liner Regression with Least Squre Fitting is Consistent

Below we offer a probabilistic justification of the combination of linear regression with least square fitting. Our justification is based the observation that convergence in L2 norm guarantees convergence in probability. Specifically we are going to show that

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\left|\widehat{\boldsymbol{\beta}}_{\text{LR}} - \boldsymbol{\beta}_0\right|\right|_2^2\right] = 0, \tag{4.3.1}$$

given that $\boldsymbol{\beta_0}$ is the "true" parameter.
As we did in maximum likelihood estimation, we first set up the data generation process using a probabilistic model. We assume that $\mathbf{y}$ is obtained by introducing a noise component $\boldsymbol{\epsilon}$ to $\mathbf{X}\boldsymbol{\beta}$ in the way that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}. \tag{4.3.2}$$

We further assume that $\boldsymbol{\epsilon}$ is a multivariate Gaussian variable with zero mean and diagonal covariance matrix $\sigma^2\mathbf{I}$ or $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. β_0 is the true parameter of linear regression.
To prove (4.3.1), we first show $\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{\text{LR}}\right] = \boldsymbol{\beta}_0$. This is true since $\widehat{\boldsymbol{\beta}}_{\text{LR}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{\text{LR}}\right] &= \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon})\right] \\
&= \mathbb{E}\left[\boldsymbol{\beta}_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right] \\
&= \boldsymbol{\beta}_0.
\end{aligned}$$

To complete our proof we have

$$\begin{aligned}
\mathbb{E}\left[\left|\left|\hat{\boldsymbol{\beta}}_{LR} - \boldsymbol{\beta}_0\right|\right|_2^2\right] &= \mathbb{E}\left[\left|\left|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right|\right|_2^2\right] \\
&= \mathbb{E}\left[\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right] \quad (\mathbf{X}^T\mathbf{X} textissymmetric!) \\
&= \mathbb{E}\left[\text{tr}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}))\right] \\
&= \sigma^2\text{tr}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) \quad (\mathbb{E}\left[\boldsymbol{\epsilon}\cdot\boldsymbol{\epsilon}^T\right] = \sigma^2\mathbf{I}) \\
&= \sigma^2\text{tr}((\mathbf{X}^T\mathbf{X})^{-1}).
\end{aligned}$$

Now we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\left|\left|\widehat{\boldsymbol{\beta}}_{\text{LR}} - \boldsymbol{\beta}_0\right|\right|_2^2\right] &= \lim_{n\to\infty} \sigma^2\text{tr}((\mathbf{X}^T\mathbf{X})^{-1}) \\
&= \lim_{n\to\infty} \sigma^2\frac{\text{tr}([\text{Cov}\left[\mathbf{x}\right]]^{-1})}{n} \\
&= 0.
\end{aligned}$$

The last step follows naturally from law of large numbers.

4.4 Least Square Fitting and Maximum Likelihood Estimation

The previous result is not surprising at all. Below we show that linear regression with least square fitting is a maximum likelihood estimation. The consistency immediately follows the fact that all MLEs are consistent (with mild assumptions). To set up the stage of MLE, we modify (4.3.2) a little bit so that the problem has a familiar parameter estimation format. In this set up we assume $\mathbf{X}$ is a constant. $\mathbf{y}$ is a random variable following a multi-dimensional Gaussian distribution with the expectation $\mathbf{X}\beta$ and the covarance matrix $\mathbf{\Sigma} = \sigma^2 \cdot \mathbf{I}$, or

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{\Sigma}). \tag{4.4.1}$$

The likelihood function of $\boldsymbol{\beta}$ is then

$$\mathcal{L}_n(\boldsymbol{\beta}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))$$

We have

$$\begin{aligned}
\widehat{\beta}_{\text{MLE}} &= \arg\max_{\boldsymbol{\beta}} \ln \mathcal{L}_n(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \Big(-\frac{p}{2}\ln(2\pi) + -\frac{1}{2}\ln(\sigma^2) + -\frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \Big) \\
&= \arg\max_{\boldsymbol{\beta}} -\frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \\
&= \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \\
&= \hat{\boldsymbol{\beta}}_{\text{LR}}.
\end{aligned}$$

The equation is true since σ and π are constants. To maximize a number x is to minimize the opposite of x.
Since all maximum likelihood estimators are consistent, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(\Big|\Big|\widehat{\beta}_{\text{LR}} - \beta_0\Big|\Big|_2^2 > \epsilon) = 0,$$
$$\text{or } \widehat{\beta}_{\text{LR}} \xrightarrow{\text{P}} \beta_0.$$

4.5 Generalization Error of Prediction Estimation

Remember that we use the lost function $L(\beta) = \Big|\Big|\mathbf{X}\hat{\beta} - y_0\Big|\Big|_2^2$, we derive the generalization error in terms of a bias and variance decomposition of the estimator.

$$\begin{aligned}
\mathbb{E}_{\text{error}} &= \mathbb{E}\left[L\right] \\
&= \mathbb{E}\left[(\hat{y} - y_0)^2\right] \\
&= \mathbb{E}\left[(\hat{y} - \mathbb{E}\left[\hat{y}\right] + \mathbb{E}\left[\hat{y}\right] - y_0)^2\right] \\
&= \mathbb{E}\left[(\hat{y} - \mathbb{E}\left[\hat{y}\right])^2\right] + \mathbb{E}\left[(\mathbb{E}\left[\hat{y}\right] - y_0)^2\right] + \mathbb{E}\left[2(\hat{y} - \mathbb{E}\left[\hat{y}\right])(\mathbb{E}\left[\hat{y}\right] - y_0)\right].
\end{aligned}$$

Clearly, we see that the first term of the result is the variance of the estimator $\hat{y}$. The second term is the squared value of the difference between the true value y_0 and the expected value of the estimation of $\hat{y}$. This difference is called the *bias* of the estimator. The last term is zero since

$\mathbb{E}[\hat{y}] - \mathbb{E}[\mathbb{E}[\hat{y}]]$ is zero. In other words the bias of an estimator θ is defined as $bias = \mathbb{E}\left[\hat{\theta} - \theta_0\right]$, where θ_0 denotes the ground truth.
Calculate the mean and variance of the prediction $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}X^T\mathbf{y}$:

$$\begin{aligned}
\mathbb{E}[\hat{\mathbf{y}}] &= \mathbb{E}\left[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon)\right] \\
&= \mathbf{X}\beta \\
&= \mathbf{y_0}, \\
\mathbb{E}[\hat{y}_i] &= y_{i0}, \\
\mathbb{V}[\hat{y}_i] &= \mathbb{E}\left[(\hat{\mathbf{y}} - \mathbf{y_0})^T(\hat{\mathbf{y}} - \mathbf{y_0})\right]/n \\
&= \mathbb{E}\left[(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon)^T(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon)\right]/n \\
&= \mathbb{E}\left[tr(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\right]/n \\
&= \sigma^2 tr(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)/n \\
&= \sigma^2 tr((\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X}))/n \\
&= \sigma^2 tr(\mathbf{I}_{p\times p})/n \\
&= \sigma^2 p/n.
\end{aligned}$$

4.6 Bibliographic Notes and Further Reading

4.7 Exercises

Chapter 5

Linear Classification

In this chapter we study a different predictive task where rather than predicting a numeric score we want to predict a categorical label. This is commonly known as a classification problem. In studying the problem, for the time being, we constrain ourself to linear models where the meaning of the those models will be clear soon.

5.1 Logistic regression

Below we introduce a conditional model towards linear classification. To start with a simple problem, we limit ourselves to the case where we have only two types of labels, known as binary classification. For the time being, we use 1 and 0. Following this we have y_i takes value either 1 or 0, with the probability depends on $\mathbf{x}_i$ and a parameter $\boldsymbol{\beta}$. $\mathbf{x}_i$ is a $p \times 1$ vector, the size of $\boldsymbol{\beta}$ is also $p \times 1$. $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$ and $\mathbf{X}$ is a $n \times p$ matrix.

$$\mathbb{P}(y_i = 1|\mathbf{x}_i, \beta) = \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}$$
$$\mathbb{P}(y_i = 0|\mathbf{x}_i, \beta) = 1 - \mathbb{P}(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}$$

Hence, we can write the likelihood function $\mathcal{L}(\boldsymbol{\beta})$ as

$$\begin{aligned}\mathcal{L}_n(\boldsymbol{\beta}) &= \mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \mathbb{P}(y_i|\mathbf{x_i}, \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} (\frac{e^{\boldsymbol{\beta}^T \mathbf{x_i}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_i}}})^{y_i} (\frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_i}}})^{1-y_i}\end{aligned}$$

Here we use the trick that $p^{y_i}(1-p)^{1-y_i}$ is p if y_i is 1 and $1-p$ if y_i is 0.
One step further, the log-likelihood function can be shown to be

$$\begin{aligned}\ell_n(\beta) &= \ln \mathcal{L}_n(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} [y_i \beta^T \mathbf{x_i} - y_i \ln(1 + e^{\beta^T \mathbf{x_i}}) - (1 - y_i) \ln(1 + e^{\beta^T \mathbf{x_i}})] \\ &= \sum_{i=1}^{n} [y_i \beta^T \mathbf{x_i} - \ln(1 + e^{\beta^T \mathbf{x_i}})]\end{aligned}$$

Now the problem turns into an optimization problem

$$\begin{aligned}\widehat{\beta}_{\text{MLE}} &= \arg\max_{\beta} \ell_n(\beta) \\ &= \arg\max_{\beta} \sum_{i=1}^{n} [y_i \beta^T \mathbf{x_i} - \ln(1 + e^{\beta^T \mathbf{x_i}})]\end{aligned} \tag{5.1.1}$$

Note that, the objective function is a convex function, which means it either has one global maximum or is unbounded. Thus, calculating the vector of the partial derivatives of $\ell_n(\boldsymbol{\beta})$ (gradient) with respect to β and setting it to 0 give us the solution to this optimization problem.

$$\frac{\partial \ell_n(\beta)}{\partial \beta} = \sum_{i=1}^{n} [y_i \mathbf{x_i} - \frac{e^{\boldsymbol{\beta}^T \mathbf{x_i}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_i}}} \cdot \mathbf{x_i}] = 0$$

Here we have a set of p equations solving for p parameters. We may have a unique solution. However, it is not easy to get a close-form solution expression from this system of equations. Instead, below we use the Newton-Raphson method to obtain the solution numerically. Newton-Raphson method is an iterative method. Initiatively β^{old} is set to zero and we use the following formula to update β:

$$\beta^{\text{new}} = \beta^{\text{old}} - \left[\frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^T}\right]^{-1} \frac{\partial \ell_n(\beta)}{\partial \boldsymbol{\beta}},$$

where $\frac{\partial^2 f}{\partial \beta \partial \beta^T}$ is the Hessian matrix of a real-valued function f.
To move forward we rewrite the gradient as

$$\begin{aligned}\frac{\partial \ell_n(\beta)}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} [y_i \mathbf{x_i} - \frac{e^{\boldsymbol{\beta}^T \mathbf{x_i}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_i}}} \cdot \mathbf{x_i}] \\ &= \sum_{i=1}^{n} [y_i - \frac{e^{\boldsymbol{\beta}^T \mathbf{x_i}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_i}}}] \mathbf{x_i} \\ &= \mathbf{X}^T (\mathbf{y} - \mathbf{p})\end{aligned}$$

where $\mathbf{p}$ is a $n \times 1$ column vector,

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} \frac{e^{\boldsymbol{\beta}^T \mathbf{x_1}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_1}}} \\ \frac{e^{\boldsymbol{\beta}^T \mathbf{x_2}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_2}}} \\ \vdots \\ \frac{e^{\boldsymbol{\beta}^T \mathbf{x_n}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x_n}}} \end{bmatrix}.$$

The last step of the new form is correct since matrix vector multiplication leads to the weighted sum of the column vectors in the matrix.

Then, second-order partial derivative (a.k.a the Hessian matrix) is calculated as

$$\begin{aligned}
\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^T} &= \frac{\partial \mathbf{X}^T(\mathbf{y}-\mathbf{p})}{\partial\beta^T} \\
&= -\mathbf{X}^T \frac{\partial \mathbf{p}}{\partial\beta^T} (\text{since } \mathbf{y} \text{ is a constant}) \\
&= -\mathbf{X}^T \begin{bmatrix} \partial(\frac{e^{\boldsymbol{\beta}^T\mathbf{x_1}}}{1+e^{\boldsymbol{\beta}^T\mathbf{x_1}}})/\partial\boldsymbol{\beta}^T \\ \partial(\frac{e^{\boldsymbol{\beta}^T\mathbf{x_2}}}{1+e^{\boldsymbol{\beta}^T\mathbf{x_2}}})/\partial\boldsymbol{\beta}^T \\ \vdots \\ \partial(\frac{e^{\boldsymbol{\beta}^T\mathbf{x_n}}}{1+e^{\boldsymbol{\beta}^T\mathbf{x_n}}})/\partial\boldsymbol{\beta}^T \end{bmatrix} \\
&= -\mathbf{X^T} \begin{bmatrix} (1+e^{\boldsymbol{\beta}^T\mathbf{x_1}})^{-2} e^{\boldsymbol{\beta}^T\mathbf{x_1}} \cdot \mathbf{x_1}^T \\ (1+e^{\boldsymbol{\beta}^T\mathbf{x_2}})^{-2} e^{\boldsymbol{\beta}^T\mathbf{x_2}} \cdot \mathbf{x_1}^T \\ \vdots \\ (1+e^{\boldsymbol{\beta}^T\mathbf{x_n}})^{-2} e^{\boldsymbol{\beta}^T\mathbf{x_n}} \cdot \mathbf{x_n}^T \end{bmatrix} \\
&= -\mathbf{X^T} \begin{bmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & p_n(1-p_n) \end{bmatrix} \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \vdots \\ \mathbf{x_n}^T \end{bmatrix} \\
&= -\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X}
\end{aligned}$$

where $\mathbf{W}$ is a $n \times n$ diagonal matrix written as

$$\mathbf{W} = \begin{bmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & p_n(1-p_n) \end{bmatrix}.$$

$\mathbf{W}$ is formed in this way since $(1+e^{\boldsymbol{\beta}^T\mathbf{x_1}})^{-2} e^{\boldsymbol{\beta}^T\mathbf{x_1}} = p(1-p)$ if $p = \frac{e^{\boldsymbol{\beta}^T\mathbf{x_1}}}{1+e^{\boldsymbol{\beta}^T\mathbf{x_1}}}$.
Putting everything together the updating step of the Newton-Raphson method is

$$\begin{aligned}
\beta^{\text{new}} &= \beta^{\text{old}} - \left[\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^T}\right]^{-1} \frac{\partial \ell_n(\beta)}{\partial\boldsymbol{\beta}} \\
&= \beta^{\text{old}} + (\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T(\mathbf{y}-\mathbf{p}).
\end{aligned} \tag{5.1.2}$$

The corresponding algorithm is hence.

5.2 Connection to Linear Regression

There is an interesting connection between linear regression and linear classification. Below we show that applying the Newton-Raphson method, the way that we solve the linear classification problem can be viewed as solving a sequence of linear regression problems. Before we do that we first study a problem of weighted linear regression problem.

Given the training data: a n by p object-feature matrix $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}]^T$, a vector of labels $\mathbf{y}$, and a vector $\mathbf{w} \in \mathbb{R}^p$ where $w_i > 0$, the weighted linear regression problem is to find a linear function that minimizes the weighted loss function, or

$$\hat{\beta} = \arg\min_{\beta} \sum w_i \cdot L(y_i, \mathbf{x_i}^T \beta) \tag{5.2.1}$$

We often the squared loss where $L(y, \hat{y}) = (y - \hat{y})^2$.
Converting the equation to an equivalent matrix format we have

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \sum w_i \cdot L(y_i, \mathbf{x_i}^T \beta) \\ &= \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta), \end{aligned} \tag{5.2.2}$$

where $\mathbf{W}$ is

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & w_n \end{bmatrix}.$$

The solution can be easily obtained using a similar approach that we performed in linear regression. We expand the equation (5.2.2), take the gradient, set it zero, and we have

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta) \\ &= \arg\min_{\beta} \beta^T (\mathbf{X}^T \mathbf{W} \mathbf{X}) \beta - 2\beta \mathbf{X}^T \mathbf{W} \mathbf{y} \\ &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \end{aligned} \tag{5.2.3}$$

Now we return to the linear regression problem that we studied before. We have

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{n} w_i (y_i - \mathbf{x_i}^T \beta)^2 \\ &= \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Again $\mathbf{W}$ is a diagonal matrix where $\mathbf{W}_{i,i} = w_i$.
We rewrite the equation (5.1.2) in the following way:

$$\begin{aligned} \beta^{\text{new}} &= \beta^{\text{old}} - \left[\frac{\partial^2 \ell_n(\beta)}{\partial \beta \partial \beta^T} \right]^{-1} \frac{\partial \ell_n(\beta)}{\partial \boldsymbol{\beta}} \\ &= \beta^{\text{old}} + (\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T (\mathbf{y} - \mathbf{p}) \\ &= (\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \cdot (\mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}((\mathbf{y} - \mathbf{p})) \\ &= (\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}, \end{aligned} \tag{5.2.4}$$

where $\mathbf{z} = \mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}((\mathbf{y} - \mathbf{p})$.
With the equation rewritten, we see that the solution of linear classification is also the solution of the related weighted linear regression problem, as specified below:

$$\beta^{\text{new}} = \arg\min_{\beta} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$$

This connection helps in deriving penalized linear classification problems in the exercises.

5.3 Problem of Logistic Regression

5.4 Bibliographic Notes and Further Reading

For the Bayesian treatment of logistic regression, see Chapter 4 of [2].

5.5 Exercises

5.1[Probit Regression] Probit regression is similar to logistic regression as a way to perform linear classification. Different from logistic regression, Probit regression uses a different probabilistic model that involves the cumulative distribution function of a standard Gaussian. The Probit regression is specified as

$$\mathbb{P}(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = F(\boldsymbol{\beta}^T\mathbf{x}_i)$$
$$\mathbb{P}(y_i = 0|\mathbf{x}_i, \boldsymbol{\beta}) = 1 - F(\boldsymbol{\beta}^T\mathbf{x}_i)$$

where $0 \leq F(a) \leq 1$ is the cumulative distribution function of a distribution. Typically we use the standard Gaussian so that $F(a) = \int_{-\infty}^{a} \mathcal{N}(x|0, 1)\ dx$.
Show that Probit regression is a linear classification model and derive the decision boundary. Develop the maximum likelihood estimator of $\boldsymbol{\beta}$ using an iterative re-weighting algorithm.

5.2[Penalized Logistic Regression] Consider the following objective function

$$\hat{\beta} = \arg\min_{\beta} -\ell_n(\beta) - \lambda\beta^T\beta.$$

Comparing to (5.1.1), we make a couple of changes. First, we change the maximization of the log likelihood function to minimization of the negation of the same value. Second we add an additional item of the L2 norm of β with a coefficient $\lambda > 0$. This is known as a penalized logistic regression problem. Derive the solution of this problem using the Newton-Raphson method.

Chapter 6

Consistency of Maximum Likelihood Estimation

In the chapter of maximum likelihood estimation we introduce a very basic property of MLE. MLE is consistent, which gives us confidence of utilizing MLE in modeling. However our proof is developed in a rather specialized situation where we only consider a finite number of possible parameters in the search for the best parameter. In general we often consider all possible parameter values in a n-dimensional space where the total number of elements is clearly infinite. Extending the search from a finite space to an infinite space and providing certain theoretic guarantee are of core value in the study of MLE, which is the aim of this chapter.
As the first step, we study the uniform strong law of large numbers.

6.1 Uniform Laws of Large Numbers

Laws of large numbers tell us that the sample mean of a statistics converges to the population mean of the statistics. In studying MLE we see that in applying the laws of large numbers, a commonly encountered case is that we search through a parameter space and we want to ensure that we still have the nice property that sample mean of a statistics converges to the population mean of the statistics when we systematically exam a parameter space. To formalize the discussion we revisit the familiar concept of laws of large numbers with two key differences. First, the statistics that we are interested in parametrized by a parameter θ. Second θ is an element in the parameter space Θ. Below we introduce the concept of uniform laws of large numbers.

Definition 6.1 (Uniform Laws of Large Numbrers)**.** Given n *i.i.d.* random variables $X = X_1, X_2, \ldots, X_n$ with the distribution function $F(x)$, let $U(x, \theta)$ be a function and $\theta \in \Theta$ is an element in a parameter space Θ. We are interested in the following statistics

$$U(X, \theta) = \frac{1}{n} \sum_{i=1}^{n} U(x_i, \theta),$$

which we assume the statistics exists and is finite for all $\theta \in \Theta$. In addition we set

$$\mu(\theta) = \mathbb{E}\left[U(X, \theta)\right] = \int U(x, \theta)\, dF(x). \tag{6.1.1}$$

By the weak law of large number we have

$$\frac{1}{n}\sum_{i=1}^{n} U(X_i,\theta) \xrightarrow{\text{P}} \mu(\theta). \tag{6.1.2}$$

Given a sample $\mathbf{x} = x_1, x_2, \ldots, x_n$ of X we define $U_s(\mathbf{x})$ as

$$U_s(\mathbf{x}) = \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} U(x_i,\theta) - \mu(\theta)\right| \tag{6.1.3}$$

$U_s(X)$ is a random variable. We say that $\frac{1}{n}\sum_{i=1}^{n} U(X_i,\theta)$ is *uniformly convergent in probability* to $\mu(\theta)$ if $U_s(X)$ converges in probability to 0. In addition $\frac{1}{n}\sum_{i=1}^{n} U(X_i,\theta)$ is *uniformly convergent almost surely* to $\mu(\theta)$ if $U_s(X)$ converges almost surely to 0.
Equivalently $\frac{1}{n}\sum_{i=1}^{n} U(X_i,\theta)$ is uniformly convergent in probability to $\mu(\theta)$ if

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} U(x_i,\theta) - \mu(\theta)\right| \xrightarrow{\text{P}} 0, \tag{6.1.4}$$

and $\frac{1}{n}\sum_{i=1}^{n} U(X_i,\theta)$ is uniformly convergent almost surely to $\mu(\theta)$ if

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} U(x_i,\theta) - \mu(\theta)\right| \xrightarrow{\text{a.s.}} 0. \tag{6.1.5}$$

Since (with mild assumptions) almost surely convergence implies convergence in probability we focus on uniform convergence almost surely below.

Example 6.1. Given a sample $\mathbf{x} = x_1, x_2, \ldots, x_n$ of n *i.i.d.* random variables $X = X_1, X_2, \ldots, X_n$, let $M_n(X,\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i;\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i;\theta)$ for $\theta \in \Theta$. If Θ is finite by the probability union bound, we have $M_n(X,\theta)$ converges uniformly in probability to $M(\theta)$ and $M_n(X,\theta)$ converges uniformly almost surely to $M(\theta)$.

The following theorem, mainly adopted from [3], state a sufficient condition to guarantee the uniform convergence almost surely.

Theorem 6.1 (Uniform Strong Law of Large Numbers)**.** *If we have*

- Θ *is compact,*
- $U(X_i,\theta)$ *is continuous for all* x *and all* $\theta \in \Theta$*,*
- *There exists a function* $K(x)$ *such that* $\mathbb{E}[K(x)] < \infty$ *and* $|U(x,\theta)| \le K(x)$ *for all* x *and* θ*.*

Then

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{1}^{n} U(X_j,\theta) - \mu(\theta)\right| = 0\Big) = 1. \tag{6.1.6}$$

By the definition of convergence almost surely, the conclusion of Theorem (6.1) can be stated as $\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} U(x_i,\theta) - \mu(\theta)\right| \xrightarrow{\text{a.s.}} 0$. Since convergence almost surely implies convergence in probability, we have $\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} U(x_i,\theta) - \mu(\theta)\right| \xrightarrow{\text{P}} 0$.
This is one of a few theorems that we do not provide a proof immediately. We delay the proof after we discuss the strong law of large numbers in Appendix. This provides the necessary background for a strong statement regarding the consistency of MLE.

6.2 Consistency of Maximum-Likelihhood Estimates

The the previous discussion, we give a sufficient condition to guarantee consistency of MLE.

Theorem 6.2 (Consistency of Maximum-Likelihood Estimates)**.** *If we have*

- Θ *is compact,*
- $\ell(x|\theta)$ *is continous in* θ *for all* x*,*
- *There exists a function* $K(x)$ *such that* $\mathbb{E}[K(x)] < \infty$ *and* $|\ell(()x, \theta)| \leq K(x)$ *for all* x *and* θ*,*
- $M(\theta) = \int \ell(x|\theta) f(x|\theta_0)\, dx$ *is continous,*
- *the distribution of* $f(x|\theta)$ *is identifiable.*

Then

$$\widehat{\theta}_{\text{MLE}} \xrightarrow{\text{P}} \theta_0. \tag{6.2.1}$$

Here $\ell(x_i|\theta)$ is the log likelihood of x. $\widehat{\theta}_{\text{MLE}}$ is the maximum likelihood estimation of θ is n samples. To prove this theorem, we first show a simple (yet very useful) property of $M(\widehat{\theta}_{\text{MLE}})$.

Theorem 6.3. *With the previous stated conditions we have* $M(\widehat{\theta}_{\text{MLE}})$ *converges to* $M(\theta_0)$ *in probability or*

$$M(\widehat{\theta}_{\text{MLE}}) \xrightarrow{\text{P}} M(\theta_0)$$

Proof. With the first three conditions of Theorem (6.2), we have $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{P}} 0$. By definition MLE maximize the likelihood function for the n samples. So we have $M_n(\widehat{\theta}_{\text{MLE}}) \geq M_n(\theta_0)$.
As proved in Theorem (3.1) for identifiable distributions as a global property of MLE we have $M(\theta_0) > M(\theta)$.
Combining the two statements, we have $M(\theta_0) \geq M(\widehat{\theta}_{\text{MLE}})$.
Alternatively we have

$$\begin{aligned} M(\theta_0) - M(\widehat{\theta}_{\text{MLE}}) &= M(\theta_0) - M_n(\theta_0) + M_n(\theta_0) - M(\widehat{\theta}_{\text{MLE}}) \\ &\leq M(\theta_0) - M_n(\theta_0) + M_n(\widehat{\theta}_{\text{MLE}}) - M(\widehat{\theta}_{\text{MLE}}) \\ &\leq M(\theta_0) - M_n(\theta_0) + \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)|\,. \end{aligned}$$

□

Clearly $M(\theta_0) - M_n(\theta_0) \xrightarrow{\text{P}} 0$ and $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{P}} 0$ and hence we have $M(\theta_0) - M(\widehat{\theta}_{\text{MLE}}) \xrightarrow{\text{P}} 0$, or $M(\widehat{\theta}_{\text{MLE}}) \xrightarrow{\text{P}} M(\theta_0)$.
If we have $M(\widehat{\theta}_{\text{MLE}}) \xrightarrow{\text{P}} M(\theta_0)$, do we automatically have $\widehat{\theta}_{\text{MLE}} \xrightarrow{\text{P}} \theta_0$ for any function M? The answer is no. We show a figure to illustrate the case below.
However since M is continuous in a compact set Θ and $M(\theta_0)$ is the only global maximum (by identifiability of the distribution family) we conclude that $M(\widehat{\theta}_{\text{MLE}}) \xrightarrow{\text{P}} M(\theta_0)$.
With similar conditions we can prove a stronger version of the consistency, i.e. $\widehat{\theta}_{\text{MLE}}$ converges almost surely to θ_0. Interested reader may see [3], chapter 17 for details.

6.3 Bibliographic Notes and Further Reading

6.4 Exercises

Part III

Learning with Penalized Likelihood Estimation

Chapter 7

Akaike Information Criterion(AIC)

7.1 Intuition of Why MLE may not be Optimal.

Let us perform a thought experiment. Let's consider a linear regression model where $y = \beta^T \mathbf{x}$. MLE does not work because it considers all features equally, with no guard against the meaningless or "garbage" features.

7.2 AIC

Claim 7.1. *$E[L_n(\hat{\theta}) - \hat{L_n}(\hat{\theta})] = \frac{-p^*}{n}$, with $p^* = 1$.*

Chapter 8

Ridge Regression

Being consistent does not guarantee that linear regression always produces "stable" parameter estimation in real life. Here stable means that with small amount of noises introduced in Equation (), the estimation $\hat{\beta}_{LR}$ does not change much. There are many reasons that a consistent estimator may not be stable. Below for linear regression we provide justification of the behavior using singular value decomposition.

8.1 Unstable Parameter Estimation in Linear Regression

Now we introduce the Singular Value Decomposition (SVD) of $\mathbf{X}$ where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are two unitary matrices, i.e. $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{I}$. $\mathbf{U}$ is an $n \times n$ matrix spanning the column space of $\mathbf{X}$, while $\mathbf{V}$ is the $p \times p$ matrix spanning the row space of $\mathbf{X}$. $\mathbf{\Sigma}$ is an $n \times p$ matrix containing p singular values of $\mathbf{X}$. It has the form that: the upper p rows is a diagonal matrix, where the diagonal elements are p sigular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ in decreasing order. The rest $n - p$ rows are all zero elements. Below we show $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^T\mathbf{\Sigma}$. We use these two matrices in the subsequent discussion:

$$\Sigma_{n\times p} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_p \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \qquad \Sigma^T\Sigma = \begin{bmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_p^2 \end{bmatrix}, \qquad (\Sigma^T\Sigma)^{-1} = \begin{bmatrix} \frac{1}{\lambda_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2^2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{\lambda_p^2} \end{bmatrix}.$$

Straightforward calculation shows that

$$\begin{aligned} \mathbf{X}^T\mathbf{X} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= (\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T)\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T, \\ (\mathbf{X}^T\mathbf{X})^{-1} &= \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})^{-1}\mathbf{V}^T, and \\ (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T &= \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^T\mathbf{U}^T. \end{aligned}$$

Hence we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{LR} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^T\mathbf{U}^T\mathbf{y}.\end{aligned}$$

Finally we have

$$\begin{aligned}\hat{\beta}_{LR} &= \mathbf{V}\begin{bmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_p} & 0 & \cdots & 0 \end{bmatrix}_{p\times n} \begin{bmatrix} \mathbf{u_1}^T\mathbf{y} \\ \mathbf{u_2}^T\mathbf{y} \\ \vdots \\ \mathbf{u_n}^T\mathbf{y} \end{bmatrix}_{n\times 1} \\ &= \mathbf{V}\cdot\frac{1}{\lambda_1}\begin{bmatrix} \mathbf{u_1}^T\mathbf{y} \\ \frac{\lambda_1}{\lambda_2}\mathbf{u_2}^T\mathbf{y} \\ \vdots \\ \frac{\lambda_1}{\lambda_p}\mathbf{u_p}^T\mathbf{y} \end{bmatrix},\end{aligned}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_p$ are p singular values in decreasing order. $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_n}]$.
Problem: Since the singular values are arranged in decreasing order, sometimes λ_p is significantly smaller than λ_1. The large ratio $\frac{\lambda_1}{\lambda_p}$ (referred as condition number of a matrix) means a large variance for that component. This is a challenging case for β estimation. Below we introduce ridge regression, a slightly changed form of linear regression with least square, to address the problem.

8.2 Ridge Regression

Define

$$\begin{aligned}\widehat{\beta}_{\mathrm{R}} &= \arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda\,||\boldsymbol{\beta}||_2^2 \\ &= \arg\min_{\beta} \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta - 2\mathbf{y}^T\mathbf{X}\cdot\beta + \mathbf{y}^T\mathbf{y},\end{aligned} \tag{8.2.1}$$

Solving the previous equation we have

$$\begin{aligned}\widehat{\beta}_{\mathrm{R}} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \text{and} \\ \hat{\mathbf{y}}_R &= \mathbf{X}\hat{\beta_R} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned} \tag{8.2.2}$$

Here λ is a Lagrangian coefficient that is usually set by cross-validation. We obtain this result either from directly differentiate with β, or substitute β with $\beta + \Delta\beta$ to derive. The most intuitive way is to construct the estimator from the general linear regression solution.
Straightforward calculation shows that

$$\begin{aligned}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} &= \mathbf{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\mathbf{V}^T + \lambda\mathbf{V}\mathbf{V}^T, \\ (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} &= \mathbf{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{V}^T, \textit{and} \\ (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T &= \mathbf{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} + \lambda\mathbf{I})^{-1}\boldsymbol{\Sigma}^T\mathbf{U}^T.\end{aligned}$$

Similar to what we did in the previous section, we introduce the singular value decomposition of $\mathbf{X}$. Now the estimator $\hat{\beta}_R$ can be expressed as

$$\hat{\beta} = \mathbf{V} \begin{bmatrix} \frac{\lambda_1}{\lambda_1^2+\lambda} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2}{\lambda_2^2+\lambda} & \cdots & 0 & 0 & \cdots & 0 \\ & & & & & & \\ 0 & 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \frac{\lambda_p}{\lambda_p^2+\lambda} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u_1}^T\mathbf{y} \\ \mathbf{u_2}^T\mathbf{y} \\ \vdots \\ \mathbf{u_n}^T\mathbf{y} \end{bmatrix}.$$

8.3 Incoporating Non-linear Relationship with Transformations

Let us revisit the linear regression model that we studied in Chapter (4). If we have a quadratic relationship between y and x, say for example $y = x^2$, how do we perform curve fitting? It turns out linear regression and ridge regression is still applicable in this case. Consider the following transformation $\phi : \mathbb{R} \to \mathbb{R}^3$ where $\phi(x) = (x^2, x, 1)^T$. We put the number 1 here to fit the linear regression model with possible offset.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1} \qquad \mathbf{X} = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix}_{n\times 3} \qquad f(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}.$$

Below we work on the general case where each sample $\mathbf{x}_i$ is a p dimensional vector. The data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ is an n by p matrix with n samples. Considering a transformation $\phi : \mathbb{R}^p \to \mathbb{R}^d$ mapping a p-dimensional vector to a $d \geq p$ dimensional vector with the possibility of d be infinite. Let $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]^T_{n\times d}$ be the data matrix after transformation.

$$\begin{aligned} \widehat{\beta}_{\mathrm{R}} &= \arg\min_{\beta} ||\mathbf{y} - \phi(\mathbf{X})\boldsymbol{\beta}||_2^2 + \lambda\,||\boldsymbol{\beta}||_2^2 \\ &= (\phi(\mathbf{X})^T\phi(\mathbf{X}) + \lambda\mathbf{I})^{-1}\phi(\mathbf{X})^T\mathbf{y}. \end{aligned} \tag{8.3.1}$$

8.4 Incoporating Non-linear Relationship with Kernerl Matrix

(8.3.1) provides a convenient approach to incorporate non-linear relationship to regression [1]. One related problem is that we may decide to transform the data to a very high dimensional space where $d \gg n$ and hence the matrix $\phi(X)^T\phi(X)$ is of size $d \times d$ which is hard to deal with. Below we present a simple trick to deal with the problem. The trick, known as the kernel trick, turns out to be a very useful one and is widely used in data analysis.

We notice that $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$. To see this we noticed that $A^{-1}B = CD^{-1}$ if and only if $BD = AC$ if both A and D are inversible. After simple algebra operations we see the relationship.

From this claim, we may have

$$\begin{aligned} \widehat{\beta}_{\mathrm{R}} &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y} \\ \widehat{\mathbf{y}}_{\mathrm{R}} &= (\mathbf{X}\mathbf{X}^T)(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{y} \\ &= \mathbf{K}\cdot(\mathbf{K} + \lambda\mathbf{I})^{-1}\cdot\mathbf{y}, \end{aligned}$$

[1] We will see a similar discussion in Support Vector Machines using transformations for classification

where $K = \mathbf{X} \cdot \mathbf{X}^T$ is the *kernel matrix*.
Below we show an approach to derive the solution rather than relying on the mathematical trick.

$$\widehat{f} = \underset{f \text{ is linear}}{\arg\min} (\sum_i (y_i - f(\phi(\mathbf{x_i})))^2 + \lambda \, ||f||_2^2).$$

As we studied before, any linear functional is presented by a vector β,

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\arg\min} (\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \phi(\mathbf{x_i}))^2 + \lambda \, ||\boldsymbol{\beta}||_2^2)$$

Claim: $\boldsymbol{\beta} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x}_i)$. Let $\phi(\mathbf{X}) = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_n)]^T$ and $\alpha = (\alpha_1, \cdots, \alpha_n)^T$. We have $\beta = \phi(\mathbf{X})^T \boldsymbol{\alpha}$ and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \phi(\mathbf{X})\beta = \phi(\mathbf{X})\phi(\mathbf{X})^T\alpha = \begin{bmatrix} \phi(\mathbf{x_1})^T\phi(\mathbf{x_1}) & \phi(\mathbf{x_1})^T\phi(\mathbf{x_2}) & \cdots & \phi(\mathbf{x_1})^T\phi(\mathbf{x_n}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\mathbf{x_n})^T\phi(\mathbf{x_1}) & \cdots & \cdots & \phi(\mathbf{x_n})^T\phi(\mathbf{x_n}) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}.$$

Thus, the solution to this optimization problem is

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= \underset{\boldsymbol{\alpha}}{\arg\min} \, ||\mathbf{y} - \mathbf{K} \cdot \alpha||_2^2 + \lambda \alpha^T \mathbf{K} \alpha \\
&= (\mathbf{K}^T\mathbf{K} + \lambda\mathbf{K})^{-1}\mathbf{K}^T \cdot \mathbf{y} \\
&= (\mathbf{K}(\mathbf{K} + \lambda\mathbf{I}))^{-1}\mathbf{K} \cdot \mathbf{y} \\
&= (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \quad \text{where } \mathbf{X} \cdot \mathbf{X}^T = \mathbf{K}, \\
\widehat{\boldsymbol{\beta}} &= \phi(\mathbf{X})^T\widehat{\boldsymbol{\alpha}} = \phi(\mathbf{X})^T(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \\
\widehat{y} &= \phi(\mathbf{X})\widehat{\boldsymbol{\beta}} = \phi(\mathbf{X})\phi(\mathbf{X})^T(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \\
&= \mathbf{K} \cdot (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}
\end{aligned}$$

This indicates that, ridge regression is just a special case of kernel regression.

8.5 Incoporating Non-linear Relationship with Function Basis

Chapter 9

Ensemble Learning

9.1 Weak classifier

Definition: One whose error rate is slightly better than random guess. (We have many weak classifiers!)

$$G(x) = \text{sign}(\sum_{m=1}^{M} \alpha_m G_m(X))$$

9.2 Ada boost

1. Procedure
(1) Initialize the observation weight $w_i = \frac{1}{N}, i = 1, 2, \cdots, N$.
(2) For $m = 1$ to k, $k < N$

- Fit a classifier $G_m(x)$ using weight w_i on the training data
- Compute the error as $\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(X))}{\sum_{i=1}^{N} w_i}$
- $\alpha_m \longleftarrow \log \frac{1-\text{err}_m}{\text{err}_m}$
- $w_i \longleftarrow w_i \cdot \exp(\alpha_m I(y_i \neq G_m(X)))$

(3) Output the classifier sign $G(X) = \text{sign}(\sum_{m=1}^{k} \alpha_m G_m(X))$.

2. Thinking in another way, $f(x) = \sum_i m_i G_i(X)$. We use exponential loss $L(y, f(x)) = \exp(-yf(x))$. (This makes sense, the sign for y and $f(x)$ should be the same, otherwise the performance will be very bad.)

$$\begin{aligned}(\beta_m, G_m) &= \underset{\beta, G}{argmin} \sum_{i=1}^{N} \exp[-y_i f_{m-1}(x_i)] \exp[-y_i \beta G(x_i)] \\ &= \underset{\beta, G}{argmin} \sum_{i=1}^{N} w_i^{(m)} \exp[-\beta y_i G(x_i)] \qquad \cdots\cdots \mathbf{I}\end{aligned}$$

where $w_i^{(m)} = \exp(-y_i f_{m-1}(x_i))$. Observation for this equation is that β and G can be obtained separately, they are independent.

9.3 Alternative approach

Transformed back to linear regression with penalized term, we can rewrite the optimization problem as

$$\beta = \underset{\beta,G}{argmin} L(y, G\beta) + \lambda||\beta||_1$$

where the loss function is still exponential loss $L(y, G\beta) = \exp(-y_i \cdot G_i\beta)$.

$$L(\beta, G) = \sum_{i=1}^{N} w_i \exp(-\beta) + [\exp(\beta) - \exp(-\beta)] \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i))$$

Best G is the one that satisfy

$$G_m = \underset{G}{argmin} \sum_{i=1}^{N} w_i^{(m)} I(y_i \neq G(x_i))$$

Then take derivative with respect to β, we can get

$$[\exp(\beta) + \exp(-\beta)] \cdot (err) = e^{-\beta} \Rightarrow e^{2\beta} = \frac{1 - (err)}{(err)} \Rightarrow \beta_m = \frac{1}{2} \ln \frac{1 - (err)}{(err)}$$

Then the iterative procedure can be implemented as

$$f_m(x) = f_{m-1}(x) + \beta_m G_m(x)$$
$$w_i^m = w_i^{(m)} e^{-\beta_m y_i G_m(x)}$$

The sign function indicates $-y_i G_m(x) = 2I(y_i \neq G_m(x_i)) - 1$, which meets with our expectation. When y_i and $G_m(x)$ take different signs, the value of left hand side is 1, which is equal to the right hand side. Similar situation to same signs.

Detailed content can be referenced in
`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.7436&rep=rep1&type=pdf`.

Part IV

Learning without Knowing Distributions

Chapter 10

Support Vector Machines

10.1 Introduction

In the previous chapters we introduced the concepts of maximum likelihood estimator and penalized maximum likelihood estimator. We now turn our attention to a group of new concepts that are commonly related to the development of a very popular machine learning algorithm called the Support Vector Machines (SVMs). In the development of machine learning algorithms, SVMs have a unique position. First in SVM we do not rely on any knowledge (or assumptions) of how data are generated nor do we assume any distributions of predictive models. In this way we perform *agnoistic learning*, meaning learning without using generative models. Second SVMs project data to a high, or often infinite, dimensional space. This practice contradicts to the insights that we gained through previous chapters where high dimensional space suggests high model complexity. Third SVMs has a quite different approach to handle non-linear relationship between features and class, which is commonly known as the "kernel trick".
We talk about Redamache complexity because it is easy to compute, it is at least as tight as VC-dimension, it takes distribution as input.
Below we first introduce the concept of large margin classification in the feature space. We then transform large margin classifiers to a more familiar format of minimizing loss function together with regularization. We also demonstrate what the "kernel trick" is and finally present support vector machines in the functional format.

10.2 Large Margin Classifier

Let us consider a very simple classification rule with a linear separating hyperplane. The rule states that if the point fall into the "right" side of the hyperplane, we label it positive and negative otherwise. Equivalently given $\mathbf{x}$ and $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$, the linear decision function is:

$$\begin{aligned} G(\mathbf{x}) &= \text{sgn}[f(\mathbf{x})] \\ &= \text{sgn}[\mathbf{x}^T\boldsymbol{\beta}] \\ &= \begin{cases} 1, & \text{if } \mathbf{x}^T\boldsymbol{\beta} \geq 0 \\ -1, & \text{otherwise.} \end{cases} \end{aligned} \tag{10.2.1}$$

Considering a quite usual case where you may find a hyperplane to "perfectly" separate positive cases and negative cases in such way that there is no mistake. One observation is that if such a hyperplane exists, there may be an infinite number of them (by turning a separating hyperplane

slightly you probably have another one that perfectly separates the data set). Now our concern is which β should we choose? One school of thinking is to select the hyperplane that have the largest margin where the margin is the minimal distance between positive cases and negative cases to the separating hyperplane.
We formalize the large margin classifier using the equation

$$\begin{aligned} &\max_{||\boldsymbol{\beta}||_2^2=1} m \\ &\text{such that } y_i(\mathbf{x_i}^T\beta) \geq m \quad \text{for all } i \in [1, N], \end{aligned} \tag{10.2.2}$$

where N is the number of samples. There is a reason why we add the constraint of $||\boldsymbol{\beta}||_2^2 = 1$. Without the constraint we could simply increase the marge m by scaling $\boldsymbol{\beta}$. The variable m is called a slack variable in optimization. Its only role is to be a place holder so that we could select the minimal one among the distances between cases to the separating hyperplane. The concept of maximizing a minimal metric is recurring. We will see a similar concept called the minimax estimator in a later chapter.
(10.2.2) is equivalent to a slightly simple format,

$$\begin{aligned} &\min ||\boldsymbol{\beta}||_2^2 \\ &\text{such that } y_i(\mathbf{x_i}^T\beta) \geq 1 \quad \text{for all } i \in [1, N]. \end{aligned} \tag{10.2.3}$$

To establish the equivalency we notice that if $\hat{\beta}_0$, $\hat{m_0}$ is the solution for (10.2.2), $\hat{\beta}_1 = \frac{\hat{\beta}_0}{\hat{m_0}}$ is the solution for (10.2.3). In addition if $\hat{\beta}_1$ is the solution for (10.2.3), $m_0 = \left|\left|\hat{\beta}_1\right|\right|_2^2$ and $\hat{\beta}_0 = \frac{\hat{\beta}_1}{||\hat{\beta}_1||_2^2}$ is the solution for (10.2.2).
We only prove the first observation. The second one could be proved similarly. To prove the first observation we use proof by contradiction. Given that $\hat{\beta}_0$, $\hat{m_0}$ is the solution of (10.2.2) and $\hat{\beta}_1 = \frac{\hat{\beta}_0}{\hat{m_0}}$, suppose that there exists a $\hat{\beta}_2$ satisfying

$$\begin{aligned} &\left|\left|\hat{\boldsymbol{\beta}}_2\right|\right|_2 < \left|\left|\hat{\boldsymbol{\beta}}_1\right|\right|_2 \quad \text{and} \\ &y_i(\mathbf{x_i}^T\hat{\boldsymbol{\beta}}_2) \geq 1 \text{ for all } i \in [1, N], \end{aligned}$$

we then have

$$y_i(\mathbf{x_i}^T\frac{\hat{\boldsymbol{\beta}}_2}{||\boldsymbol{\beta}_2||_2}) \geq \frac{1}{||\beta_2||_2} > \frac{1}{||\beta_1||_2} = \hat{m_0}.$$

Basically, we find $\beta_3 = \frac{\hat{\beta}_2}{||\beta_2||_2}$, such that

$$\begin{aligned} ||\beta_3||_2^2 = 1 \quad \text{and} \\ y_i(\mathbf{x_i}^T\beta_3) > \hat{m_0} \text{ for all } i \in [1, N]. \end{aligned}$$

The existance of such β_3 contradicts the assumption that $\hat{\beta}_0$, $\hat{m_0}$ is the solution of (10.2.2).
For linear non-separate cases, we may introduce a technique called “soft margin” as specified below:

$$\begin{aligned} &\min ||\boldsymbol{\beta}||_2^2 \\ \text{such that } &y_i(\mathbf{x_i}^T\beta) \geq 1-\xi_i \\ &\xi_i \geq 0, \sum \xi_i \leq \tau_0, \end{aligned}$$

where τ_0 is a hyper-parameter. $1-\xi_i$ is the so-called soft margin. If $\xi_i = 0$ for every i, the data is well separated.
Now a commonly used SVM formula (c-SVM) follows as

$$\begin{aligned} &\min \frac{1}{2}||\beta||_2^2 + c\cdot\sum_1^N \xi_i \\ \text{such that } &y_i(\mathbf{x_i}^T\beta) \geq 1-\xi_i \\ &\xi_i \geq 0, \end{aligned} \tag{10.2.4}$$

where i ranges from 1 to N. $\frac{1}{2}$ is used by convention. We solve the constraint optimization problem (10.2.4) using Lagrange multipliers. The Lagrangian L_p is

$$L_p = \frac{1}{2}||\beta||^2 + c\cdot\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i[y_i(\mathbf{x_i}^T\beta) - (1-\xi_i)] - \sum_{i=1}^N \mu_i\xi_i. \tag{10.2.5}$$

To optimize L_p we take the gradients and set them to zero. We have

$$\begin{aligned} \frac{\partial L_p}{\partial \beta} &= \beta - \sum_{i=1}^N \alpha_i y_i \mathbf{x_i} = 0 \quad \Rightarrow \beta = \sum_{i=1}^N \alpha_i y_i \mathbf{x_i} \\ \frac{\partial L_p}{\partial \xi_i} &= c - \alpha_i - \mu_i = 0 \quad \Rightarrow \mu_i = c - \alpha_i \end{aligned}$$

In addition the KKT (Karush-Kuhn-Tucker) conditions of (10.2.5) are

$$y_i\mathbf{x_i}^T\boldsymbol{\beta} - (1-\xi_i) \geq 0 \tag{10.2.6}$$
$$\xi_i \geq 0 \tag{10.2.7}$$
$$\alpha_i \geq 0 \tag{10.2.8}$$
$$\mu_i \geq 0 \tag{10.2.9}$$
$$\alpha_i(y_i\mathbf{x_i}^T\boldsymbol{\beta} - (1-\xi_i)) = 0 \tag{10.2.10}$$
$$\mu_i\xi_i = 0 \tag{10.2.11}$$

i ranges from 1 to N. (10.2.6) and (10.2.7) are inherited from the original problem. (10.2.8) and (10.2.9) are there because the Lagrangian multipliers for inequality constrains must be non-negative. (10.2.10) and (10.2.11) take care of boundary conditions. We discuss further details of constrained optimization problems in Chapter E.
We substitute the results back to objective function (10.2.5) and obtain the dual problem L_d of (10.2.5). Noticing (10.2.5) is convex in terms of β and ξ, we hence maximize the dual problem.

$$
\begin{aligned}
L_d &= \frac{1}{2}\alpha_i\alpha_j \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \mathbf{x_i}^T\mathbf{x_j} + c\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i y_i \mathbf{x_i}^T \sum_{j=1}^{N}\alpha_j y_j \mathbf{x_j} + \sum_{i=1}^{N}(1-\xi_i)\alpha_i - \sum_{i=1}^{N}\mu_i\xi_i \\
&= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x_i}^T\mathbf{x_j}, \qquad 0 \le \alpha_i \le c.
\end{aligned}
\tag{10.2.12}
$$

Maximizing L_d is achieved using quadratic programming.
With this results we want to classify all training samples to three categories based on their contribution to the decision function $\boldsymbol{\beta}$. We notice that for those training samples $\mathbf{x}_i$ such that $\xi_i > 0$, we have $\mu_i = 0$ (10.2.11), $y_i\mathbf{x_i}^T\boldsymbol{\beta} - (1-\xi_i) = 0$(10.2.6), and $\alpha_i = c$(10.2.6). The results have two interpretations. First for those samples on the "wrong" side of the decision function, their contribution to the decision function is c. Second SVM has a natural way to deal with outliers in that the contribution of a sample to the decision function is capped at c.
If $\xi_i = 0$, following (refeq:svm:kkt6) we have $\mu_i \neq 0$. We have two possible cases. If $y_i\mathbf{x_i}^T\boldsymbol{\beta} - 1 = 0$, we have $0 < \alpha_i < c$. If $y_i\mathbf{x_i}^T\boldsymbol{\beta} - 1 > 0$, we have $\alpha_i = 0$. All those is where $\alpha_i \neq 0$ are called *support vectors*, which contribute to the decision function. All those is where $\alpha_i = 0$ are none support vectors and they have no contribution to the decision function.

10.3 Hinge Loss and Regularized Loss Function Minimization

In the previous section we discussed large margin classifier and the related optimization problems. Here we show the connection of large margin classifier to regularized loss functions.
We first define a new function that is a one-sided linear function. That is

$$
f_+(x) = \begin{cases} x, & \text{if } x \ge 0 \\ 0, & \text{otherwise.} \end{cases}
\tag{10.3.1}
$$

With the function we may find that the following two objective function are actually equivalent,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{SVM}} &= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\boldsymbol{\beta}||_2^2 + c\sum_i \xi_i \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\boldsymbol{\beta}||_2^2 + c\sum_i L_{\text{HL}}(y_i, \mathbf{x_i}, \boldsymbol{\beta}) \\
&= \arg\min_{\boldsymbol{\beta}} L_{\text{HL}}(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2
\end{aligned}
$$

where $L_{\text{HL}}(y_i, \mathbf{x_i}, \boldsymbol{\beta})$ is called the *hinge loss function* as $L_{\text{HL}}(y_i, \mathbf{x_i}, \boldsymbol{\beta}) = f_+(1 - y_i\mathbf{x_i}^T\boldsymbol{\beta})$. $L_{\text{HL}}(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = \sum L_{\text{HL}}(y_i, \mathbf{x_i}, \boldsymbol{\beta})$.
Comparing to ridge regression,

$$
\hat{\boldsymbol{\beta}}_{\text{R}} = \arg\min_{\boldsymbol{\beta}} \; L_{\text{SL}}(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2
$$

where $L_{\text{SL}}(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$. We found that SVM can also be expressed in the generic form of regularized loss function minimization, just replace the squared loss by hinge loss.
Below we revisit logistic regression to show the related loss function for logistic regression.

We have y_i can only take either -1 or 1. Readers may notice that this setting is different from what we specified in Chapter 5.1 where y_i takes value of 0 or 1. We do so to better illustrate the main point of comparing different loss functions. As we did before $\mathbf{x_i}$ denotes a $p \times 1$ vector, the size of $\boldsymbol{\beta}$ is also $p \times 1$. $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$.

$$p(y_i = 1|\mathbf{x_i}) = \frac{e^{\beta^T \mathbf{x_i}}}{1 + e^{\beta^T \mathbf{x_i}}}$$
$$p(y_i = -1|\mathbf{x_i}) = 1 - p(y_i = 1|\mathbf{x_i}, \beta) = \frac{1}{1 + e^{\beta^T \mathbf{x_i}}}$$

Hence, we can write the likelihood function $L(\beta)$ as

$$L(\beta) = \prod_{i=1}^{n} (\frac{e^{\beta^T \mathbf{x_i}}}{1 + e^{\beta^T \mathbf{x_i}}})^{\frac{y_i+1}{2}} (\frac{1}{1 + e^{\beta^T \mathbf{x_i}}})^{\frac{1-y_i}{2}}$$

One step further, the log-likelihood function can be shown to be

$$\begin{aligned} l(\beta) = \log L(\beta) &= \sum_{i=1}^{n} [\frac{y_i + 1}{2} \beta^T \mathbf{x_i} - \frac{y_i + 1}{2} \ln(1 + e^{\beta^T \mathbf{x_i}}) + \frac{y_i - 1}{2} \ln(1 + e^{\beta^T \mathbf{x_i}})] \\ &= \sum_{i=1}^{n} [\frac{y_i + 1}{2} \beta^T \mathbf{x_i} - \ln(1 + e^{\beta^T \mathbf{x_i}})] \\ &= \sum_{i=1}^{n} - \ln(1 + e^{-y_i \mathbf{x_i}^T \beta}) \end{aligned} \tag{10.3.2}$$

The last step of (10.3.2) is true since y_i can take value of either -1 or 1. Interested reader may verify the correctness of the derivation. With this results the logistic regression is formalized as:

$$\hat{\boldsymbol{\beta}}_{\text{LogR}} = \arg\min_{\boldsymbol{\beta}} \sum_i L_{\text{LogitL}}(y_i, \mathbf{x_i}, \boldsymbol{\beta})$$

where $L_{\text{LogitL}}(y_i, \mathbf{x_i}, \boldsymbol{\beta}) = -\ln(1 + e^{-y_i \mathbf{x_i}^T \boldsymbol{\beta}})$ is called the *logit loss*.
In Figure (), we compare hinge loss, logit loss, squared loss, and one more loss function (exponential loss). We discuss exponential loss in Chapter **??**.

10.4 Transformations and the Kernel Trick

Let us consider a transformation function $\phi : \mathbb{R}^p \to \mathbb{R}^{p_2}$ that maps $\mathbf{x}$ to a high dimensional space $\phi(\mathbf{x})$. Revisiting (10.2.4) we have

$$\begin{aligned} &\min \frac{1}{2} ||\boldsymbol{\beta}||_2^2 + c \cdot \sum_1^N \xi_i \\ &\text{s.t.} \quad y_i(\phi(\mathbf{x_i})^T \beta) \geq 1 - \xi_i \quad \text{for all } i \in [1, N] \\ &\xi_i \geq 0 \end{aligned} \tag{10.4.1}$$

Using Lagrangian L_p and we solve the dual problem we have

$$\begin{aligned} L_d &= \frac{1}{2}\alpha_i\alpha_j \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \phi(\mathbf{x_i})^T\phi(\mathbf{x_j}) + c\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i y_i \phi(\mathbf{x_i})^T \sum_{j=1}^{N}\alpha_j y_j \phi(\mathbf{x_j}) + \sum_{i=1}^{N}(1-\xi_i)\alpha_i - \sum_{i=1}^{N}\mu_i\xi_i \\ &= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(\mathbf{x_i})^T\phi(\mathbf{x_j}), \quad 0 \le \alpha_i \le c. \end{aligned} \tag{10.4.2}$$

Comparing (10.2.12 and (10.4.2) we see that the only difference is that instead of evaluating the dot product of $\mathbf{x_i}$ and $\mathbf{x_j}$ in the feature space we evaluate the dot product $\phi(\mathbf{x_i})$ and $\phi(\mathbf{x_j})$ in the transformed space.
When the transformation is implicit where we do not specific $\phi()$ but rather providing a kernel function K evaluating the dot product $\phi(\mathbf{x_i})$ and $\phi(\mathbf{x_j})$, (10.4.2) can be readily changed to

$$\begin{aligned} L_d &= \frac{1}{2}\alpha_i\alpha_j \sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \phi(\mathbf{x_i})^T\phi(\mathbf{x_j}) + c\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i y_i \phi(\mathbf{x_i})^T \sum_{j=1}^{N}\alpha_j y_j \phi(\mathbf{x_j}) + \sum_{i=1}^{N}(1-\xi_i)\alpha_i - \sum_{i=1}^{N}\mu_i\xi_i \\ &= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}), \quad 0 \le \alpha_i \le c. \end{aligned} \tag{10.4.3}$$

As we did the kernel regression we see that the previous discussion can be easily extended to a much general problem of learning a functional in a kernel space. That is

$$\begin{aligned} &\min \frac{1}{2}\,||f||^2 + c\cdot\sum_{1}^{N}\xi_i, \\ &\text{s.t.} \quad y_i f(\phi(\mathbf{x_i})) \ge 1-\xi_i, \\ &\xi_i \ge 0, \quad \text{for all } i \in [1, N]. \end{aligned} \tag{10.4.4}$$

f is a linear functional for $\phi(\mathbf{x})$. Comparing (10.4.4) and (10.2.4) we see that the major difficulty is that we do not have the definition of $||f||$.

10.5 Exercises

10.1[SVM-regression] Support Vector Machines are applicable to regression models. Below we present a formalization to it. Derive the Lagrangian, the primal form and the dual form. Show that the problem is solvable using quadratic programming. The SVM regression is formalized as

$$\min \frac{1}{2}\,||\boldsymbol{\beta}||_2^2 + c\cdot\sum_{1}^{N} L_\epsilon(y_i, \boldsymbol{\beta}^T\mathbf{x}_i)$$

where L_ϵ is the ϵ-sensitive loss function in that $L_\epsilon(y_i, \hat{y}_i) = f_+(|y_i - \hat{y}_i| - \epsilon)$.
Hint: you may introduce two slack variables ξ_i^+ and ξ_i^- for the ϵ-sensitive loss function.

10.2[ν-SVM] A different way to formalize large margin SVM is the ν-SVM, which is formalized as

$$\begin{aligned} &\min \frac{1}{2}\,||\boldsymbol{\beta}||_2^2 - \nu\rho + \frac{1}{N}\sum_1^N \xi_i \\ &\text{s.t.} \quad y_i(\mathbf{x_i}^T\boldsymbol{\beta}) \geq \rho - \xi_i \\ &\xi_i \geq 0, \rho \geq 0 \quad \text{for all } i \in [1, N] \end{aligned}$$

Once you derive the dual format you should find $\sum_1 N\alpha_i \leq \nu$ is a constraint rather than showing up in the objective function. Since the sum of α_i should be less than ν, this provides a more intuitive way to control the support vectors.

Chapter 11

Statistical Learning Theory

As we discussed in the chapter (7) of Akaike Information Criterion, high dimensional data usually leads to overfitting. If we follow that discussion the practice of mapping data to an infinite dimensional space (as we did in the support vector machines) does not sound a wise choice. In fact one of the major advantages of SVM is that using L_2 regularization we manage to prove that the chance of overfitting is quite limited. In addition as shown below we are able to prove the results without any assumption of the distributions of data. Such type of learning theory is called *agnostic learning* meaning that we have approvable bound regardless the distribution of the data. Retrospectively laws of large numbers are perfect examples of agnostic "learning" where we show that the sample mean converges to the population mean with probability (or almost surely) without specifying the distribution of the original data.
Below we show another way to perform agnostic learning, using a model called Probably Approximately Correct (PAC). Although the first glance shows that PAC is quite similar to what we discussed in laws of large numbers. Conceptually there are two important differences: (1) PAC typically does not provide equivalent results to that of converging almost surely and (2) PAC deals with finite number of samples rather than infinite samples as studied in laws of large numbers.

11.1 Probably Approximately Correct (PAC) Learning

Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and n *i.i.d.* random variables on the probability space with finite expectation μ and finite variance σ^2, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. With the Weak Law of Large Numbers we have

$$\mathbb{P}(|\bar{X} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Let $\delta = \frac{\sigma^2}{n\epsilon^2}$. Solving the equation we have $\epsilon = \sqrt{\frac{\sigma^2}{n\delta}}$. Therefore the previous equation is equivalent to the following statement. With probability $1 - \delta$, we have

$$\bar{X} - \sqrt{\frac{\sigma^2}{n\delta}} \leq \mu \leq \bar{X} + \sqrt{\frac{\sigma^2}{n\delta}}. \tag{11.1.1}$$

Although the two equations are equivalent, the first format leads to the familiar weak law of large numbers and the later format leads to a quite different interpretation. In this interpretation we emphasize that we have only finite number of samples. We show the bounds regarding the difference between our estimation and the expected value of a random variable (i.e. the true parameter).

In this chapter our major objective is to proof that with L_2 regularization as we did in SVMs, with probability $1-\delta$, we have

$$\text{Generalization error} \leq \text{Training error} + \frac{2\,||\beta||_2}{n}\sqrt{\text{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{11.1.2}$$

where n is the number of samples, $\mathbf{K}$ denotes the kernel function matrix.

In order to establish PAC bound for SVMs, one starting point can be the uniform strong law of large numbers as we discussed in Chapter (6). However the direct application of the Theorem (6.1) to regression problems and classification problems has many difficulties. The continuity requirement is one of such difficulties. For example in classification where we often use 0-1 loss where $L(y, \hat{y}_\theta(x)) = \mathcal{I}(y, \hat{y}_\theta(x))$ is not continuous. Below we show a way to provide similar claim to that of uniform strong law of large numbers. The concept was originally explored by Vapnik and his colleagues and this particular form of proof was adopted from [8].

11.2 Rademacher Complexity

The Rademacher complexity measures the complexity of a set of functions. This is a rather new concepts. In Chapter (7) we study complexity of modeling algorithms. One conclusion is that we should control the dimensionality of the modeling algorithms. The insights provided there is not applicable to kernel spaces since the potential dimensionality may be infinity.

Definition 11.1 (Empirical Rademacher complexity)**.** $\mathcal{S} = \{x_1, \ldots, x_n\}$ be a fixed sample generated from the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $\mathcal{F}$ be a class of real-valued functions defined on Ω. The *empirical Rademacher complexity* of the set $\mathcal{F}$ with respect to $\mathcal{S}$ is

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right].$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$ are *i.i.d.* uniform discrete random variables taking on values in $\{-1, 1\}$, or *i.i.d.* Rademacher random variables with probability 0.5.

Definition 11.2 (Rademacher complexity)**.** The *Rademacher complexity* is the set $\mathcal{F}$ is

$$\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}\left[\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})\right].$$

At the first glance, the involvement of the Rademacher random variables in the definition of Rademacher complexity is rather questionable. We use the following example to illustrate the role of the Rademacher random variables.

Example 11.1. Let $\mathcal{S} = \{x_1, x_2\}$ where $x_1 = 0$ and $x_2 = 1$ be a sample from a Bernoulli distribution with probability 0.5. Let f_1 maps the set $\{0, 1\}$ to $\mathbb{R}$ in the way that $f_1(0) = 0$ and $f_1(1) = 1$. Let f_2 also maps the set $\{0, 1\}$ to $\mathbb{R}$ in the way that $f_2(0) = 1$ and $f_2(1) = 0$. We have $n = 2$. $\boldsymbol{\sigma} = \{\{1, 1\}, \{1, 0\}, \{0, 1\}, \{0, 0\}\}$ where each element has probability $\frac{1}{4}$. We calculate the empirical Rademacher complexity of $\mathcal{F} = \{f_1, f_2\}$ below.

$$\begin{aligned}\widehat{\mathfrak{R}}_{\{S\}}(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] \\ &= \frac{1}{4}\sup_{f\in\mathcal{F}}\{|f_i(x_1)+f_i(x_2)|\} + \frac{1}{4}\sup_{f\in\mathcal{F}}\{|-f_i(x_1)+f_i(x_2)|\} \\ &+ \frac{1}{4}\sup_{f\in\mathcal{F}}\{|f_i(x_1)-f_i(x_2)|\} + \frac{1}{4}\sup_{f\in\mathcal{F}}\{|-f_i(x_1)-f_i(x_2)|\}.\end{aligned}$$

Since the set $\mathcal{F}$ is finite, the supremum is the maximum of the two elements. Plug in the numbers we obtain that the empirical Rademacher complexity $\widehat{\mathfrak{R}}_{\{S\}}(\mathcal{F})$ is 1. We can then calculate the Rademacher complexity $\widehat{\mathfrak{R}}_n(\mathcal{F})$, which is left as an exercise.
The following example illustrate the connection of Rademacher variables and reproducing kernel spaces.

Example 11.2. Let $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ a fixed sample generated from the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where $\Omega \subset \mathbb{R}^p$ is a set of finite dimensional vectors. Let $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a kernel function. Let $\phi : \mathbb{R}^p \to \mathcal{H}$ such that $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ for all possible i and js. We intensionally do not specify $\mathcal{H}$ nor do we mention the dimensionality of the set. For the time being, all we need to know is that $\mathcal{H}$ is a vector space (or an abstract vector space, see Appendix (B) for details) with an inner product $\langle\, , \rangle$ defined.
Let $\mathcal{F} \subset \{f \mid f : \mathcal{H} \to \mathbb{R}\}$ be a set of bounded linear functions that map elements in $\mathcal{H}$ to real numbers. A function is a bounded linear function with bound M if it is linear and $f(\mathbf{x}) \leq M\,||\mathbf{x}||_2$ for all elements in the domain.
With the set up we calculate an upper bound of the empirical Rademacher complexity of $\mathcal{F}$ below.

$$\begin{aligned}\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\phi(\mathbf{x}_i))\right|\right] \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i M\,||\phi(\mathbf{x}_i)||_2\right|\right] \\ &= \frac{2M}{n}\sum_{i=1}^{n}||\phi(\mathbf{x}_i)||_2 \\ &= \frac{2M}{n}\sum_{i=1}^{n}\sqrt{k(\mathbf{x}_i,\mathbf{x}_i)}.\end{aligned}$$

Since $\sqrt{x+y} \leq \sqrt{x}+\sqrt{y}$, a tighter bound can be obtained by the fact that functions in $\mathcal{F}$ is linear.

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\phi(\mathbf{x}_i))\right|\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n} f(\sum_{i=1}^{n}\sigma_i \phi(\mathbf{x}_i))\right|\right] && (f \text{ is linear}), \\
&\leq \frac{2M}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\left|\sum_{i=1}^{n}\sigma_i\phi(\mathbf{x}_i)\right|\right|_2\right] && (f \text{ is bounded}), \\
&= \frac{2M}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sum_{i,j=1}^{n}\sigma_i\sigma_j\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle\right)^{\frac{1}{2}}\right] && (11.2.1) \\
&\leq \frac{2M}{n}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle\right]\right)^{\frac{1}{2}} \\
&= \frac{2M}{n}\left(\sum_{i=1}^{n}\kappa(\mathbf{x}_i,\mathbf{x}_i)\right)^{\frac{1}{2}} \\
&= \frac{2M}{n}\sqrt{\mathrm{tr}(\mathbf{K})}.
\end{aligned}$$

$\mathbf{K}$ is the kernel matrix where $k_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. In the derivation we use a fact established by Jensen's inequality, which states that for every concave function g, including the square root function, we have $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$.

We now study the property of Rademacher Complexity so that we could use it to provide practical bound for generalization error.

Theorem 11.1. *Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, Let $\mathcal{F}$ be a class of functions that map Ω to $[a, a+1]$. Then with probability at least $1-\delta/2$, we have*

$$\mathfrak{R}_n(\mathcal{F}) \leq \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 2\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{11.2.2}$$

Proof. In order to prove the statement, we use the McDiarmid's theorem, as stated below.

Theorem 11.2 (McDiarmid's Theorem)**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values from $\mathcal{A}$. Suppose that there exist numbers $c_1, \ldots, c_n$ such that the function $f : \mathcal{A}^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i}\left(|f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,\hat{x}_i,\ldots,x_n)|\right) \leq c_i$$

for all $1 \leq i \leq n$. Then for any $\epsilon > 0$,

$$\mathbb{P}(f(X_1,\ldots,X_n) - \mathbb{E}[f(X_1,\ldots,X_n) \geq \epsilon]) \leq \exp\left\{\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right\}.$$

We delay the proof of the McDiarmid's Theorem to the Appendix.

Since $|x| - |y| \leq |x - y|$, given a $\boldsymbol{\sigma}$, for each g, we have

$$\begin{aligned}\left|\frac{2}{n}\sum_{\omega\in S}\sigma_i g(\omega)\right| - \left|\frac{2}{n}\sum_{\omega\in S'}\sigma_i g(\omega)\right| &\leq \frac{2}{n}\left|\sum_{\omega\in S}\sigma_i g(\omega) - \sum_{\omega\in S'}\sigma_i g(\omega)\right| \\ &= \frac{2}{n}\left|g(z_i) - g(z_i')\right| \\ &\leq \frac{2}{n}.\end{aligned}$$

Given two sets $\mathcal{Q}, \mathcal{Q}'$, if for each element $q \in \mathcal{Q}$, we could find $q' \in \mathcal{Q}'$ such that $q \leq q'$, we have $\sup(\mathcal{Q}) \leq \sup(\mathcal{Q}')$. Applying this simple fact we conclude that

$$\begin{aligned}\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{\omega\in S}\sigma_i f(\omega)\right| - \sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{\omega\in S'}\sigma_i f(\omega)\right| &\leq \frac{2}{n}, \\ \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{\omega\in S}\sigma_i f(\omega)\right|\right] - \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{\omega\in S'}\sigma_i f(\omega)\right|\right] &\leq \frac{2}{n}, \text{ or} \\ \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) - \widehat{\mathfrak{R}}_{\mathcal{S}'}(\mathcal{F}) &\leq \frac{2}{n}.\end{aligned}$$

Similarly we have

$$\widehat{\mathfrak{R}}_{\mathcal{S}'}(\mathcal{F}) - \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) \leq \frac{2}{n}.$$

□

Therefore we conclude that

$$\left|\widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) - \widehat{\mathfrak{R}}_{\mathcal{S}'}(\mathcal{F})\right| \leq c_i,$$

where $c_i = \frac{2}{n}$.
Applying the McDiarmid's theorem, set $\exp(\frac{-2\epsilon^2}{\sum_{i=1} nc_i^2}) = \frac{\delta}{2}$, solving ϵ, we have with probability $1 - \frac{\delta}{2}$

$$\mathfrak{R}_n(\mathcal{F}) \leq \widehat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F}) + 2\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

11.3 Generalization Error Bound with Rademacher Complexity for Bounded Functions

Begin with two different samples that differ only by a single point(in this case z_m), these are given as: $S = (z_1, \ldots, z_{i-1}, z_i, z_{i+1}, z_m)$ and $S' = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, z_m)$ $(1 \leq i \leq m)$.
Define the function Φ as:

$$\Phi(S) = \sup_{h\in\mathcal{H}}(\mathbb{E}\left[h\right] - \mathbb{E}_S\left[h\right])$$

We then have that:

$$\left|\Phi(S') - \Phi(S)\right| \leq \left|\sup_{h\in\mathcal{H}}(\mathbb{E}_S\left[h\right] - \mathbb{E}_{S'}\left[h\right])\right| = \left|\sup_{h\in\mathcal{H}}(\frac{h(z_i) - h(z_i')}{m})\right| \leq \frac{1}{m}$$

By the McDiarmid Inequality, we have with probability $1 - \frac{\delta}{2}$, we have:

$$\Phi(S) \leq \mathbb{E}_s\left[\Phi(S)\right] + \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{11.3.1}$$

We bound the first element at the right side of the previous inequality below.

$$\begin{aligned}
\mathbb{E}_S\left[\Phi(S)\right] &= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(\mathbb{E}\left[h\right]-\mathbb{E}_S\left[h\right])\right] \\
&= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}(\mathbb{E}_{S'}\left[h\right]-\mathbb{E}_S\left[h\right])\right] \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}(\mathbb{E}_{S'}\left[h\right]-\mathbb{E}_S\left[h\right])\right] \\
&= \underset{S,S'}{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}(h(z_i')-g(z_i))\right] \\
&= \underset{\boldsymbol{\sigma},S,S'}{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(z_i')-g(z_i))\right] \\
&\leq \underset{\boldsymbol{\sigma},S'}{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i')\right] - \underset{\boldsymbol{\sigma},S}{E}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\cdot -h(z_i)\right] \\
&= 2\mathbb{E}_{\boldsymbol{\sigma},S}\left[\sup_{h\in\mathcal{H}}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i\cdot h(z_i))\right] \\
&= \mathfrak{R}_n(\mathcal{F}).
\end{aligned}$$

11.4 Generalization Error Bound with Rademacher Complexity for Kernel-Based Hypotheses

There are a few minor technical challenges in order to apply the Theorem (11.4) to derive PAC bound for SVMs. First the loss function must be bounded while hinge loss is not. Second if we use a new loss function, we should be able to calculate its Rademacher Complexity. Fortunately these are remedies for the two points and we present one below.
We define a truncated hinge loss function, which is a continuous approximation of the 0-1 loss function.

Definition 11.3. A truncated hing loss function $L_{\text{THL}}(x)$ with the parameter γ is

$$L_{\text{THL}}(x) = \begin{cases} 0 & \text{if } x > \gamma; \\ 1-\frac{x}{\gamma} & \text{if } \gamma \geq x \geq 0; \\ 1 & \text{other wise.} \end{cases}$$

Clearly the loss function $L_{\text{THL}}(x)$ is continuous and bounded (between 0 and 1). We will use the two properties in our subsequent study.
In addition we see that hinge loss dominates the truncated hinge loss function point wise and that truncated hinge loss function dominates the 0-1 loss point wise. This observation is summarized in the following inequalities.

$$\begin{aligned} L_{\text{HL}}(x) &\geq L_{\text{THL}}(x), \\ L_{\text{THL}}(x) &\geq L_{\text{0-1}}(x) \text{ for each } x \in \mathbb{R}. \end{aligned} \tag{11.4.1}$$

We then calculate the Rademacher Complexity for the new loss function.
Putting everything together we have the following theorem.

Theorem 11.3 (PAC bound for Kernel-Based Hypotheses with Truncated Hinge Loss)**.**

Proof.

□

Theorem 11.4 (PAC Bound with Rademacher Complexity)**.** *Let $\mathbb{H}$ be the R.K.H.S. and $\Phi : \mathbf{x} \to \mathbb{H}$ be the feature mapping associated with the kernel matrix $\mathbf{K}$. Each sample $\mathcal{S} = (\mathbf{x_1}, ..., \mathbf{x_n})$ are drawn from the distribution $\mathcal{D}$. Each σ_i are the Rademacher variables, independent and drawn from a fair Binomial distribution. Finally, we consider all hypothesis $h \in \mathcal{H}$ from the hypothesis space of linear functions such that: $\mathcal{H} = \{\mathbf{x} \to \mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}) : ||\mathbf{w}||_\mathbf{2}^\mathbf{2} < \gamma\}$.*
Then with probability at least $1 - \frac{\delta}{2}$, we have that:

$$\widehat{\mathfrak{R}}_s(\mathcal{H}) \leq \frac{\gamma}{n}\sqrt{Tr(\mathbf{K})}$$

Proof.

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{n} E_\sigma \left[\sup_{||w|| \leq \gamma} < w, \sum_{i=1}^{n} \sigma_i \Phi(x_i) > \right] \\
&= \frac{1}{n} E_\sigma \left[||w|| \cdot || \sum_{i=1}^{n} \sigma_i \Phi(x_i)|| \cos(\theta) \right] \\
&= \frac{\gamma}{n} E_\sigma \left[\cdot || \sum_{i=1}^{n} \sigma_i \Phi(x_i)|| \right] && \text{choose } ||w|| = \gamma, \cos(\theta) = 1 \text{ to max.} \\
&= \frac{\gamma}{n} E_\sigma \left[\cdot || \sum_{i=1}^{n} \sigma_i \Phi(x_i)||^2 \right]^{\frac{1}{2}} \\
&\leq \frac{\gamma}{n} E_\sigma \left[\cdot || \sum_{i=1}^{n} \sigma_i \Phi(x_i)||^2 \right]^{\frac{1}{2}} && \text{Jensen's Inequality} \\
&= \frac{\gamma}{n} E_\sigma \left[\cdot \sum_{i=1}^{n} K(x_i, x_i) \right]^{\frac{1}{2}} && (i \neq j \implies E_\sigma[\sigma_i \sigma_j] = 0] \\
&= \frac{\gamma}{n} \sqrt{\mathrm{Tr}(K)}
\end{aligned}$$

To complete the proof, we notice that with probability $1 - \frac{\delta}{2}$ we bound the Rademacher complexity by its empirical value.

$$\mathfrak{R}_n(\mathcal{F}) \leq \widehat{\mathfrak{R}}_S(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

With the probability union bound we have

$$\begin{aligned}
\Phi(S) &\leq \mathbb{E}_s \left[\Phi(S)\right] + \sqrt{\frac{\ln(2/\delta)}{2n}} \\
&\leq \mathfrak{R}_S(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}} \\
&\leq \widehat{\mathfrak{R}}_S(\mathcal{F}) + 2\sqrt{\frac{\ln(2/\delta)}{2n}},
\end{aligned}$$

and hence the proof of the theorem. □
Since we have that truncated hinge loss function dominates 0-1 loss, we have $\mathbb{P}_D(y \neq sign(g(x))) = \mathbb{P}_D(L_{\text{0-1}}())yg(x) \leq \mathbb{P}_D(L_{\text{THL}}())yg(x)$. Similarily we have $L_{\text{THL}}(yg(x)) \leq L_{\text{HL}}(yg(x))$

Corollary 11.1.

11.5 Bibliographic Notes and Further Reading

11.6 Exercises

11.1[More Example of Loss Functions] Search for definitions of the following loss function definition. Using matlab to plot them. For each loss function, figure whether it is continuous and bounded. Figure out any dominating relationship.

- 0-1 Loss
- Exponential Loss
- Hinge Loss
- Logit Loss
- Smoothed Hinge Loss
- Truncated Hinge Loss
- Modified Huber Loss
- Truncated Quadratic Loss

11.2 [Rademacher Complexity] Let $\{S\} = \{x_1, x_2\}$ where $x_1 = 0$ and $x_2 = 1$ be a sample from a Bernoulli distribution with probability p. Let f_1 maps the set $\{0, 1\}$ to $\mathbb{R}$ in the way that $f_1(0) = 0$ and $f_1(1) = 1$. Let f_2 also maps the set $\{0, 1\}$ to $\mathbb{R}$ in the way that $f_2(0) = 1$ and $f_2(1) = 0$. We have $n = 2$. $\boldsymbol{\sigma} = \{\{1, 1\}, \{1, 0\}, \{0, 1\}, \{0, 0\}\}$ where each element has probability $\frac{1}{4}$. Calculate the empirical Rademacher complexity and Rademacher complexity of $\mathcal{F} = \{f_1, f_2\}$.

Part V

Bayesian Learning

Chapter 12

Statistical Decision Theory

In the previous chapter we have discussed a few classification and regression algorithms and their connections to different learning theories. We showed that L_1 regularization (e.g. sparse learning) and L_2 regularization (e.g. SVMs) are justified by different theories. This observation leads to an interesting question regarding whether other types of regularizations can also be justified. Bayesian theory sheds light on this general problem and offers an elegant solution. The direct introduction of Bayesian theory however has a tiny conceptual difficulty in that there is typically no long-term interpretation of Bayesian estimation. For example suppose that we use Baye's formula to calculate the posterior distribution of a Bernoulli distribution parameter in a dice rolling experiment. That posterior distribution reflects our "belief" of the model parameter distribution although in this case arguably the parameter itself does not change in our experiments. As a result we simply cannot state that that if we watch the system long enough the posterior distribution tells us the "chance" that we observe the parameter.

I believe the full Bayesian treatment in data analytics should involve the discussion of the subjective interpretation of probability, which we plan to discuss in this part but not in this chapter. What we want to discuss in this chapter is to introduce a theory called statistical decision theory. Our discussion starts with the common set-up of maximum likelihood estimator where data is generated by a "machine" with a fixed but known parameter. We want to "reverse-engineering" the machine with samples from it by estimating the parameter. MLE tells us that if we are following the rule and if we have infinite number of samples our estimation of the parameter converges to the true parameter. However in reality we do not have infinite number of samples. What can we do then? Studying this question we see a natural place for Bayesian estimation. In this way we switch gradually from physical probability to evidential probability without an extensive discussion on what is evidential probability. As part of our discussion we also show the connection between Bayesian estimation and another important class of estimation called minimax estimation.

12.1 Risk of Paramter Learning

Definition 12.1 (Loss function)**.** A *loss function* is a function from $\Theta \times \{A\}$ into $\mathbb{R}^+ \cup 0$ where Θ is a parameter space and $\{A\}$ is the set of possible values of an estimator or decision rule. The set $\{A\}$ is also called the *action space*.

Example 12.1. Common loss functions are listed below:

$$\begin{aligned}
L_{\mathrm{S}}(\theta,\widehat{\theta}) &= (\theta-\widehat{\theta})^2 && \text{(Squared loss)},\\
L_{\mathrm{A}}(\theta,\widehat{\theta}) &= \left|\theta-\widehat{\theta}\right| && \text{(Absolute loss)},\\
L_{L_p}(\theta,\widehat{\theta}) &= \left|\theta-\widehat{\theta}\right|^p && (L_p \text{ loss}),\\
L_{0\text{-}1}(\theta,\widehat{\theta}) &= \begin{cases} 0 & \theta=\widehat{\theta} \\ 1 & \text{otherwise,} \end{cases} && \text{(Zero-one loss)}\\
L_{\mathrm{KL}}(\theta,\widehat{\theta}) &= \int \log\frac{f(x;\theta)}{f(x;\widehat{\theta})} f(x;\theta)dx && \text{(Kullback-Leibler loss)}.
\end{aligned}$$

$\hat{\theta}$ is an estimation of the true parameter θ. For simplicity and following convention $\hat{\theta}$ and θ are assumed to be scalars. Our discussion applies to vector case where we simply replace the related notation by vector norms.
Clearly $\hat{\theta}$ is a (none negative) random variable that depends on the data that we have. To understand the behavior of $\hat{\theta}$ we compute the expectation of $\hat{\theta}$. Such value is called the *risk* of our estimation.

Definition 12.2 (Risk)**.** The *risk* of an estimator $\widehat{\theta}$ is

$$R(\theta,\widehat{\theta}) = \mathrm{E}_\theta(L(\theta,\widehat{\theta})) = \int L(\theta,\widehat{\theta}) f(x;\theta)\,dx.$$

With the concept of risk, we could compare different estimators and we illustrate such comparisons using the following example.

Example 12.2. Let $X_1,\ldots,X_n$ be an i.i.d. random sample from the Bernoulli distribution with parameter p. We'll assume the squared loss function. There can be many kinds of estimators:

$$\begin{aligned}
\widehat{p}_{\mathrm{MLE}} &= \frac{\sum_i X_i}{n}\\
\widehat{p}_1 &= \frac{\sum_{i=1}^m X_i}{m}, \quad m<n\\
\widehat{p}_2 &= \frac{\sum_i X_i+\alpha}{n+\alpha+\beta}, \quad \alpha>0,\ \beta>0\\
\widehat{p}_3 &= \frac{\sum_i X_i+1}{n+2}\\
\widehat{p}_4 &= \frac{\sum_i X_i+\frac{\sqrt{n}}{2}}{\sqrt{n}+n}
\end{aligned}$$

Using squared loss the risk for $\widehat{p}_{\mathrm{MLE}}$ is

$$\begin{aligned}
R(p,\widehat{p}_{\mathrm{MLE}}) &= \mathbb{E}\left[\left(\frac{\sum_i X_i}{n}-p\right)^2\right]\\
&= \mathbb{V}\left[\frac{\sum_i X_i}{n}\right]\\
&= \frac{\sum_i \mathbb{V}[X_i]}{n^2}\\
&= \frac{p(1-p)}{n}.
\end{aligned}$$

Similarly, the risk for p_1 is

$$R(p, \widehat{p}_1) = \frac{p(1-p)}{m}.$$

Observe that

$$R(p, \widehat{p}_{\text{MLE}}) \leq R(p, \widehat{p}_1)$$

for any $0 \leq p \leq 1$ since $m < n$.
The risk for $\hat{p}_2$ is

$$\begin{aligned} R(p, \hat{p}_2) &= \mathbb{E}\left[\left(p - \frac{\sum_i X_i + \alpha}{n + \alpha + \beta}\right)^2\right] \\ &= \text{Bias}^2\left(\frac{\sum_i X_i + \alpha}{n + \alpha + \beta}\right) + \mathbb{V}\left[\left(\frac{\sum_i X_i + \alpha}{n + \alpha + \beta}\right)\right] \\ &= (\frac{np + \alpha}{n + \alpha + \beta} - p)^2 + \frac{np(1-p)}{(n + \alpha + \beta)^2} \\ &= (\frac{\alpha - \alpha p - \beta p}{n + \alpha + \beta})^2 + \frac{np(1-p)}{(n + \alpha + \beta)^2} \\ &= \frac{(\alpha - \alpha p - \beta p)^2 + np(1-p)}{(n + \alpha + \beta)^2}. \end{aligned}$$

For the other two estimators, we can just plug in the corresponding values of α and β and obtain the risk.

In Figure **??** we show the risk function of all the estimators.
Among all the estimators p_1 is quite special in that we find another estimator whose risk is no more than that of p_1 no matter the value of p. We call such estimators *inadmissible*. If an estimator is inadmissible it probably should not be used in any estimation since there exists another one that performs always no worse than the inadmissible one.
Similarly we define *admissible* estimators below.

Definition 12.3 (Admissible Estimators)**.** An estimator $\hat{\theta}$ is *admissible* if there exists no estimators $\tilde{\theta}$ such that $R(\theta, \tilde{\theta}) \leq (\theta, \hat{\theta})$ for all possible θ.

Clearly a reasonable requirement for an estimator is that it must be admissible. A way to find such estimators (or to prove that an estimator is admissible) is discussed at the end of the following section.

12.2 Comparing Estimator Risks

As illustrated in the previous example for a given estimator, its risk is a function of the true parameter. In order to compare different estimators, one way is to somehow "summarize" the risks and use a single value for each estimator. Below we present two approaches to perform such summarization.
The first is to identify the worst senior (where the risk is maximized among all possible θ) and aims to obtain the best worst senior. This rule is the minimax rule.

Definition 12.4 (Minimax rule)**.** The *maximum risk* is defined as

$$\overline{R}(\widehat{\theta}) = \max_{\theta} R(\theta, \widehat{\theta}).$$

An estimator $\widehat{\theta}_{mm}$ that minimizes the maximum risk among all possible estimators is called a *minimax rule*. Formally,

$$\widehat{\theta}_{mm} = \arg\min_{\widehat{\theta}} \left(\max_{\theta} R(\theta, \widehat{\theta}) \right).$$

The other approach is to compute the "weighted" area of curve for the risk function $(\theta, \hat{\theta})$ and to minimize such value. This rule is the Bayes rule.

Definition 12.5 (Bayes rule)**.** The *Bayes risk* is defined as

$$r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta)\, d\theta$$

where $f(\theta)$ is a non-negative function that assigns "weight" to different θ. An estimator $\widehat{\theta}_B$ that minimizes the Bayes risk is called a *Bayes rule*. Formally,

$$\widehat{\theta}_B = \arg\min_{\widehat{\theta}} r(f, \widehat{\theta}).$$

At the first glance applying both the miniMax rule and the Bayes rule looks very difficult since we have (at conceptually) to go through *all* possible estimators. How could we rule out the possibility that there does not exists another estimator that has lower maximal risk or Bayes risk? Below we present a theorem that shows such applying such rules are completely possible.

Theorem 12.1. *The Bayes risk satisfies*

$$r(f, \widehat{\theta}) = \int r(\widehat{\theta}|x) m(x)\, dx$$

where

$$r(\widehat{\theta}|x) = \int L(\theta, \widehat{\theta}) f(\theta|x)\, d\theta,$$

is called the posterior risk and

$$m(x) = \int f(x|\theta) f(\theta)\, d\theta.$$

is the marginal distribution.

Proof.

$$\begin{aligned}
r(f, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}) f(\theta)\, d\theta \\
&= \int \left(\int L(\theta, \widehat{\theta}) f(x|\theta)\, dx \right) f(\theta)\, d\theta \\
&= \int \int L(\theta, \widehat{\theta}) f(x, \theta)\, dx\, d\theta \\
&= \int \int L(\theta, \widehat{\theta}) f(\theta|x) m(x)\, dx\, d\theta \qquad \text{(by Bayes' Theorem)} \\
&= \int \left(\int L(\theta, \widehat{\theta}) f(\theta|x)\, d\theta \right) m(x)\, dx \\
&= \int r(\widehat{\theta}|x) m(x)\, dx.
\end{aligned}$$

□

12.3 Minimax Estimator

12.4 Stein's Paradox and Asymptotic Behavior of Estimators

In this chapter we discussed admissibility of estimators. By definition, if a minimax estimator is unique (meaning there exists only one minimax estimator) it must be admissible since it has the lowest maximal loss. Following the same logic unique Bayesian estimator for an arbitrary prior is always admissible since it minimizes the Bayes' risk function.
An interesting question is whether MLE is admissible. Below we present Steins Paradox, which gives a clear negative answer to the question.
Retrospectively "consistency" played a central role in the development of maximum likelihood estimation. MLE estimator is consistent. As we studied before MLE may not always lead to definite answers. For example in logistic regression, MLE may not lead to definite answer if the data is well separated. In addition in linear regression, if the data matrix $\mathbf{X}$ is not full ranked, $\mathbf{X}^T\mathbf{X}$ may not be have an inverse and hence we do not have a definite MLE estimator. There are multiple such cases. However if MLE does lead to definite answer, any Bayesian estimator and minimax estimator are consistent under mild conditions.

Definition 12.6 (Well-behaved Loss Function)**.** A loss function is *well-behaved* if $L(\theta, \hat{\theta}) = 0$ when $\theta = \hat{\theta}$ and $L(\theta, \hat{\theta}) > 0$ otherwise.

Theorem 12.2. *For estimation problems where MLE estimator exists, if the loss function is well-behaved, the Bayesian estimator is consistent. In addition any minimax estimator is also consistent.*

Proof. Since MLE is consistent, if the number of samples n goes infinite, the related risk function converges point-wise to zero. It immediately follows that minimax estimator must also be consistent. Consider an arbitrary Bayesian estimator that minimizes the Bayes' risk for a given prior distribution. Since there exists an estimator whose risk function converges point-wise to zero, the risk function for the Bayesian estimator must converge point-wise to zero, which suggests that the Bayesian estimator is consistent. □

12.5 Bibliographic Notes and Further Reading

Much of the discussion is from classical Bayesian where our setup is that data are from an unknown true model. Further reading of statistical decision theory may be found at [10]. Asymptotic analysis of Beyesian estimators could be found at Bernstein-von Mises theorem [9]. Doob-style consistency theorem for stationary models is discussed in [6]. A review on consistency of Bayesian estimators was presented in [4].
In the chapter **??** we discuss subjective probability where our focus is on predictive distribution of future data (a.k.a. an operationalist perspective).

Chapter 13

Multivariate Gaussian and Conjugate Prior

13.1 Multivariate Gaussian Distribution

Definition 13.1 (Multivariate Gaussian Distribution)**.** Given $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x}$ is a multivariate gaussian random variable if the density function of $\mathbf{x}$ has the format of

$$\begin{aligned}\mathbf{x} &\sim \mathcal{N}(x|\mu, \mathbf{\Sigma}) \\ &= \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))\end{aligned}$$

where $\boldsymbol{\mu}$ is a p dimensional vector called the mean. $\mathbf{\Sigma}$ is a $p \times p$ positive semi-definite matrix called the covariance matrix. $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$, $\mathbf{\Sigma}^{-1}$ is the so-called *precision matrix*.

We usually set the matrix $\mathbf{\Sigma}$ to be symmetric. If $\mathbf{\Sigma}$ is not with $(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^T\frac{1}{2}(\mathbf{S}^{-1}+(\mathbf{S}^{-1})^T)(\mathbf{x}-\boldsymbol{\mu})$ we could simply define a symmetric matrix $\mathbf{\Sigma}_1^{-1} = \frac{1}{2}(\mathbf{\Sigma}^{-1}+(\mathbf{\Sigma}^{-1})^T)$ to replace $\mathbf{\Sigma}$ and keep the density function unchanged.
We should notice that multivariate Gaussian distribution has only one peak in the density function and hence it is unimodal. In general we should exercise caution when using multivariate Gaussian distribution if data may have multiple peaks in its density function. On the other hand, multivariate Gaussian distribution are quite flexible. It has a total of $\frac{1}{2}p(p+1)+p$ free variables (p dimensional mean and $p \times p$ symmetric covariance matrix). We also notice that any Gaussian distributions are determined completely by $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$.
Below we show why we call $\mathbf{\Sigma}$ the covariance matrix.
With straightforward calculation we have the following theorem.

Theorem 13.1. *Given* $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\mathbf{0}, \sigma^2\mathbf{I})$*, we have* $\mathbb{E}[\mathbf{x}] = 0$ *and* $\mathrm{Cov}[\mathbf{x}] = \sigma^2\mathbf{I}$

Proof.

$$\begin{aligned}f(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2}|\mathbf{S}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \\ &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{x_i^2}{\sigma^2}) \quad (|\mathbf{\Sigma}| = \sigma^{2p})\end{aligned}$$

Since the normal probability density function is symmetric with respect to y-axis, we conclude that $\mathbb{E}[\mathbf{x}] = \int \mathbf{x}f(\mathbf{x})\ d\mathbf{x} = 0$.

For the covariance matrix, we let $\mathbf{V} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]$ and we have

$$\begin{aligned}\mathbf{V}(i,j) &= \int x_i x_j \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{x_i^2}{\sigma^2})\, d\mathbf{x} \\ &= \begin{cases} 0, & \text{if } i \neq j \\ \sigma^2, & \text{if } i = j. \end{cases}\end{aligned}$$

Obviously, $\mathbf{V} = \mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right] = \mathrm{Cov}\left[\mathbf{x}\right] = \sigma^2\mathbf{I} = \mathbf{\Sigma}$. The results are easily extended to cases where $\boldsymbol{\mu} \neq 0$. □

Consider now $\mathbf{y} = \mathbf{A}\mathbf{x}$, $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$, $\mathbf{A}$ is a $p \times p$ full rank matrix, then we have

$$\begin{aligned}f(\mathbf{y}) &= f(\mathbf{x})\frac{d_x\mathbf{V}}{d_y\mathbf{V}} \\ &= \frac{1}{(2\pi)^{D/2}} \exp(-\frac{1}{2}\mathbf{y}^T(\mathbf{A}^{-1})^T(\mathbf{A}^{-1})\mathbf{y})\frac{1}{|\det(\mathbf{A})|} \\ &= \frac{1}{(2\pi)^{D/2}|\det(\mathbf{A})|} \exp(-\frac{1}{2}\mathbf{y}^T\mathbf{S}^{-1}\mathbf{S}) \\ &= \frac{1}{(2\pi)^{D/2}|\mathbf{S}|^{1/2}} \exp(-\frac{1}{2}\mathbf{y}^T\mathbf{S}^{-1}\mathbf{y})\end{aligned}$$

since $(\mathbf{A}^{-1})^T(\mathbf{A}^{-1}) = \mathbf{S}^{-1}$, $\det(\mathbf{S}) = \det(\mathbf{A})\det(\mathbf{A}^T) = (\det(\mathbf{A}))^2$. Then we have,

$$\begin{aligned}\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] &= \mathbb{E}\left[\mathbf{A}\mathbf{x}\mathbf{x}^T\mathbf{A}\right] \\ &= \mathbf{A}\mathbb{E}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{A}^T \\ &= \mathbf{A}\mathbf{A}^T \\ &= \mathbf{\Sigma}\end{aligned}$$

For any Gaussian, $\mathbb{E}\left[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\right] = \mathbf{S}$, $f(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{S})$.

13.2 Conditional Distribution of Multivariate Gaussian Distributions

We know that the logarithm of the pdf of a multivariate Gaussian random variable is a quadratic function of a multi-dimensional vector. The following theorem states that the inverse is also correct. This theorem is very important for the rest of the discussion when we want to prove that a random variable is a multi-variate Gaussian.

Theorem 13.2. *Given a random vector* $\mathbf{x}$ *if the logarithm of its pdf function takes the following quadratic form*

$$\ln f(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T{\mathbf{A}_0}^{-1}\mathbf{x} + \mathbf{x}^T\mathbf{b}_0 + C \tag{13.2.1}$$

where C *is a constant that is no depends on* $\mathbf{x}$. $\mathbf{A}_0$ *is a symmetric positive definite matrix.* $\mathbf{x}$ *must be a multivariate Gaussian with the pdf function* $f(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x})$ *and*

$$\begin{aligned}\mathbf{\Sigma}_\mathbf{x} &= \mathbf{A}_0, \\ \boldsymbol{\mu}_\mathbf{x} &= \mathbf{A}_0 \cdot \mathbf{b}_0.\end{aligned} \tag{13.2.2}$$

Proof.

□

As an example of Theorem 13.2, we show the following important property of multivariate Gaussian distributions.

Theorem 13.3. *Suppose* $\mathbf{x} = [\mathbf{x}_a^T, \mathbf{x}_b^T]^T$ *has a joint Gaussian distribution, the conditional distribution* $\mathbf{x_a}$ *given* $\mathbf{x_b}$ *is also a multivariate Gaussian distribution.*

Proof. Let $f(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}.$$

We have the log pdf function as

$$\ln f(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu) + C_0 \tag{13.2.3}$$

where $C_0 = -\frac{1}{2}|\boldsymbol{\Sigma}| - \frac{p}{2}\ln(2\pi)$ is a constant.
By definition $f(\mathbf{x}_a|\mathbf{x}_b) = f(\mathbf{x}_a, \mathbf{x}_b)/f(\mathbf{x}_b)$ where $f(\mathbf{x}_b)$ is the pdf of the marginal distribution of $\mathbf{x}_b$. Hence we have

$$\begin{aligned} \ln f(\mathbf{x}_a|\mathbf{x}_b) &= \ln f(\mathbf{x}) - \ln f(\mathbf{x}_b) \\ &= -\frac{1}{2}(\mathbf{x} - \mu)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu) + C_0 - \ln f(\mathbf{x}_b) \\ &= -\frac{1}{2}(\mathbf{x_a} - \mu_a)^T \boldsymbol{\Lambda}_{aa}(\mathbf{x_a} - \mu_a) - \frac{1}{2}(\mathbf{x_a} - \mu_a)^T \boldsymbol{\Lambda}_{ab}(\mathbf{x_b} - \mu_b) \\ &\quad -\frac{1}{2}(\mathbf{x_b} - \mu_b)^T \boldsymbol{\Lambda}_{ba}(\mathbf{x_a} - \mu_a) - \frac{1}{2}(\mathbf{x_b} - \mu_b)^T \boldsymbol{\Lambda}_{bb}(\mathbf{x_b} - \mu_b) + C_1 \end{aligned} \tag{13.2.4}$$

where $C_1 = C_0 - \ln f(\mathbf{x}_b)$ is a constant that does not depend on $\mathbf{x}_a$.
In calculating the pdf for the conditional distribution $\mathbf{x_b}, \mu_a, \mu_b$ are fixed. The above expression is a quadratic function with respect to $\mathbf{x_a}$. If we rewrite (13.2.4) we obtain:

$$\ell(\mathbf{x}) = -\frac{1}{2}\mathbf{x_a}^T \boldsymbol{\Lambda}_{aa}\mathbf{x_a} + \mathbf{x_a}^T(\boldsymbol{\Lambda}_{aa}\mu_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x_b} - \boldsymbol{\mu}_b)) + C_2 \tag{13.2.5}$$

where $C_2 = -\frac{1}{2}\boldsymbol{\mu}_a\boldsymbol{\Lambda}_{aa}\mu_a + \mu_a^T\boldsymbol{\Lambda}_{ab}(\mathbf{x_b} - \mu_b) - \frac{1}{2}(\mathbf{x_b} - \mu_b)^T\boldsymbol{\Lambda}_{bb}(\mathbf{x_b} - \mu_b) + C_1$ is a constant. In deriving (13.2.5) we use the facts that $\boldsymbol{\Lambda}_{ab} = \boldsymbol{\Lambda}_{ba}$ since the precision matrix is symmetric.
Applying Theorem 13.2 we have:

$$\begin{aligned} f(\mathbf{x_a}|\mathbf{x_b}) &= \mathcal{N}(\mu_{a|b}, \boldsymbol{\Sigma}_{a|b}), \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1}, \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{\mathbf{a|b}}(\boldsymbol{\Lambda}_{\mathbf{aa}}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x_b} - \boldsymbol{\mu}_b)) \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x_b} - \boldsymbol{\mu}_b). \end{aligned} \tag{13.2.6}$$

□

Schur's lemma computes the inverse of a partitioned matrix as follows

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{bmatrix} \tag{13.2.7}$$

where $\mathbf{M} = (\mathbf{A} - \mathbf{BD^{-1}C})^{-1}$. Applying Schur's lemma we obtain

$$\begin{aligned}\mathbf{\Lambda} &= \mathbf{\Sigma}^{-1} \\ &= \begin{bmatrix}\mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb}\end{bmatrix}^{-1} \\ &= \begin{bmatrix}\mathbf{M} & -\mathbf{M}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1} \\ -\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba}\mathbf{M} & \mathbf{\Sigma}_{bb}^{-1} + \mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma_{ba}}\mathbf{M}\mathbf{\Sigma_{ab}}\mathbf{\Sigma}_{bb}^{-1}\end{bmatrix}.\end{aligned}$$

$\mathbf{M} = (\mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}{\mathbf{\Sigma}_{bb}}^{-1}\mathbf{\Sigma}_{ba})^{-1}$
Clearly we have $\mathbf{\Lambda}_{aa} = \mathbf{M}$ and $\mathbf{\Lambda}_{ab} = -\mathbf{M}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}$.
Plug results into (13.2.6) we have

$$\begin{aligned}f(\mathbf{x_a}|\mathbf{x_b}) &= \mathcal{N}(\mu_{a|b}, \mathbf{\Sigma}_{a|b}), \\ \mathbf{\Sigma}_{a|b} &= \mathbf{\Lambda}_{aa}^{-1} \\ &= \mathbf{M}^{-1} \\ &= \mathbf{\Sigma}_{aa} - \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}\mathbf{\Sigma}_{ba} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x_b} - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a - \mathbf{M}^{-1}(-\mathbf{M}\mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1})(\mathbf{x_b} - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \mathbf{\Sigma}_{ab}\mathbf{\Sigma}_{bb}^{-1}(\mathbf{x_b} - \boldsymbol{\mu}_b).\end{aligned} \tag{13.2.8}$$

Using a similar trick (See Exercise 13.1) we obtain the following conclusion regarding the marginal distribution of $\mathbf{b}$ and $\mathbf{a}$.

$$\begin{aligned}f(\mathbf{x_b}) &= \mathcal{N}(\nu_b, \mathbf{\Sigma}_b), \\ \mathbf{\Sigma}_b &= \mathbf{\Sigma}_{bb}, \\ \boldsymbol{\nu}_b &= \boldsymbol{\mu}_b, \\ f(\mathbf{x_a}) &= \mathcal{N}(\nu_a, \mathbf{\Sigma}_a), \\ \mathbf{\Sigma}_a &= \mathbf{\Sigma}_{aa}, \\ \boldsymbol{\nu}_a &= \boldsymbol{\mu}_a.\end{aligned} \tag{13.2.9}$$

Basically the marginal distributions have a simple form by picking up related components from the joint distribution.

13.3 Joint Distributions of Multivariate Gaussian Distributions

Theorem 13.4. *Suppose we have* $f(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{S_0})$, $f(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{S_1})$, *then the marginal distribution* $f(\mathbf{y})$ *is a multivariate Gaussian* $\mathcal{N}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{\Sigma}_\mathbf{y})$ *and the conditional distribution* $f(\mathbf{x}|\mathbf{y})$ *is again a multivariate Gaussian* $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{x}|y})$, *where*

$$\begin{aligned}\boldsymbol{\mu}_\mathbf{y} &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \\ \mathbf{\Sigma}_\mathbf{y} &= \mathbf{S_1} + \mathbf{AS_0}\mathbf{A}^T, \\ \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} &= \mu + \mathbf{\Sigma}_{\mathbf{x}|\mathbf{y}}\mathbf{A}^T{\mathbf{S_1}}^{-1}(\mathbf{y} - \mathbf{A}\mu - \mathbf{b}), \\ \mathbf{\Sigma}_{\mathbf{x}|\mathbf{y}} &= ({\mathbf{S_0}}^{-1} + \mathbf{A}^T{\mathbf{S_1}}^{-1}\mathbf{A})^{-1}.\end{aligned} \tag{13.3.1}$$

Proof. Let $\mathbf{z} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$. The logarithm of the probability density function of $\mathbf{z}$ can be expressed through conditional probability as

$$\begin{aligned}\ln f(\mathbf{z}) =& \ln f(\mathbf{x}) + \ln f(\mathbf{y}|\mathbf{x})\\ =& -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T {\mathbf{S_0}}^{-1}(\mathbf{x}-\boldsymbol{\mu})\\ &-\frac{1}{2}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b})^T {\mathbf{S_1}}^{-1}(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}) + C\end{aligned} \tag{13.3.2}$$

Where C is a constant that does not depends on $\mathbf{x}$ or $\mathbf{y}$.
Expanding the quadratic formula, regrouping similar items, and ignoring constants (13.3.2) may be rewritten as $\ln f(\mathbf{z}) = \mathrm{I} + \mathrm{II}$ where I is the quadratic component and II the linear component as

$$\begin{aligned}\mathrm{I} &= -\frac{1}{2}\mathbf{x}^T(\mathbf{S}_0^{-1} + \mathbf{A}^T\mathbf{S}_1^{-1}\mathbf{A})\mathbf{x} + -\frac{1}{2}\mathbf{y}^T\mathbf{S}_1^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{S}_1^{-1}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{S}_1^{-1}\mathbf{y}\\ &= -\frac{1}{2}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^T \begin{bmatrix}\mathbf{S}_0^{-1} + \mathbf{A}^T\mathbf{S}_1^{-1}\mathbf{A} & -\mathbf{A}^T\mathbf{S}_1^{-1}\\ -\mathbf{S}_1^{-1}\mathbf{A} & \mathbf{S}_1^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix},\\ \mathrm{II} &= \mathbf{x}^T\mathbf{S}_0^{-1}\boldsymbol{\mu} - x^T\mathbf{A}^T\mathbf{S}_1^{-1}\mathbf{b} + \mathbf{y}^T\mathbf{S}_1^{-1}\mathbf{b}\\ &= \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}^T \begin{bmatrix}\mathbf{S}_0^{-1}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{S}_1^{-1}\mathbf{b}\\ \mathbf{S}_1^{-1}\mathbf{b}.\end{bmatrix}\end{aligned} \tag{13.3.3}$$

Applying Theorem 13.2 we immediately conclude that $\mathbf{z}$ is a multivariate Gaussian with pdf as $f(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{z}, \Sigma_\mathbf{z})$ where

$$\begin{aligned}\Lambda_\mathbf{z} &= \begin{bmatrix}{\mathbf{S_0}}^{-1} + \mathbf{A}^T{\mathbf{S_1}}^{-1}\mathbf{A} & -\mathbf{A}^T{\mathbf{S_1}}^{-1}\\ -{\mathbf{S_1}}^{-1}\mathbf{A} & {\mathbf{S_1}}^{-1}\end{bmatrix},\\ \Sigma_\mathbf{z} &= \Lambda_\mathbf{z}^{-1}\\ &= \begin{bmatrix}\mathbf{S_0} & \mathbf{S_0}\mathbf{A}^T\\ \mathbf{A}\mathbf{S_0} & \mathbf{S_1} + \mathbf{A}\mathbf{S_0}\mathbf{A}^T\end{bmatrix},\\ \boldsymbol{\mu}_\mathbf{z} &= \Sigma_\mathbf{z}\begin{bmatrix}{\mathbf{S_0}}^{-1}\boldsymbol{\mu} - \mathbf{A}^T{\mathbf{S_1}}^{-1}\mathbf{b}\\ {\mathbf{S_1}}^{-1}\mathbf{b}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b}\end{bmatrix}.\end{aligned} \tag{13.3.4}$$

$\Lambda_\mathbf{z}$ is the precision matrix of the Gaussian variable $\mathbf{z}$.
Following (13.2.9) we obtain the marginal distribution of $\mathbf{y}$ as

$$\begin{aligned}f(\mathbf{y}) &\sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y}),\\ \boldsymbol{\mu}_\mathbf{y} &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b},\\ \boldsymbol{\Sigma}_\mathbf{y} &= \mathbf{S_1} + \mathbf{A}\mathbf{S_0}\mathbf{A}^T.\end{aligned} \tag{13.3.5}$$

Following (13.2.8) we obtain the conditional distribution of $\mathbf{x}|\mathbf{y}$ as

$$\begin{aligned}f(\mathbf{x}|\mathbf{y}) &\sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|y}),\\ \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} &= \mu + \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\mathbf{A}^T{\mathbf{S_1}}^{-1}(\mathbf{y} - \mathbf{A}\mu - \mathbf{b}),\\ \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} &= ({\mathbf{S_0}}^{-1} + \mathbf{A}^T{\mathbf{S_1}}^{-1}\mathbf{A})^{-1}.\end{aligned} \tag{13.3.6}$$

□

13.4 Exercises

13.1[Marginal Distribution of Multivariate Gaussian Distributions] Using the formula $\ln f(\mathbf{b}) = \ln f(\mathbf{a}, \mathbf{b}) - \ln f(\mathbf{a}|\mathbf{b})$ to prove that

$$\begin{aligned} f(\mathbf{x_b}) &= \mathcal{N}(\nu_b, \mathbf{\Sigma}_b), \\ \mathbf{\Sigma}_b &= \mathbf{\Sigma}_{bb}, \\ \boldsymbol{\nu}_b &= \boldsymbol{\mu}_b. \end{aligned} \tag{13.4.1}$$

Hint: $\ln f(\mathbf{a}|\mathbf{b})$ may be written as $-\frac{1}{2}((\mathbf{x}_a - \mu_a) - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x}_b - \mu_b))^T \mathbf{\Lambda}_{aa}((\mathbf{x}_a - \mu_a) - \mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab}(\mathbf{x}_b - \mu_b))$ added with a constant. You may also use the fact that $\mathbf{\Lambda}_{bb} - \mathbf{\Lambda}_{ba}\mathbf{\Lambda}_{aa}^{-1}\mathbf{\Lambda}_{ab} = \mathbf{\Sigma}_{bb}^{-1}$.

Chapter 14

Bayesian Linear Regression

14.1 Bayesian Linear Regression Specification

We have the model

$$\mathbf{y_i} = \mathbf{x_i}^T\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

ϵ_i and ϵ_j are independent if $i \neq j$. $i, j \in [1, n]$ where n is the total number of training samples that we have.
An equivalent representation is using multivariate Gaussian. That is

$$f(\mathbf{y}|\beta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

Assume that the desired parameter β also have some prior distribution, and for simplicity, we say normal distribution, as $f(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2\mathbf{I})$. This is a hierarchical model as shown in Figure ().
As discussed in Chapter 13, since $\boldsymbol{\beta}$ is a multivariate Gaussian and $\mathbf{y}|\boldsymbol{\beta}$ is also a multivariate Gaussian, we know the joint distribution and the posterior distribution are both Gaussian. Before we perform the full Bayesian treatment, we first study a rather special case where we only pick up one point, the mode, from the posterior distribution.

14.2 Maximum A Posterior Estimation

Definition 14.1 (Maximum A Posterior Estimation)**.** The maximum a posterior estimation of a parameter β or $\widehat{\beta}_{\text{MAP}}$ is the parameter that maximize the posterior distribution. That is

$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} f(\beta|\mathbf{y}) \tag{14.2.1}$$

Example 14.1. For Bayesian linear regression where $f(\mathbf{y}|\beta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $f(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2\mathbf{I})$ we have $\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} f(\beta|\mathbf{y})$. Since logarithm is a monotonic transformation we also have

$$\begin{aligned}
\widehat{\beta}_{\text{MAP}} &= \arg\max_{\beta} f(\beta|\mathbf{y}) \\
&= \arg\max_{\beta} \ln f(\beta|\mathbf{y}) \\
&= \arg\max_{\beta} \ln f(\boldsymbol{\beta}) + \ln f(\mathbf{y}|\beta) - \ln f(\mathbf{y}).
\end{aligned}$$

Since $f(\mathbf{y})$ is a constant that does not rely on the value of β we have

$$\begin{aligned}\widehat{\beta}_{\text{MAP}} &= \arg\max_{\beta} \ln f(\boldsymbol{\beta}) + \ln f(\mathbf{y}|\beta) \\ &= \frac{\beta^T\beta}{\sigma_\beta^2} + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} + C\end{aligned}$$

where $C = -p\ln\sqrt{2\pi} - p\ln\sigma_\beta - n\ln\sqrt{2\pi} - n\ln\sigma$.
Rearrange the previous equation we have

$$\widehat{\beta}_{\text{MAP}} = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\beta^T\beta$$

where $\lambda = \frac{\sigma^2}{\sigma_\beta^2}$.
We see the same formalization in Ridge regression (e.g. 8.2.1). Using the Ridge regression formula (8.2.2) we have

$$\widehat{\beta}_{\text{MAP}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{14.2.2}$$

$\frac{\sigma^2}{\sigma_\beta^2}$ quantifies the noise level of the data. If we use λ to denote this ratio, the result given by Bayesian Linear Regression assuming Gaussian prior is equivalent to that of the Ridge regression.

14.3 Posterior Distribution of Model Parameter

In the MAP estimation we obtain the mode of the posterior distribution. Here we show that the posterior distribution is of a quit familiar format: that of multivariate Gaussian. The result is a natural application of what we discussed in the chapter of multivariate Gaussian. Given $f(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2\mathbf{I})$ and $f(\mathbf{y}|\beta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, set the following variables in Theorem 13.4

$$\begin{aligned}\mathbf{A} &= \mathbf{X} \\ \mathbf{x} &= \boldsymbol{\beta} \\ \mathbf{b} &= \mathbf{0} \\ \boldsymbol{\mu} &= \mathbf{0} \\ \mathbf{S}_1 &= \sigma^2\mathbf{I} \\ \mathbf{S}_0 &= \sigma_\beta^2\mathbf{I}.\end{aligned} \tag{14.3.1}$$

Applying the Theorem 13.4 we have $f(\boldsymbol{\beta}|\mathbf{y}) = \mathcal{N}(\mu_{\boldsymbol{\beta}|\mathbf{y}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}})$ is a multivariate Gaussian. The covariance matrix and expectation are

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} &= (\mathbf{S}_0^{-1} + \mathbf{A}^T\mathbf{S}_1^{-1}\mathbf{A})^{-1} \\ &= \sigma^2(\frac{\sigma^2}{\sigma_\beta^2}\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}, \\ \boldsymbol{\mu}_{\boldsymbol{\beta}|\mathbf{y}} &= \mu + \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\mathbf{A}^T{\mathbf{S}_\mathbf{1}}^{-1}(\mathbf{y} - \mathbf{A}\mu - \mathbf{b}) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}|\mathbf{y}} \cdot \mathbf{X}^T\sigma^{-2}\mathbf{I}\mathbf{y} \\ &= (\frac{\sigma^2}{\sigma_\beta^2}\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

Since the posterior distribution is Gaussian, the mode is the posterior mean as we see in the MAP estimation.

14.4 Predictive Posterior Distribution

After estimating a parameter (or its distribution in Bayesian analysis) our focus turns to making predictions. One approach is to utilize the posterior distribution of the parameter by evaluating the posterior distribution of the parameter $\boldsymbol{\beta}$. For each β we plug it into a generative model to evaluate the distribution of observable. The overall distribution of the observable is the combination of the two distributions, or

$$f(y_{n+1}|\mathbf{X}, \mathbf{y}_n) = \int f(y_{n+1}|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}_n)\, d\boldsymbol{\beta}. \tag{14.4.1}$$

For Bayesian linear regression we have

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}_n) &= \mathcal{N}(\boldsymbol{\mu}_{\beta|\mathbf{y}_n}, \boldsymbol{\Sigma}_{\beta|\mathbf{y}_n}) \\ f(y_{n+1}|\boldsymbol{\beta}) &= \mathcal{N}(y_{n+1}|\mathbf{x}_{n+1}^T\boldsymbol{\beta}, \sigma^2), \end{aligned} \tag{14.4.2}$$

applying Theorem 13.4 we have

$$\begin{aligned} f(y_{n+1}|\mathbf{X}, \mathbf{y}_n) &= \int \mathcal{N}(y_{n+1}|\mathbf{x}_{n+1}^T\boldsymbol{\beta}, \sigma^2)\mathcal{N}(\boldsymbol{\mu}_{\beta|\mathbf{y}_n}, \boldsymbol{\Sigma}_{\beta|\mathbf{y}_n})\, d\boldsymbol{\beta} \\ &= \mathcal{N}(\boldsymbol{\mu}_{y_{n+1}}, \sigma^2_{y_{n+1}}), \end{aligned} \tag{14.4.3}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{y_{n+1}} &= \mathbf{x}_{n+1}^T\mu_{\beta|\mathbf{y}_n} \\ &= \mathbf{x}_{n+1}^T(\frac{\sigma^2}{\sigma_\beta^2}\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \sigma^2_{y_{n+1}} &= \sigma^2 + \mathbf{x}_{n+1}^T\Sigma_{\boldsymbol{\beta}|\mathbf{y}_n}\mathbf{x}_{n+1} \\ &= \sigma^2 + \sigma^2\mathbf{x}_{n+1}^T(\frac{\sigma^2}{\sigma_\beta^2}\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{n+1}. \end{aligned} \tag{14.4.4}$$

The reason we could obtain the predictive distribution without evaluate the integration is because of Theorem 13.4. In this case we could "view" $f(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}_n)$ as the prior distribution of $\boldsymbol{\beta}$ and $f(y_{n+1}|\mathbf{X}, \mathbf{y}_n)$ as the marginal distribution of y_{n+1} when we apply Theorem 13.4.

Chapter 15

Gaussian Process

15.1 Nonparametric Models

Gaussian process is a nonparametric Bayesian method. Suppose that the observable data is conditional on a hidden variable. That is

$$y_n = z_n + \epsilon_n \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2).$$

Let random variable $\mathbf{z}_n = [z_1, z_2, \cdots, z_n]^T$ and has the distribution $f(\mathbf{z}_n) = \mathcal{N}(\mathbf{0}, \mathbf{K})$. $\mathbf{K}$ is a $n \times n$ covariance matrix, which is also the kernel matrix of $\mathbf{x}$ or

$$\mathbf{K} = (k_{i,j})_{i,j=1}^N, \quad \text{where} \quad k_{i,j} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

Given the training data $\mathbf{y}_n = [y_1, y_2, \cdots, y_n]^T$ and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ our goal is to derive the predictive distribution of $f(y_{n+1}|\mathbf{y}_n)$ from the model that we specified.
With $f(\mathbf{y}_{n+1}|\mathbf{z}_{n+1}) = \mathcal{N}(\mathbf{z}_{n+1}, \sigma^2\mathbf{I}_{n+1})$, and $f(\mathbf{z}_{n+1}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{n+1})$ by Theorem 13.4 we conclude that $f(\mathbf{y}_{n+1}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y})$ where

$$\begin{aligned} \boldsymbol{\mu}_\mathbf{y} &= \sigma^2\mathbf{I}_{n+1} \cdot \mathbf{0} + \mathbf{0} \\ &= \mathbf{0}, \\ \boldsymbol{\Sigma}_\mathbf{y} &= \sigma^2\mathbf{I}_{n+1} + \mathbf{I}_{n+1} \cdot \mathbf{K}_{n+1} \cdot \mathbf{I}_{n+1}^T \\ &= \sigma^2\mathbf{I}_{n+1} + \mathbf{K}_{n+1}. \end{aligned} \tag{15.1.1}$$

With Theorem 13.3, we know that the conditional distribution of $f(y_{n+1}|\mathbf{y}_n)$ is also a Gaussian. Before we present the format that $f(y_{n+1}|\mathbf{y}_n) = \mathcal{N}(\mu_{y_{n+1}|\mathbf{y}_n}, \sigma^2_{y_{n+1}|\mathbf{y}_n})$ we first represent the information of the join Gaussian as

$$\boldsymbol{\mu}_\mathbf{y} = \begin{bmatrix} \mu_{y_{n+1}} \\ \boldsymbol{\mu}_{\mathbf{y}_n} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_\mathbf{y} = \begin{bmatrix} \mathbf{K}(\mathbf{x}_{n+1}, \mathbf{x}_{n+1}) + \sigma^2 & \mathbf{K}_{n+1,n} \\ \mathbf{K}_{n,n+1} & \mathbf{K}_{nn} + \sigma^2\mathbf{I}_n \end{bmatrix}^{-1}.$$

Using (13.2.7), we obtain the precision matrix of the covariance matrix as

$$\begin{aligned} \boldsymbol{\Lambda}_\mathbf{y} = \boldsymbol{\Sigma}_\mathbf{y}^{-1} &= \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}. \end{aligned}$$

Following (13.2.6) we have

$$\begin{aligned}
\mu_{y_{n+1}|\mathbf{y}_n} &= \sigma^2_{y_{n+1}|\mathbf{y}_n}(\sigma^2\mathbf{I})^{-1}(\mathbf{y}_n - \mathbf{I0} - \mathbf{0}) \\
&= \sigma^2(\sigma^2\mathbf{K}_{n+1}^{-1} + \mathbf{I})^{-1}\sigma^{-2}\mathbf{y}_n \\
&= (\sigma^2\mathbf{K}_{n+1}^{-1} + \mathbf{I})^{-1}\mathbf{y}_n, \\
\sigma^2_{y_{n+1}|\mathbf{y}_n} &= (\mathbf{K}_{n+1}^{-1} + \mathbf{I}^T(\sigma^2\mathbf{I})^{-1}\mathbf{I})^{-1} \\
&= (\mathbf{K}_{n+1}^{-1} + (\sigma^2\mathbf{I})^{-1})^{-1} \\
&= \sigma^2(\sigma^2\mathbf{K}_{n+1}^{-1} + \mathbf{I})^{-1}.
\end{aligned} \tag{15.1.2}$$

15.2 Connections to Ridge Regression

from the sequential relationship between $\mathbf{K_n}$ and $\mathbf{K_{n+1}}$, we can have $p(z_{n+1}) = \mathcal{N}(\mathbf{0}, \mathbf{K_{n+1}})$ and $p(Y_{n+1}|Z_{n+1}) = \mathcal{N}(Z_{N+1}, \sigma^2\mathbf{I})$.
From conditional Gaussian, we know that the joint distribution of $p(Y_{n+1}, Z_{n+1})$ is Gaussian, the marginal distribution of $p(y_{n+1})$ is a Gaussian with $\mu_{y_{n+1}} = 0$ and $\mathbf{S_{y_{n+1}}} = \sigma^2\mathbf{I} + \mathbf{K}_{n+1}$. And the update probability distribution of $p(y_{n+1}|y_n)$ is also a Gaussian,

$$\mu_{y_{n+1}|y_n} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) = K \cdot (K_n + \sigma^2\mathbf{I})^{-1}y_n$$

where $\mu_a = 0$, $\mu_b = 0$. $\Sigma_{aa} = K$.
Equivalently, we can write two expressions below,

$$\begin{aligned}
\mathbf{y}_{n+1} &= \mathbf{x}_{n+1}^T\mathbf{x}^T(\mathbf{x}\mathbf{x}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y}_n \\
\hat{\mathbf{y}}_{n+1} &= \mathbf{x}_{n+1}^T\beta \qquad \beta = (\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I})^{-1}\mathbf{x}^T\mathbf{y}
\end{aligned}$$

2. ***Claim:*** $(\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I})^{-1}\mathbf{x}^T = \mathbf{x}^T(\mathbf{x}\mathbf{x}^T + \lambda\mathbf{I})^{-1}$.
Hint: First left multiply $\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I}$ to make the inverse on the left hand side disappear, then right multiply $\mathbf{x}\mathbf{x}^T + \lambda\mathbf{I}$. The claim is obvious after you expand out the equation.

15.3 Connections to SVMs

15.4 Gaussian Processes for Classification

15.5 Bayesian Nonparametric and MLE

Chapter 16

Exchangeability and Subjective Probability

In the previous chapters we introduced the concept of prior distribution in a rather technical way in order to obtain the weighted area under curve for an estimator. We see that we are allowed to use any prior distribution to express our perception of where we should weight more or less in comparing risk functions for estimators. However the overall set up still follows the idea that data are generated by a distribution with a fixed (but unknown) parameter. With the observed data we want to have an estimator that is close to the true parameter as much as possible using a loss function. In addition the closeness may be evaluated if we obtain many sets of data or computing the expectation of the loss function.
In this chapter we start to deviate completely from the long-term interpretation of probability. We want to have a complete Bayesian treatment of probability and learning and show probability is a nice way (and somewhat only way) to express our "belief" of uncertainty.
In order to do so the main theoretic observation is based on two theorems: that of de Finetti's presentation theory and that of Cox's probability theorem. For de Finetti's presentation theory we do not present the most general format but focusing on the special case of Bernoulli random variable. The more general case using measure theory is presented in the Appendix.

16.1 Exchangeability vs. Independence

We are about to introduce a concept that may be considered as a generalization of the independence concept in probability. Let us first re-exam some of the properties of Bernoulli random variables. For one Bernoulli random variable let we have $\mathbb{P}(H) = \theta$. By probability axioms as discussed in Chapter 2 we have $\mathbb{P}(T) = 1-\theta$ since the sum of the probability of complementary events has to be 1. Let us consider two *i.i.d.* Bernoulli random variables. We have $\mathbb{P}(HT) = \mathbb{P}(TH) = \theta(1-\theta)$, $\mathbb{P}(HH) = \theta^2$ and $\mathbb{P}(TT) = (1-\theta)^2$. All the calculations are based on the concept of independence. In addition we have $\mathbb{P}(HT) = \sqrt{\mathbb{P}(HH)\mathbb{P}(TT)}$, although the requirement that $\mathbb{P}(HT) = \sqrt{\mathbb{P}(HH)\mathbb{P}(TT)}$ seems strong. An interesting question is that do we have a way to assign probability such that $\mathbb{P}(HT) = \mathbb{P}(TH)$ (due to symmetry) but does not necessarily have the algebraic structure such that $\mathbb{P}(HT) = \sqrt{\mathbb{P}(HH)\mathbb{P}(TT)}$?
Let us consider a prior distribution for θ. We choose uniform distribution between 0 and 1 for simplicity. Simple calculation shows that $\mathbb{P}(H) = \int f(\theta)\mathbb{P}(X = H|\theta)d\theta = \int \theta d\theta = \frac{1}{2}$. $\mathbb{P}(T) = 1 - \mathbb{P}(H) = \frac{1}{2}$. Further calculation shows that $\mathbb{P}(HH) = \int \theta^2 d\theta = \frac{1}{3}$. $\mathbb{P}(TT) = \int (1-\theta)^2 d\theta = \frac{1}{3}$. $\mathbb{P}(HT) = \mathbb{P}(TH) = \int \theta(1-\theta)d\theta = \frac{1}{6}$. Clearly in this case we still have $\mathbb{P}(HT) = \mathbb{P}(TH)$ but we do

not have $\mathbb{P}(HT) = \sqrt{\mathbb{P}(HH)\mathbb{P}(TT)}$. In addition we notice that if we change the prior distribution $\mathbb{P}(H)$ may change accordingly.
Is this result reasonable? If we adopt the long-term frequency interpretation of probability we tend to believe that $\mathbb{P}(H)$ is a fixed but unknown parameter that is associated with the experiment settings, a.k.a the dice (or the dice and the way that the dice is casted). However imaging that we are performing a though experiments where we believe θ could take many values with different probability. If we consider all the possible values that θ may take, the calculation that we just did starts to make sense. In addition we notice that the calculation follows the axioms of probability and it satisfies our intuition where the event "HT" and "TH" should have the same probability.
Below we present general discussions regarding why such probability assignment is reasonable. In addition we show that in order to keep the property such that $\mathbb{P}(HT) = \mathbb{P}(TH)$ the *only* possible approach is through a prior distribution. We first introduce the concept of exchangeability to formalize our intuition of symmetry and to show that exchangeability is a natural extension of *i.i.d.* .
A *permutation function* is a one-to-one function maps integers from 1 to N $(1 \leq N)$ to 1 to N. Below we define exchangeability of random variables.

Definition 16.1 (Exchangeability)**.** Let $\pi(\dot{})$ be a permutation function and let $X_1, X_2, ..., X_n$ be random variables. $X_1, X_2, ..., X_n$ are *exchangeable* according to the probability function $\mathbb{P}()$ if:

$$\mathbb{P}(X_1, X_2, ..., X_n) = \mathbb{P}(X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(n)})$$

for any permutation function π.

Example 16.1. For Bernoulli random variables X_1 and X_2 they are exchangeable if we have:

$$\mathbb{P}(X_1, X_2) = \mathbb{P}(X_2, X_1),$$

or $\mathbb{P}$ (HT) = $\mathbb{P}$ (TH).

Example 16.2. For Bernoulli random variables X_1, X_2, X_3 they are exchangeable if we have:

$$\mathbb{P}(X_1, X_2, X_3) = \mathbb{P}(X_1, X_3, X_2) = \mathbb{P}(X_2, X_1, X_3) = \mathbb{P}(X_2, X_3, X_1) = \mathbb{P}(X_3, X_1, X_2) = \mathbb{P}(X_3, X_2, X_1)$$

or $\mathbb{P}$ (HHT) = $\mathbb{P}$ (HTH)=$\mathbb{P}$ (THH) and $\mathbb{P}$ (HTT) = $\mathbb{P}$ (THT)=$\mathbb{P}$ (TTH).

Example 16.3. Given three exchangeable random variables X_1, X_2, X_3 we have

$$\begin{aligned}
\mathbb{P}(X_1, X_2) &= \int \mathbb{P}(X_1, X_2, X_3) d(X_3) \\
\mathbb{P}(X_1, X_3) &= \int \mathbb{P}(X_1, X_2, X_3) d(X_2) \\
&= \int \mathbb{P}(X_1, X_3, X_2) d(X_2) \\
&= \mathbb{P}(X_1, X_2)
\end{aligned}$$

Clearly for a set of exchangeable random variables any subset of the set is also exchangeable.

Example 16.4. Given three exchangeable random variables X_1, X_2, X_3 we have

$$\begin{aligned}\mathbb{P}(X_1, X_2|X_3) =& \frac{\mathbb{P}(X_1, X_2, X_3)}{\mathbb{P}(X_3)} \\ \mathbb{P}(X_2, X_1|X_3) =& \frac{\mathbb{P}(X_2, X_1, X_3)}{\mathbb{P}(X_3)} \\ =& \frac{\mathbb{P}(X_1, X_2, X_3)}{\mathbb{P}(X_3)} \\ =& \mathbb{P}(X_1, X_2|X_3)\end{aligned}$$

So with three exchangeable random variables, any two random variables conditional on the third one are still exchangeable

16.2 de Finetti's Probability Representtion Theory

Theorem 16.1 (de Finetti's Probability Representtion Theory)**.** *For a group of exchangeable Bernoulli random variable $X_1, \cdots, X_n$, any valid expression of the join distribution must adopt the format that*

$$\mathbb{P}(X_1, \cdots, X_n) = \int_0^1 \theta^{S_n}(1-\theta)^{n-S_n} dQ(\theta), \tag{16.2.1}$$

where $Q(\theta) = \lim_{N\to\infty} \mathbb{P}(S_N/N \leq \theta)$ is the cumulative distribution function. $S_m = \sum_{i=1}^m X_i$.

Proof. We denote π as a permutation function, With exchangeability we have

$$\begin{aligned}\mathbb{P}(X_1 + X_2 + \cdots + X_n = y_n) =& \mathbb{P}(\sum_{i=1}^n X_i = y_n) \\ =& \binom{n}{y_n} \mathbb{P}(X_{\pi(1)}, \cdots, X_{\pi(n)})\end{aligned} \tag{16.2.2}$$

For example, let $n = 2$, $y_n = 1$, we have

$$\begin{aligned}\mathbb{P}(X_1 + X_2 = 1) =& \mathbb{P}(1,0) + \mathbb{P}(0,1) \\ =& \binom{2}{1}\mathbb{P}(1,0) \\ =& \binom{2}{1}\mathbb{P}(0,1).\end{aligned}$$

Similarly let $n = 3$, $y_n = 2$, we have

$$\begin{aligned}\mathbb{P}(X_1 + X_2 = 2) =& \mathbb{P}(1,1) \\ =& \binom{2}{2}\mathbb{P}(1,1).\end{aligned}$$

Let us perform a "thought" experiment. Let's imagine that we have one more random variable X_3 that is exchangeable with X_1, X_2. Let $N = 3$ and with the symbols Y_N defined before we have

$$\begin{aligned}\mathbb{P}(X_1 + X_2 = 1) =& \mathbb{P}(1,0,0) + \mathbb{P}(0,1,0) + \mathbb{P}(1,0,1) + \mathbb{P}(0,1,1) \\ =& \mathbb{P}(X_1 + X_2 = 1|X_1 + X_2 + X_3 = 1) \cdot \mathbb{P}(X_1 + X_2 + X_3 = 1) + \\ & \mathbb{P}(X_1 + X_2 = 1|X_1 + X_2 + X_3 = 2) \cdot \mathbb{P}(X_1 + X_2 + X_3 = 2)\end{aligned} \tag{16.2.3}$$

As we studied in Chapter 2 for the classical probability for this conditional probability we have the hypergeometric distribution as

$$\mathbb{P}(X_1+X_2=1|X_1+X_2+X_3=1)=\frac{\binom{1}{1}\binom{2}{1}}{\binom{3}{2}}=\frac{2}{3} \tag{16.2.4}$$

Similarly we have

$$\mathbb{P}(X_1+X_2=1|X_1+X_2+X_3=2)=\frac{\binom{2}{1}\binom{1}{1}}{\binom{3}{2}}=\frac{2}{3} \tag{16.2.5}$$

Putting (16.2.4) and (16.2.5) back to (16.2.3) we have

$$\mathbb{P}(X_1+X_2=1)=\frac{2}{3}\mathbb{P}(X_1+X_2+X_3=1)+\frac{2}{3}\mathbb{P}(X_1+X_2+X_3=2)$$

If we continue the thought experiments to extend n exchangeable random variable to N such ones we obtain the following relationship

$$\begin{aligned}
&\mathbb{P}(X_1+X_2+\cdots+X_n=y_n)\\
=&\mathbb{P}(\sum_{i=1}^{n}X_i=y_n)\\
=&\sum_{y_N=y_n}^{N-(n-y_n)}\mathbb{P}(\sum_{i=1}^{n}X_i=y_n|\sum_{i=1}^{N}X_i=y_N)\cdot\mathbb{P}(\sum_{i=1}^{N}X_i=y_N)\\
=&\sum_{y_N=y_n}^{N-(n-y_n)}\frac{\binom{y_N}{y_n}\binom{N-y_N}{n-y_n}}{\binom{N}{n}}\mathbb{P}(\sum_{i=1}^{N}X_i=y_N)\\
=&\frac{n!}{y_n!(n-y_n)!}\sum_{y_N=y_n}^{N-(n-y_n)}\frac{y_N\cdot(y_N-1)\cdots(y_N-y_n+1)\cdot(N-y_n)\cdot(N-y_n-1)\cdot(N-y_n-(n-1))}{N(N-1)\cdots(N-n+1)}\\
=&\binom{n}{y_n}\sum_{y_N}\frac{\theta N(\theta N-1)\cdots(\theta N-y_n+1)(1-\theta)N((1-\theta)N-1)\cdots((1-\theta)N-n-y_n+1)}{N(N-1)\cdots(N-n+1)},
\end{aligned} \tag{16.2.6}$$

where $\theta=\frac{y_N}{N}$, $y_N=\theta N$, and $N-y_N=(1-\theta)N$.
Let us continue the thought experiments and let $N\to\infty$, summation becomes integral and we have

$$\mathbb{P}(X_1+X_2+\cdots+X_n=y_n)=\binom{n}{y_n}\int\theta^{y_n}(1-\theta)^{n-y_n}d(Q(\theta)), \tag{16.2.7}$$

where $Q(\theta)$ is defined before.
To obtain (16.2.7) note that $\frac{\theta N-y_n+1}{N}\le a_N\le\frac{\theta N}{N-n+1}$ and $\frac{(1-\theta)N-n-y_n+1}{N}\le b_N\le\frac{(1-\theta)N}{N-n+1}$. Let $N\to\infty$ we have $\theta\le a_N\le\theta$ and $1-\theta\le b_N\le 1-\theta$. Hence we have $\lim_{N\to\infty}a_N=\theta, b_N=1-\theta$. Plug the results to (16.2.6) we have (16.2.7).
Comparing (16.2.2) and (16.2.7) we have the proof completed. □
The interpretation of theorem (16.1) is quite interesting. This theorem tells us that if a group of Bernoulli random variables are exchangeable the joint distribution MUST follow (16.2.1) where $Q(\theta)$ can be any accumulative probability distribution selected by subjective judgement. This justifies the usage of prior distribution in Bayesian statistics.

16.3 Cox's Theorem

16.4 Exercises

16.1 Derive similar results for exchangeable Poisson random variables

Appendix A

Real Numbers and Number Systems

A.1 Natural number

Definition: $\mathbb{N} = \{1, 2, \dots, \}$.

Operation defined: $+$, $*$.

A.2 Integers

Definition: $\mathbb{Z} = \{[(a, b)]\}$, $a, b \in \mathbb{N}$. $(a_1, b_1) \sim (a_2, b_2)$ if $a_1 + b_2 = a_2 + b_1$.
For example,

$$
\begin{aligned}
&(1,1) \sim (2,2) \quad \checkmark \\
&(1,1) \sim (2,3) \quad \times \\
&(1,2) \sim (2,3) \quad \checkmark \\
&(2,1) \sim (3,2) \sim (4,3). \quad \checkmark
\end{aligned}
$$

Operation defined: $+$, $-$, $*$.

$$
\begin{aligned}
&Z_1 = (a, b) \quad Z_2 = (c, d) \\
&Z_1 + Z_2 = (a + c, b + d) \\
&Z_1 * Z_2 = (ac + bd, ad + bc) \\
&Z_1 - Z_2 = (a + d, b + c)
\end{aligned}
$$

For example,

$$
\begin{aligned}
&(1,1) + (2,1) = (3,2) = (2,1) \\
&(1,2) + (2,1) = (3,3) = (2,2) = (1,1) \\
&(1,2) * (1,2) = (5,4) = (2,1) \\
&(2,1) * (2,1) = (5,4) = (2,1)
\end{aligned}
$$

Note that $(1, 1) = 0$, $(1, 2) = -1$, $(2, 1) = 1$.

A.3 Rational number

Definition: $\mathbb{Q} = \{[(a,b)]\}$ $a, b \in \mathbb{Z}, b \neq 0$. $\quad (a,b) \sim (c,d)$ if $ad = bc$.
For example,

$$(1,3) \sim (2,6) \sim (3,9) \ \checkmark$$
$$(1,3) \sim (1,2) \ \times$$

Operation defined: $+,\ -,\ *,\ /$.

$$Z_1 = (a,b) \quad Z_2 = (c,d)$$
$$Z_1 + Z_2 = (ad + bc, bd)$$
$$Z_1 - Z_2 = (ad - bc, bd)$$
$$Z_1 * Z_2 = (ac, bd)$$
$$Z_1 \ / \ Z_2 = (ad, bc) \quad (c \neq 0)$$

Cresis! A sequence of rational numbers converge to an irrational number.

$$x_{n+1} = \frac{1}{2}(x_n + \frac{2}{x_n})$$
$$x_0 = 2,$$
$$x_1 = 1.5,$$
$$x_2 = 1.416,$$
$$x_3 = 1.4142156,$$
$$\sqrt{2} = 1.4142135.$$

Claims:
1. $x_n \geq 0$
2. $x_n \geq \sqrt{2}$

$$\begin{aligned} x_{n+1} &= \frac{1}{2}(x_n + \frac{2}{x_n}) \\ &\geq \frac{1}{2} \cdot 2 \cdot \sqrt{x_n \cdot \frac{2}{x_n}} \\ &= \sqrt{2} \end{aligned}$$

3. x_n is rational
4. x_n is monotonic decreasing

$$x_{n+1} - x_n = \frac{1}{x_n} - \frac{1}{2}x_n = \frac{2 - x_n^2}{2x_n} \leq 0$$

5. $\lim_{n\to\infty} x_n = \sqrt{2}$ ($\sqrt{2}$ is not a rational number).

$$x_\infty = \frac{1}{2}(x_\infty + \frac{2}{x_\infty})$$

A.4 Real number

1. Definition(Cauchy sequence):
$(x_n)_{n=1}^{\infty}, x_n \in \mathbb{Q}$ is a Cauchy sequence if for every $\epsilon > 0$, $\exists\ n_\epsilon$ *s.t.*

$$|x_m - x_{m'}| < \epsilon, \quad \text{for all } m, m' \geq n_\epsilon.$$

For example, $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$ is a Cauchy sequence.

2. Definition(Real number):
Two Cauchy sequences $S = (S_n)_{n=1}^{\infty}$, $t = (t_n)_{n=1}^{\infty}$, $S \sim t$ if $S - t \to 0$.
$R = \{[S]\}$, S is a rational Cauchy sequence. $[S]$ is the equivalent class defined by " $\sim$" on rational Cauchy sequence.
Operation defined: $+,\ -,\ *,\ /$.

$$\begin{aligned} S + t &= (S_n + t_n)_1^{\infty} \\ S - t &= (S_n - t_n)_1^{\infty} \\ S * t &= (S_n * t_n)_1^{\infty} \\ S/t &= (S_n/t_n)_m^{\infty} \end{aligned}$$

3. Given $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_n > \cdots > 0$, *s.t.* $\lim_{n\to\infty} \epsilon_n = 0$. Choose $S^{1'} = S_1^{1'}, S_2^{1'}, \ldots, S_n^{1'}$ for S^1, *s.t.* for ϵ_i,

$$|S_m^{1'} - S_{m'}^{1'}| \leq \epsilon_i, \quad \text{for all } m, m' \geq i.$$

4. **Claim:**

- t is a rational Cauchy sequence
- $\lim_{n'\to\infty} S^{n'} = t$.

Proof sketch: 1. t is a rational sequence for sure. For Cauchy sequence, for any $\epsilon > 0$,

$$\begin{aligned} &\text{Find } n_1 \ \ s.t. \ \ \epsilon_{n_1} \leq \frac{\epsilon}{3} \\ &\text{Find } n_2 \ \ s.t. \ \ |S^{m_1'} - S^{m_2'}| \leq \frac{\epsilon}{3} \\ &n = \max(n_1, n_2), \end{aligned}$$

$$\begin{aligned} |S_{m_1}^{m_1} - S_{m_2}^{m_2}| &= |S_{m_1}^{m_1} - S^{m_1} + S^{m_1} - S^{m_2} + S^{m_2} - S_{m_2}^{m_2}| \\ &\leq |S_{m_1}^{m_1} - S^{m_1}| + |S^{m_1} - S^{m_2}| + |S^{m_2} - S_{m_2}^{m_2}| \\ &< \epsilon \qquad \text{for all } m_1, m_2 \geq n. \end{aligned}$$

5. Definition(Real Cauchy sequence): $S = (S_n)_{n=1}^{\infty}$, $S_n \in R$. A sequence of real number is a Cauchy sequence if for any $\epsilon > 0$, $\exists\ n_\epsilon$ *s.t.*

$$|S_m - S_{m'}| \leq \epsilon, \quad \text{for all } m, m' \geq n_\epsilon.$$

6. **Claims:**

- Any real Cauchy sequence converges to a real number. (Completeness of $\mathbb{R}$)

Appendix B

Abstract Vector Spaces

Here we present a rather abstract definition of vector spaces. The main purpose is to provide a uniform treatment of vectors, matrices, and (to some extend) functions under this treatment. To do so we first introduce the formal definition of field as below.

B.1 Field

Definition B.1 (Field)**.** A *field* is a set $\{F\}$ with two operations $+$, $*$, such that

$$a + b \in \{F\}, \qquad a * b \in \{F\} \quad \text{for all} \quad a, b \in \{F\}.$$

For the addition operation '+', we have:

$$\begin{aligned}
&(a + b) + c = a + (b + c),\\
&a + b = b + a,\\
&\exists\ 0_{\{F\}} \in \{F\} \quad s.t. \quad a + 0_{\{F\}} = 0_{\{F\}} + a = a,\\
&\exists\ c \in \{F\} \quad s.t. \quad a + c = c + a = 0.
\end{aligned}$$

For the multiplication operation $+$, we have:

$$\begin{aligned}
&(a * b) * c = a * (b * c),\\
&a * b = b * a,\\
&\exists\ 1_{\{F\}} \quad s.t. \quad a * 1_{\{F\}} = 1_{\{F\}} * a = a,\\
&\exists\ c\ \forall\ a \neq 0 \quad s.t. \quad a * c = c * a = 1,\\
&\forall\ a, b, c \in \{F\} \quad (a + b) * c = a * c + b * c,
\end{aligned}$$

B.2 Vector Space

We introduce an abstract definition of vectors in the following discussion. The definition, known as the *abstract vector space*, extends our notation of vectors as directed line segments in an Euclidian space to ones that include linear transformations and continuous functions. There are multiple reasons for this extension. The most important one is that we want to adopt a rather unified treatment of vector, matrices, and functions. Such unified treatment greatly simplifies the work when we introduce important learning concepts such as the Support Vector Machines.

Definition B.2 (Abstract Vector Space)**.** A set $\{V\}$ is a vector space for a field $\{F\}$ if there exists two binary operations $+,\ *$, where $+:\ \{V\}\times\{V\}\to\{V\}$ and $*:\ \{F\}\times\{V\}\to\{V\}$ such that

$$\begin{aligned}
&\forall \mathbf{u},\mathbf{v}\in\{V\},\\
&\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u},\\
&\exists\ 0_{\{V\}}\ \ s.t.\ \ \mathbf{u}+0_{\{V\}}=0_{\{V\}}+\mathbf{u}=\mathbf{u},\\
&\forall\ \mathbf{u}\in\{V\},\ \exists\ \mathbf{v}\ \ s.t.\ \ \mathbf{u}+\mathbf{v}=0_{\{V\}},\\
&\forall \alpha,\beta\in\{F\},\mathbf{u},\mathbf{v}\in\{V\},\\
&\alpha(\mathbf{u}+\mathbf{v})=\alpha\mathbf{u}+\alpha\mathbf{v},\\
&(\alpha+\beta)\mathbf{u}=\alpha\mathbf{u}+\beta\mathbf{u},\\
&(\alpha\cdot\beta)\mathbf{u}=\alpha\cdot(\beta\mathbf{u}),\\
&\forall\ \mathbf{u}\ \ 1\cdot\mathbf{u}=\mathbf{u},
\end{aligned}\tag{B.2.1}$$

where the symbol $\forall\ x$ reads as for all x and the symbol $\exists\ x$ means there exists an x.

Example B.1. Examples of vector space:
(1) $\mathbf{V}=\mathbb{R}^p$, all p-dimension vectors in $\mathbb{R}$.
(2) $\mathbf{V}=\{(\mathbf{X})_{m\times n}\}$, all $m\times n$ matrices in $\mathbb{R}^{m\times n}$.

For functions $g:\mathbb{R}^p\to\mathbb{R}$ we introduce two generic operations $+$ and $*$ as

$$\begin{aligned}
(g_1+g_2)(\mathbf{x}) &:= g_1(\mathbf{x})+g_2(\mathbf{x}),\\
(\alpha g)(\mathbf{x}) &:= \alpha\cdot g(\mathbf{x}).
\end{aligned}$$

With the definition we immediately have

$$\begin{aligned}
&g_1+g_2=g_2+g_1,\\
&\alpha(g_1+g_2)=\alpha g_1+\alpha g_2,\\
&(\alpha+\beta)g=\alpha g+\beta g,\\
&(\alpha\cdot\beta)g=\alpha\cdot(\beta g),\\
&\forall\ g\ \ 1\cdot g=g,
\end{aligned}\tag{B.2.2}$$

$$\begin{aligned}
&\exists 0_v, s.t., g+0_v=0_v+g=g,\\
&\text{For any} g,\exists g_-, s.t. g+g_-=g_{-1}+g=0_v.
\end{aligned}\tag{B.2.3}$$

Although promising the fact cited above does not guarantee that an arbitrary set of functions is a vector space, with some restrictions we are able to establish that some of the sets are. See examples below.

Example B.2. With the definition it is easy to show that the following two examples are vector spaces.

- All linear functional of $\mathbb{R}^p$. $\mathcal{F}=\{g|g:\mathbb{R}^p\to\mathbb{R}$ and g is linear$\}$.
- $\mathcal{I}=[0,1]$, $P^n=\{g|g$ is a polynomial of degree up to n defined on $\mathcal{I}\ \}$.

B.3 Vector Norms

A very important concept of vector is its norm.

B.3.1 ℓ_p Norm of Vectors

Definition B.3 (ℓ_p norm of vectors)**.** Given a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, the ℓ_p norm of $\mathbf{x}$ is

$$\ell_p \text{ norm: } ||\mathbf{x}||_p := (\sum_{i=1}^{n} x_i^p)^{\frac{1}{p}}. \tag{B.3.1}$$

Example B.3. A few commonly used norm of vectors for $p \in \{0, 1, 2, \infty\}$ are listed below.

$$\begin{aligned}
\ell_0 \text{ norm: } & ||\mathbf{x}||_0 := \sum_{i=1}^{n} (x_i \neq 0) \\
\ell_1 \text{ norm: } & ||\mathbf{x}||_1 := \sum_{i=1}^{n} |x_i| \\
\ell_2 \text{ norm: } & ||\mathbf{x}||_2 := \sqrt{\sum_{i=1}^{n} x_i^2} \\
\ell_\infty \text{ norm: } & ||\mathbf{x}||_\infty := \max |x_i|
\end{aligned}$$

B.3.2 Norm of Matrices

Definition B.4. Given a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, and a vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_r)^T$ are the singular values of X ordered in a non-increasing order such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$

$$\begin{aligned}
&\text{Trace Norm: } ||\mathbf{X}||_* := \text{tr}(\sqrt{X^T X}) = \sum_{i=1}^{r} \lambda_i = ||\boldsymbol{\lambda}||_1 \\
&\text{Frobenius Norm: } ||\mathbf{X}||_F := \sqrt{\text{tr}(X^T X)} = \sqrt{\sum_{i=1}^{r} \lambda_i^2} = ||\boldsymbol{\lambda}||_2 \\
&\text{Spectral Norm: } ||\mathbf{X}||_2 := \sqrt{\lambda_{\max}(X^T X)} = \lambda_1 = ||\boldsymbol{\lambda}||_\infty
\end{aligned} \tag{B.3.2}$$

Appendix C

Function Spaces

Following the discussion of abstract vector spaces, we show that the following sets of functions have the same linear algebra property.

C.1 Function Spaces as Abstract Vector Space

Definition C.1 (Function Spaces)**.** Let $g : \mathcal{I} \to \mathbb{R}$. $\mathcal{I}$ is a closed interval. An example of such intervals is $\mathcal{I} = [0, 1]$ which we alway use for illustration. $g \in C^n[\mathcal{I}]$ if g^n is defined on $\mathcal{I}$ where g^n is the n_{th} deferential of g.

Example C.1. With the definition we explain three commonly used symbols as

- $C^0[\mathcal{I}]$: all continuous function defined on $\mathcal{I}$
- $C^1[\mathcal{I}]$: all continuous and differentiable function defined on $\mathcal{I}$
- $C^2[\mathcal{I}]$: all continuous, differentiable, and second-order differentiable function defined on $\mathcal{I}$

It is easy to show that linear combination of continuous function is still continuous and scaling of a continuous functions is still continuous. So the operation $+$ and $*$ are well defined on C^0. It is quite easy to extend the observation to C^n. In addition we can show that C^n is a vector space.

Clearly for the previous examples we have

- $g_1 + g_2 = g_2 + g_1$.
- $\exists 0_v, s.t.,\ g + 0_v = 0_v + g = g$.
- For any g, $\exists g_-, s.t.,\ g + g_- = g_{-1} + g = 0_v$.

C.2 Function convergence

Definition C.2 (Point-wise convergence)**.** Given a sequence of functions $F = (f_1, f_2, \cdots, f_n)$, the sequence converges to a function g point-wisely if

$$\lim_{n\to\infty} f_n(x) = f(x) \quad \text{for all} \quad x \in \mathcal{I}.$$

Example C.2. Let a sequence of function defined as $f_n(x) = x^n$ defined on an interval $\mathcal{I} = [0, 1]$. Let

$$g(x) = \begin{cases} 0, & \text{if } \ x \neq 1 \\ 1, & \text{other wise.} \end{cases}$$

We have f_n converges to g point-wisely.

Example C.3. $q_{n+1} = \frac{1}{2}(x^2 + 2q_n - q_n^2)$ is a system of functions defined on an interval $\mathcal{I} = [-1, 1]$. $q_0 = \mathbf{1}$. We have several **claims**:

- q_n is a polynomial.
- $q_n \geq |x|$.
  Proof sketch: We know $q_0 \geq |x|$, for $x \in \mathcal{I}$. Let $q_n \geq |x|$, then

$$\begin{aligned} q_{n+1} &= \frac{1}{2}x^2 + q_n - \frac{1}{2}q_n^2 \\ q_{n+1} - |x| &= (q_n - |x|) - \frac{1}{2}(q_n^2 - x^2) \\ &= \frac{1}{2}(q_n - |x|)[(1 - q_n) + (1 - |x|)] \\ &\geq 0 \end{aligned}$$

- $q_n \geq q_{n+1}$

$$q_{n+1} - q_n = \frac{1}{2}x^2 - \frac{1}{2}q_n^2 \leq 0$$

- $1 \geq q_n$.
- $|x| = q_\infty(x)$

$$\begin{aligned} q_{n+1}(x) &= \frac{1}{2}x^2 + q_n(x) - \frac{1}{2}q_n^2(x^2) \\ n \to \infty &\Rightarrow \frac{1}{2}x^2 = \frac{1}{2}q_n^2(x) \\ &\Rightarrow |x| = q_n(x) \end{aligned}$$

(2) **Claim:** $p(\mathcal{I})$ is the set of polynomial functions including a constant function on a closed interval $\mathcal{I}$, any $f \in C^0(\mathcal{I})$ can be approximated by a sequence of functions in $p(\mathcal{I})$.

Definition C.3 (Converges with probability and Converges almostly surely)**.**

$$\lim_{n\to\infty} \mathbb{P}(|f_n(x) - f(x)| > \epsilon) = 0 \quad \text{for } \ x \in \mathcal{I}.$$

Converge almost surely

$$p(\lim_{n\to\infty} f_n(x) \neq f(x)) = 0 \quad \text{for } \ x \in \mathcal{I}.$$

C.3 Reproducing Hilbert Spaces

Mercer' Kernel

Appendix D

Advanced Probability and SLLN

Before the conclusion of the book we want to present a rigorous treatment of probability, random variable, and convergences of random variable. The culminating point of the discussion is the presentation and proof of strong law of large numbers. With the stronger version of convergence theory we revisit the convergence of Bayesian posterior distribution and Bayesian estimators.

D.1 Measure Theory

D.1.1 Measurable Sets and Events

Definition D.1 (σ-Algebra)**.** Let Ω be a set. An *algebra* is a collection of subsets $\mathcal{A}$ of $\{\Omega\}$ such that

$$\begin{aligned}&\emptyset \in \mathcal{A}\\ &\text{if } \{A\} \in \mathcal{A}, \{A\}^c \in \mathcal{A}\\ &\text{if } \{A\}, \{B\} \in \mathcal{A}, \{A\} \cup \{B\} \in \mathcal{A}\end{aligned}$$

An algebra $\mathcal{A}$ is a *σ-algebra* if it has the property such that

$$\text{if } \{A\}_1, \{A\}_2, \cdots \in \mathcal{A}, \bigcup_{i=1}^{\infty} \{A\}_i \in \mathcal{A}.$$

A *measurable space* $(\Omega, \mathcal{A})$ is a set Ω together with a σ-algebra $\mathcal{A}$ on the set. Each $\{A\} \in \mathcal{A}$ is called a *measurable set* or an *event* in the context of probability theory.

Claim D.1. *If $\mathcal{A}$ is an algebra, then $\{\Omega\} \in \mathcal{A}$.*

Proof. Left as exercise. □

Claim D.2. *For an algebra $\mathcal{A}$, if $\{A\}, \{B\} \in \mathcal{A}$, then $\{A\} \cap \{B\} \in \mathcal{A}$.*

Proof. Left as exercise. □

Claim D.3. *For a σ-algebra $\mathcal{A}$, if $\{A\}_1, \{A\}_2, \ldots$ are in $\mathcal{A}$, then $\bigcap_{i=1}^{\infty} \{A\}_i \in \mathcal{A}$.*

Proof. Left as exercise. □

Definition D.2. Let $\mathcal{C}$ be any collection of subsets of $\{\Omega\}$. $\sigma(\mathcal{C})$ is called the *σ-algebra generated by $\mathcal{C}$* and is the smallest σ-algebra that contains $\mathcal{C}$.

Definition D.3 (Borel σ-algebra)**.** Let $\mathcal{T}$ be the collection of open sets of $\mathbb{R}$. The *Borel σ-algebra* on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra that contains $\mathcal{T}$, i.e. $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{T})$.

Claim D.4. *The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is also generated from the following collection of sets:*

1. $\{(a, b) : a, b \in \mathbb{R}\}$,
2. $\{[a, b] : a, b \in \mathbb{R}\}$,
3. $\{(a, b] : a, b \in \mathbb{R}\}$,
4. $\{(-\infty, b] : b \in \mathbb{R}\}$,

Proof. Left as exercise. □

D.1.2 Measures and Probability

Definition D.4 (Measure)**.** Let $\{\Omega\}$ be a set, and let $\mathcal{A}$ be a σ-algebra on $\{\Omega\}$. A *measure* on $\mathcal{A}$ is a function $\mu : \mathcal{A} \to [0, \infty]$ such that

1. $\mu(\emptyset) = 0$.
2. If $\{A\}_1, \{A\}_2, \ldots$ are disjoint sets in $\mathcal{A}$, then $\mu\left(\bigcup_{i=1}^{\infty} \{A\}_i\right) = \sum_{i=1}^{\infty} \mu\left(\{A\}_i\right)$.

μ is a *probability measure* if in addition

3. $\mu(\{\Omega\}) = 1$.

The triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. The range of μ is $[0, \infty]$. Including ∞ in $\mathbb{R}$ is called the extended real numbers. For probability we restrict the range of μ to $[0, 1]$ and we call such measure a *probability measure*. If we use a probability measure in a measure space, we have a *probability space*. In other words, a probability space is a measure space $(\Omega, \mathcal{A}, \mu)$ where the range of μ is $[0, 1]$.

D.1.3 Measurable Functions and Random Variables

Definition D.5 (Measurable Function)**.** Let $(\{\Omega\}_1, \mathcal{A}_1)$ and $(\{\Omega\}_2, \mathcal{A}_2)$ be measurable spaces and let $f : \{\Omega\}_1 \to \{\Omega\}_2$. f is a *measurable function* if $f^{-1}(\{A\}) \in \mathcal{A}_1$ for each $A \in \mathcal{A}_2$. In particular, if $(\{\Omega\}_2, \mathcal{A}_2) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, f is called a *Borel measurable function.*

Claim D.5. *Let $(\{\Omega\}, \mathcal{A})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be measurable spaces and let $f : \{\Omega\} \to \mathbb{R}$. The follow conditions are equivalent.*

1. *f is Borel measurable.*
2. *$\{\omega : f(\omega) > a\} \in \mathcal{A}$ for all $a \in \mathbb{R}$.*
3. *$\{\omega : f(\omega) \leq a\} \in \mathcal{A}$ for all $a \in \mathbb{R}$.*
4. *$\{\omega : f(\omega) < a\} \in \mathcal{A}$ for all $a \in \mathbb{R}$.*
5. *$\{\omega : f(\omega) \geq a\} \in \mathcal{A}$ for all $a \in \mathbb{R}$.*

Proof. Left as exercise. □

Definition D.6 (Random Variable)**.** A *random variable* X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a Borel measurable function $X : \{\Omega\} \to \mathbb{R}$.

Definition D.7 (Induced Probability Measure)**.** Let X be a random variable on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The *probability measure induced* by X is the probability measure $\mathbb{P}_X$ on $\mathcal{B}(\mathbb{R})$ given by

$$\mathbb{P}_X(\{B\}) = \mathbb{P}(\{\omega : X(\omega) \in \{B\}\}), \quad \{B\} \in \mathcal{B}(\mathbb{R}).$$

D.2 Types of Convergence

Since a random variable is a function the convergence of a random variable should follow the discussion that we presented in the Chapter **??**. These convergences include point-wise convergence, convergence in distance, convergence in measure, and convergence almost everywhere.

D.2.1 Convergence of Events

Definition D.8 (Set Limit)**.** Let $\{\{A\}_n\}_{n=1}^{\infty}$ be a sequence of subsets of a set Ω. The *upper limit* of the sequence $\{\{A\}_n\}_{n=1}^{\infty}$ is

$$\limsup_{n\to\infty} \{A\}_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{A\}_k,$$

and the *lower limit* of the sequence $\{\{A\}_n\}_{n=1}^{\infty}$ is

$$\liminf_{n\to\infty} \{A\}_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \{A\}_k.$$

Furthermore, if

$$\limsup_{n\to\infty} \{A\}_n = \liminf_{n\to\infty} \{A\}_n = \{A\},$$

then $\{A\}$ is said to be the *limit* of the sequence $\{\{A\}_n\}_{n=1}^{\infty}$ and is denoted

$$\lim_{n\to\infty} \{A\}_n = \{A\}.$$

Definition D.9. Let $\{\{A\}_n\}_{n=1}^{\infty}$ be a sequence of subsets of a set Ω. We say $\{\{A\}_n\}_{n=1}^{\infty}$ form an *increasing* sequence of sets if $\{A\}_1 \subset \{A\}_2 \subset \dots$ and form a *decreasing* sequence of sets if $\{A\}_1 \supset \{A\}_2 \supset \dots$.

Claim D.6. *An increasing (or decreasing) sequence of events in a σ-algebra always converge to an event. In fact, if $\{\{A\}_n\}_{n=1}^{\infty}$ is increasing, then*

$$\lim_{n\to\infty} \{A\}_n = \bigcup_{n=1}^{\infty} \{A\}_n,$$

and if $\{\{A\}_n\}_{n=1}^{\infty}$ is decreasing, then

$$\lim_{n\to\infty} \{A\}_n = \bigcap_{n=1}^{\infty} \{A\}_n,$$

Proof. Left as exercise. □

Proposition D.1. *Given a probability space $(\{\Omega\}, \mathcal{A}, p)$, let $\{\{A\}_n\}_{n=1}^{\infty}$ be an increasing (or decreasing) sequence of events in $\mathcal{A}$, then*

$$\lim_{n\to\infty} p(\{A\}_n) = p\left(\lim_{n\to\infty} \{A\}_n\right).$$

Proof. Suppose $\{\{A\}_n\}_{n=1}^{\infty}$ is an increasing sequence of events. Let $\{B\}_1 = \{A\}_1$ and $\{B\}_i = \{A\}_i - \{A\}_{i-1}$ for each $i \geq 2$. Observe that the sequence of events $\{\{B\}_n\}_{n=1}^{\infty}$ are pairwise disjoint,

$$A_n = \bigcup_{i=1}^{n} \{B\}_i, \quad \text{and} \quad \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} \{B\}_i.$$

Since p is a measure on $\mathcal{A}$,

$$\mathbb{P}(\{A\}_n) = \mathbb{P}(\bigcup_{i=1}^{n} \{B\}_n) = \sum_{i=1}^{n} \mathbb{P}(\{B\}_n).$$

Therefore,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(\{A\}_n) &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(\{B\}_i) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(\{B\}_i) \\
&= \mathbb{P}(\bigcup_{i=1}^{\infty} \{B\}_i) \\
&= \mathbb{P}(\bigcup_{i=1}^{\infty} \{A\}_i) \\
&= \mathbb{P}(\lim_{n\to\infty} \{A\}_n).
\end{aligned}$$

If $\{\{A\}_n\}_{n=1}^{\infty}$ is a decreasing sequence of events, then $\{\{A\}_n^c\}_{n=1}^{\infty}$ is an increasing sequence of events. Therefore, using the results from the first part

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(\{A\}_n) &= \lim_{n\to\infty} (1 - \mathbb{P}(\{A\}_n^c)) \\
&= 1 - \lim_{n\to\infty} \mathbb{P}(\{A\}_n^c) \\
&= 1 - \mathbb{P}(\lim_{n\to\infty} \{A\}_n^c) \\
&= 1 - \mathbb{P}(\bigcap_{n=1}^{\infty} \{A\}_n^c) \\
&= \mathbb{P}(\bigcup_{n=1}^{\infty} \{A\}_n) \\
&= \mathbb{P}(\lim_{n\to\infty} \{A\}_n).
\end{aligned}$$

□

D.2.2 Convergence of Random Variables

Definition D.10 (Almost sure convergence)**.** A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ *converges almost surely* to X, denoted $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

Theorem D.1. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{\text{P}} X$

Proof. Left as exercise. □

Lemma D.1. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. Then $X_n \xrightarrow{\text{a.s.}} X$ if and only if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_m - X| < \epsilon \text{ for all } m \geq n) = 1.$$

Proof. Let $\epsilon_1 > \epsilon_2 > \epsilon_3 > \dots$ be a sequence of decreasing numbers such that $\lim_{n\to\infty} \epsilon_n = 0$. For each $n \in \mathbb{N}$ and $L \in \mathbb{N}$, let $\{A\}_L^n$ be the set

$$\{A\}_L^n = \{\omega : \ |X_m(\omega) - X(\omega)| < \epsilon_L \text{ for all } m \geq n\}.$$

Observe that $\{A\}_L^n \subset \{A\}_L^{n+1}$ for all $n \in \mathbb{N}$. Therefore, by Claim D.6, the set

$$\begin{aligned}
\{A\}_L &= \lim_{n\to\infty} \{A\}_L^n \\
&= \bigcup_{n=1}^{\infty} \{A\}_L^n \\
&= \{\omega : \text{ there exists } n \in \mathbb{N} \text{ such that } |X_m(\omega) - X(\omega)| < \epsilon_L \text{ for all } m \geq n\}
\end{aligned}$$

exists for each $L \in \mathbb{N}$. Similarly, observe that $\{A\}_L \supset \{A\}_{L+1}$ for all $L \in \mathbb{N}$. Therefore, the set

$$\begin{aligned}
\{A\} &= \lim_{L\to\infty} \{A\}_L \\
&= \bigcap_{L=1}^{\infty} \{A\}_L \\
&= \{\omega : \text{ for all } L \geq 1, \text{ there exists } n \in \mathbb{N} \text{ such that } |X_m(\omega) - X(\omega)| < \epsilon_L \text{ for all } m \geq n\} \\
&= \{\omega : \text{ for all } \epsilon > 0, \text{ there exists } n \in \mathbb{N} \text{ such that } |X_m(\omega) - X(\omega)| < \epsilon \text{ for all } m \geq n\} \\
&= \left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}
\end{aligned}$$

exists.
($\Leftarrow$) Assume for every $\epsilon > 0$,

$$\lim_{n\to\infty} p\left(|X_m - X| < \epsilon \text{ for all } m \geq n\right) = 1.$$

Then

$$\begin{aligned} p\left(\lim_{n\to\infty} X_n = X\right) &= p\left(\{A\}\right) \\ &= p\left(\lim_{L\to\infty} \{A\}_L\right) \\ &= \lim_{L\to\infty} p\left(\{A\}_L\right) \\ &= \lim_{L\to\infty} p\left(\lim_{n\to\infty} \{A\}_L^n\right) \\ &= \lim_{L\to\infty} \lim_{n\to\infty} p\left(\{A\}_L^n\right) \\ &= \lim_{L\to\infty} \lim_{n\to\infty} p\left(|X_m - X| < \epsilon_L \text{ for all } m \geq n\right) \\ &= \lim_{L\to\infty} 1 \\ &= 1. \end{aligned}$$

Hence, $X_n \overset{a.s.}{\to} X$.
($\Rightarrow$) Conversely, assume $X_n \overset{a.s.}{\to} X$ and let $\epsilon > 0$. Since

$$\{A\} = \bigcap_{L=1}^{\infty} \{A\}_L \subset \{A\}_L$$

for all $L \in \mathbb{N}$, it follows

$$1 = p\left(\{A\}\right) \leq p\left(\{A\}_L\right) \leq 1$$

for all $L \in \mathbb{N}$. Therefore, $p\left(\{A\}_L\right) = 1$ and

$$\begin{aligned} p\left(\{A\}_L\right) &= \lim_{n\to\infty} p\left(\{A\}_L^n\right) \\ &= \lim_{n\to\infty} p\left(|X_m - X| < \epsilon_L \text{ for all } m \geq n\right). \end{aligned}$$

Since $\epsilon_n \to 0$ as $n \to \infty$, choose L such that $\epsilon_L < \epsilon$. Hence,

$$1 = \lim_{n\to\infty} p\left(|X_m - X| < \epsilon_L < \epsilon \text{ for all } m \geq n\right).$$

□

Definition D.11 (Almost complete convergence)**.** A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ *converges almost completely* to X, denoted $X_n \overset{c}{\to} X$, if for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} p\left(|X_n - X| \geq \epsilon\right) < \infty.$$

Lemma D.2. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. If $X_n \overset{c}{\to} X$, then $X_n \overset{a.s.}{\to} X$.*

Proof. Assume for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} p\left(|X_n - X| \geq \epsilon\right) < \infty.$$

Let

$$\{A\}_n = \{\omega : \ |X_m(\omega) - X(\omega)| < \epsilon \text{ for all } m \geq n\}.$$

Then

$$\begin{aligned}\{A\}_n^c &= \{\omega : \ |X_m(\omega) - X(\omega)| \geq \epsilon \text{ for some } m \geq n\} \\ &= \bigcup_{m=n}^{\infty} \{\omega : \ |X_m(\omega) - X(\omega)| \geq \epsilon \ \}.\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} p\left(\{A\}_n^c\right) &= \lim_{n\to\infty} p\left(\bigcup_{m=n}^{\infty} \{\omega : \ |X_m(\omega) - X(\omega)| \geq \epsilon \ \}\right) \\ &\leq \lim_{n\to\infty} \sum_{m=n}^{\infty} p\left(|X_m - X| \geq \epsilon\right) \\ &= 0.\end{aligned}$$

Hence,

$$\begin{aligned}\lim_{n\to\infty} p\left(|X_m - X| < \epsilon \text{ for all } m \geq n\right) &= \lim_{n\to\infty} p\left(\{A\}_n\right) \\ &= \lim_{n\to\infty} 1 - p\left(\{A\}_n^c\right) \\ &= 1.\end{aligned}$$

So by Lemma **??**, $X_n \overset{a.s.}{\to} X$.

□

D.3 Strong Law of Large Numbers

Theorem D.2 (Strong Law of Large Numbers)**.** *Let* $\{X_n\}_{n=1}^{\infty}$ *be an i.i.d. sequence with* $\mathrm{E}\left(X^4\right) < \infty$*. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathrm{E}(X).$$

Proof. First, we'll assume $\mathrm{E}(X_i) = 0$. Observe that

$$\mathrm{E}\left(\bar{X}^4\right) = \mathrm{E}\left(\frac{(X_1 + \cdots + X_n)^4}{n^4}\right) = \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathrm{E}(X_i X_j X_k X_\ell).$$

By independence, the only nonzero terms are of the form $\mathrm{E}\left(X_i^4\right)$ or $\mathrm{E}\left(X_i^2 X_j^2\right)$ where $i \neq j$. By counting all possible nonzero terms, we get

$$\mathrm{E}\left(\bar{X}^4\right) = \frac{1}{n^4}\left[n\,\mathrm{E}\left(X_1^4\right) + 3n(n-1)\,\mathrm{E}\left(X_1^2 X_2^2\right)\right].$$

By Holder's inequality,

$$\mathrm{E}\left(X_1^2 X_2^2\right) \leq \sqrt{\mathrm{E}\left(X_1^4\right)}\sqrt{\mathrm{E}\left(X_2^4\right)} = \mathrm{E}\left(X_1^4\right).$$

Therefore,

$$\begin{aligned}\mathrm{E}\left(\bar{X}^4\right) &\leq \frac{1}{n^4}\left[(n+3n(n-1))\,\mathrm{E}\left(X_1^4\right)\right] \\ &\leq \frac{1}{n^4}\left[3n^2\,\mathrm{E}\left(X_1^4\right)\right] \\ &= \frac{3\,\mathrm{E}\left(X_1^4\right)}{n^2}.\end{aligned}$$

Let $\epsilon > 0$. By Markov's inequality,

$$p\left(|\bar{X}| \geq \epsilon\right) = p\left(\bar{X}^4 \geq \epsilon^4\right) \leq \frac{\mathrm{E}\left(\bar{X}^4\right)}{\epsilon^4}.$$

Hence,

$$\begin{aligned}\sum_{n=1}^{\infty} p\left(|\bar{X}| \geq \epsilon\right) &\leq \sum_{n=1}^{\infty} \frac{\mathrm{E}\left(\bar{X}^4\right)}{\epsilon^4} \\ &\leq \sum_{n=1}^{\infty} \frac{3\,\mathrm{E}\left(X_1^4\right)}{\epsilon^4 n^2} \\ &= \frac{3\,\mathrm{E}\left(X_1^4\right)}{\epsilon^4} \sum_{n=1}^{\infty} \frac{1}{n^2} \\ &< \infty.\end{aligned}$$

So by Lemma **??**, $\bar{X} \overset{a.s.}{\to} 0$.
Now suppose $\mathrm{E}(X_i) \neq 0$. Then

$$\begin{aligned}\bar{X} - \mathrm{E}(X_i) &= \frac{X_1 + \cdots + X_n - n\,\mathrm{E}(X_i)}{n} \\ &= \frac{(X_1 - E(X_1)) + \cdots + (X_n - \mathrm{E}(X_n))}{n} \\ &\overset{a.s.}{\to} 0.\end{aligned}$$

Hence, $\bar{X} \overset{a.s.}{\to} \mathrm{E}(X_i)$.

□

Bibliographic Notes and Further ReadingThe main results of this chapter could be found in [1].

Appendix E

Numeric Optimizations

Optimization plays a key-role in machine learning. Many machine learning problems are formalized as optimization problems. Below we focus on constrained optimization problems where we have an objective function to optimize and a set of constraints to satisfy and the so-called Karush-Kuhn-Tucker(KKT) Conditions.

Definition E.1 (Constratined Optimization Problem)**.** We consider the standard form for nonlinear programming problems. The objective function is denoted by $f(\mathbf{x})$ while the equality and inequality constraints are shown as vector-valued functions.

$$\begin{aligned} \text{minimize } & f(\mathbf{x}) \\ \text{s.t. } & \mathbf{g}(\mathbf{x}) = 0 \\ & \mathbf{h}(\mathbf{x}) \geq 0 \end{aligned} \tag{E.0.1}$$

Definition E.2 (Regular Point)**.** If the point $\mathbf{x^*}$ satisfies the following constraints, then $\mathbf{x^*}$ is a *regular point*.

$$\begin{aligned} \mathbf{h(x^*) = 0} \\ \mathbf{g(x^*)} \leq 0 \end{aligned}$$

Definition E.3 (Karush-Kuhn-Tucker(KKT) Conditions)**.** Let $\mathbf{x^*}$ be a relative minimum for a nonlinear programming problem in the standard form and is also a regular point of the constraints. Then there is a vector $\boldsymbol{\lambda}$ and a vector $\boldsymbol{\mu} \geq \mathbf{0}$ such that:

$$\begin{aligned} \boldsymbol{\nabla} f(\mathbf{x^*}) + \boldsymbol{\lambda} \boldsymbol{\nabla} \mathbf{h}(\mathbf{x}^*) + \boldsymbol{\mu}^T \mathbf{g} \boldsymbol{\nabla}(\mathbf{x}^*) = \mathbf{0} \\ \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}^*) = 0 \end{aligned}$$

Definition E.4 (Complementary Slackness)**.** Note that $\boldsymbol{\mu} \geq \mathbf{0}$ and $\mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$ is a *complementary slackness* condition. Meaning that a component of the $\boldsymbol{\mu}$ may be nonzero only if the matching constraint if active.

$$\begin{aligned} \mathbf{g}(\mathbf{x}^*) < 0 &\implies \mu_i = 0 \\ \mu_i > 0 &\implies \mathbf{g}(\mathbf{x}^*) = 0 \end{aligned}$$

Bibliography

[1] Patrick Billingsley. *Probability and Measure*. Wiley, 1986.

[2] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] Thomas S. Ferguson. *A Course in Large Sample Theory*. Chapman & Hall, 1996.

[4] Subhashis Ghoshal. A review of consistency and convergence rates of posterior distribution. *Proceedings of Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1996.

[5] Keigh Knight. *Mathematical Statistics*. Chapman & Hall, 2000.

[6] Antonio Lijoi, Igor Prunster, and Stephen G. Walker. Bayesian consistency for stationary models. *Econometric Theory*, 23:749–759, 2007.

[7] Kevin Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

[8] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[9] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[10] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.