

# 微积分与概率论基础

---

3月机器学习在线班 邹博

2015年3月7日

# 回忆知识

---

□ 求S的值：

$$S = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots + \frac{1}{n!} + \cdots$$



# 复习微积分：两边夹定理

---

□ 当  $x \in U(x_0, r)$  时，有  $g(x) \leq f(x) \leq h(x)$  成立，  
并且  $\lim_{x \rightarrow x_0} g(x) = A$ ， $\lim_{x \rightarrow x_0} h(x) = A$ ，那么

$$\lim_{x \rightarrow x_0} f(x) = A$$

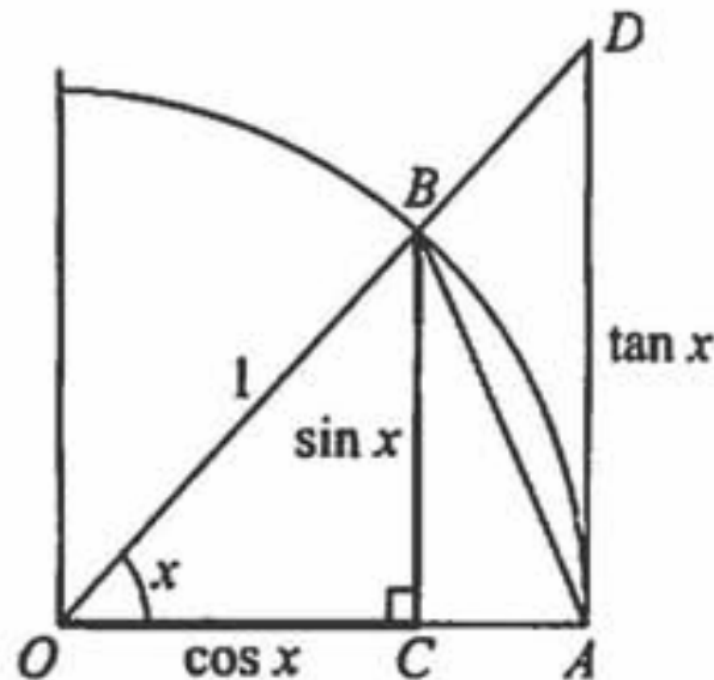


# 极限

- 由右图： $\sin x < x < \tan x$ ，  
 $x \in U(0, \varepsilon)$
- 从而： $1 < x/\sin x < 1/\cos x$
- 即： $\cos x < \sin x/x < 1$
- 因为： $\lim_{x \rightarrow 0} \cos x = \cos 0 = 1$
- 从而：

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$$

- 该式将三角函数和多项式建立了极限关系



# 思考

---

□ 该式的极限是多少？

$$\lim_{x \rightarrow 0} \frac{\sin^2 x}{x^2}$$



# 复习微积分：极限存在定理

---

□ 单调有界数列必有极限

■ 单增数列有上界，则其必有极限



# 构造数列 $\{x_n\}$

$$\begin{aligned}x_n &= \left(1 + \frac{1}{n}\right)^n \\&= 1 + C_n^1 \frac{1}{n} + C_n^2 \frac{1}{n^2} + C_n^3 \frac{1}{n^3} + \cdots + C_n^n \frac{1}{n^n} \\&= 1 + n \cdot \frac{1}{n} + \frac{n(n-1)}{2!} \cdot \frac{1}{n^2} + \frac{n(n-1)(n-2)}{3!} \cdot \frac{1}{n^3} + \cdots + \frac{n(n-1)(n-2)\cdots 1}{n!} \cdot \frac{1}{n^n} \\&= 1 + 1 + \frac{1}{2!} \cdot \left(1 - \frac{1}{n}\right) + \frac{1}{3!} \cdot \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) + \cdots + \frac{1}{n!} \cdot \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{n-1}{n}\right) \\&< 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} \\&< 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-1}} \\&= 3 - \frac{1}{2^{n-1}} \\&< 3\end{aligned}$$



# 自然常数

---

□ 根据  $\left(1 + \frac{1}{n+1}\right)^n < \left(1 + \frac{1}{x}\right)^x < \left(1 + \frac{1}{n}\right)^{n+1}$

□ 从而公式  $\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x$  的极限存在，定义为  $e$ 。

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x = e$$





# 导数



- 简单的说，导数就是曲线的斜率，是曲线变化快慢的反应
- **二阶导数**是斜率变化快慢的反应，表征曲线的**凸凹性**
  - 在GIS中，往往一条二阶导数连续的曲线，我们称之为“**光顺**”的。
  - 还记得高中物理老师时常念叨的吗：**加速度的**方向总是指向轨迹曲线凹的一侧



# 常用函数的导数

---

$$(1) \quad C' = 0 \quad (C \text{ 为常数});$$

$$(2) \quad (x^n)' = nx^{n-1} \quad (n \in Q);$$

$$(3) \quad (\sin x)' = \cos x;$$

$$(4) \quad (\cos x)' = -\sin x;$$

$$(5) \quad (a^x)' = a^x \ln a;$$

$$(6) \quad (e^x)' = e^x;$$

$$(7) \quad (\log_a x)' = \frac{1}{x} \log_a e;$$

$$(8) \quad (\ln x)' = \frac{1}{x}.$$

$$(u + v)' = u' + v'$$

$$(uv)' = u'v + uv'$$



# 应用

---

- 已知函数  $f(x)=x^x$ ,  $x>0$
- 求  $f(x)$  的最小值
  - 领会幂指数函数的一般处理套路

- 附:  $N^{\frac{1}{\log N}} = ?$

- 在计算机算法跳跃表Skip List的分析中, 用到了该常数。
- 背景: 跳表是支持增删改查的动态数据结构, 能够达到与平衡二叉树、红黑树近似的效率, 而代码实现简单。



# 求解 $x^x$

---

$$t = x^x$$

$$\rightarrow \ln t = x \ln x$$

$$\xrightarrow{\text{两边对}x\text{求导}} \frac{1}{t} t' = \ln x + 1$$

$$\xrightarrow{\text{令}t'=0} \ln x + 1 = 0$$

$$\rightarrow x = e^{-1}$$

$$\rightarrow t = e^{-\frac{1}{e}}$$



# Taylor公式 – Maclaurin公式

---

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_n(x)$$

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + o(x^n)$$



# Taylor公式的应用

---

□ 数值计算：初等函数值的计算(在原点展开)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \cdots + (-1)^{m-1} \frac{x^{2m-1}}{(2m-1)!} + R_{2m}$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + R_n$$

□ 在实践中，往往需要做一定程度的变换。



# 计算 $e^x$

---

- 给定正实数 $x$ , 计算 $e^x=?$
- 一种可行的思路:
- 求整数 $k$ 和小数 $r$ , 使得
  - $x = k \cdot \ln 2 + r, |r| \leq 0.5 \cdot \ln 2$
- 从而:
$$e^x = e^{k \cdot \ln 2 + r}$$
$$= e^{k \cdot \ln 2} \cdot e^r$$
$$= 2^k \cdot e^r$$



# Taylor公式的应用

---

□ 考察基尼指数的图像、熵、分类误差率三者之间的关系

■ 将 $f(x)=-\ln x$ 在 $x_0=1$ 处一阶展开，忽略高阶无穷小，得到 $f(x) \approx 1-x$

■ 从而：
$$H(X) = -\sum_{k=1}^K p_k \ln p_k$$

$$\approx \sum_{k=1}^K p_k (1-p_k)$$

■ 上述结论，在决策树章节中会进一步讨论





# 方向导数

---

- 如果函数  $z=f(x,y)$  在点  $P(x,y)$  是可微分的，那么，函数在该点沿任一方向  $L$  的方向导数都存在，且有：

$$\frac{\partial f}{\partial l} = \frac{\partial f}{\partial x} \cos \varphi + \frac{\partial f}{\partial y} \sin \varphi$$

- 其中， $\psi$  为  $x$  轴到方向  $L$  的转角。



# 梯度

---

- 设函数 $z=f(x,y)$ 在平面区域 $D$ 内具有一阶连续偏导数，则对于每一个点 $P(x,y) \in D$ ，向量

$$\left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

为函数 $z=f(x,y)$ 在点 $P$ 的梯度，记做 $\text{grad}f(x,y)$

- 梯度的方向是函数在该点变化最快的方向
  - 考虑一座解析式为 $H(x,y)$ 的山。在 $(x_0,y_0)$ 点的梯度是在该点坡度最陡的方向。
- 梯度下降法
  - 思考：如果下山方向和梯度呈 $\theta$ 夹角，下降速度是多少？



# 凸函数

---

- $f(x)$ 在区间 $I$ 上连续，如果对 $I$ 上任意两点 $x_1, x_2$ ，恒有 $f((x_1+x_2)/2) < (f(x_1)+f(x_2))/2$ ，则称 $f(x)$ 在 $I$ 上是凸的。
- 注：中国大陆数学界某些机构关于函数凹凸性定义和国外的定义是相反的。Convex Function在某些中国大陆的数学书中指凹函数。Concave Function指凸函数。但在中国大陆涉及经济学的很多书中，凹凸性的提法和其他国家的提法是一致的，也就是和数学教材是反的。举个例子，同济大学高等数学教材对函数的凹凸性定义与习惯定义正好相反。另外，也有些教材会把凸定义为上凸，凹定义为下凸。



# 凸函数的判定

---

- 定理： $f(x)$ 在区间 $[a,b]$ 上连续，在 $(a,b)$ 内二阶可导，那么：
  - 若 $f''(x) > 0$ ，则 $f(x)$ 是凸的；
  - 若 $f''(x) < 0$ ，则 $f(x)$ 是凹的
  
- 即：一元二阶可微的函数在区间上是凸的，当且仅当它的二阶导数是非负的



# 凸函数

---

□ 凸函数更一般的表述

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

$f$  为凸函数，则有：

$$f(\theta_1 x_1 + \dots + \theta_n x_n) \leq \theta_1 f(x_1) + \dots + \theta_n f(x_n)$$

其中  $0 \leq \theta_i \leq 1, \theta_1 + \dots + \theta_n = 1$ .

□ 意义：可以在确定函数的凸凹性之后，对函数进行不等式替换。



# 凸性质的应用

---

- 设 $p(x)$ 、 $q(x)$ 是在 $X$ 中取值的两个概率分布，给定如下定义式：

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

- 试证明： $D(p \parallel q) \geq 0$

- 上式在最大熵模型等内容中会详细讨论。



# 注意到 $y=\log x$ 在定义域上是凹函数

---

$$\begin{aligned} D(p \parallel q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= -\sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\geq -\log \sum_x \left( p(x) \cdot \frac{q(x)}{p(x)} \right) \\ &= -\log \sum_x q(x) \\ &= -\log 1 \\ &= 0 \end{aligned}$$



# 概率论

---

□ 对概率的认识:  $P \in [0,1]$

■  $P=0$

□ 事件出现的概率为0 → 事件不会发生?

■ 将位于 $[0,1]$ 的函数 $y=f(x)$ 看成 $x$ 对应 $y$ 事件的概率

□ 要求 $f(x)$ 在定义域 $[0,1]$ 的积分为1

□ 古典概型

■ 排列组合

□ 概率密度函数Probability Density Function

□ 累计分布函数





# 古典概型

---

- 举例：将 $n$ 个不同的球放入 $N(N \geq n)$ 个盒子中，假设盒子容量无限，求事件 $A = \{\text{每个盒子至多有1个球}\}$ 的概率。



解  $P(A) = \frac{P_N^n}{N^n}$

---

□ 基本事件总数:

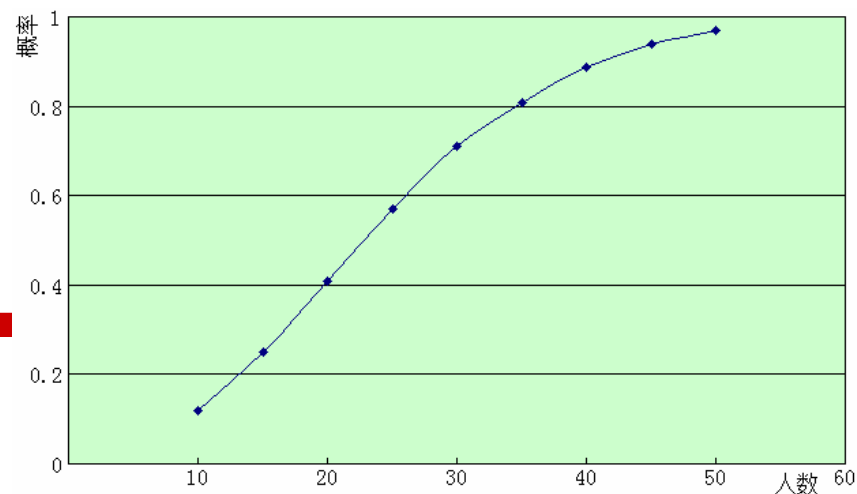
- 第1个球, 有N种放法;
- 第2个球, 有N种放法;
- .....
- 共:  $N^n$ 种放法。

□ 每个盒子至多放1个球的事件数:

- 第1个球, 有N种放法;
- 第2个球, 有N-1种放法;
- 第3个球, 有N-2种放法;
- .....
- 共:  $N(N-1)(N-2)\cdots(N-n+1) = P_N^n$



# 实际问题



□ 某班上有50位同学，至少有2人生日相同的概率是多少？

n	10	15	20	25	30	35	40	45	50
P	0.12	0.25	0.41	0.57	0.71	0.81	0.89	0.94	0.97



# 装箱问题

---

- 将12件正品和3件次品随机装在3个箱子中。  
每箱中恰有1件次品的概率是多少？



# 解

---

- 将15件产品装入3个箱子，每箱装5件，共有  $15!/(5!5!5!)$  种装法；
- 先把3件次品放入3个箱子，有  $3!$  种装法。对于这样的每一种装法，把其余12件产品装入3个箱子，每箱装4件，共有  $12!/(4!4!4!)$  种装法；
- $P(A) = (3! * 12! / (4!4!4!)) / (15! / (5!5!5!)) = 25/91$



# 与组合数的关系

---

- 把n个物品分成k组，使得每组物品的个数分别为 $n_1, n_2, \dots, n_k$ ，( $n = n_1 + n_2 + \dots + n_k$ )，则不同的分组方法有 $\frac{n!}{n_1! n_2! n_3! \cdots n_k!}$ 种。
- 上述问题的简化版本，即n个物品分成2组，第一组m个，第二组n-m个，则分组方法有 $\frac{n!}{m!(n-m)!}$ ，即： $C_n^m$ 。



# 概率

---

□ 条件概率:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

□ 全概率公式:

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

□ 贝叶斯(Bayes)公式:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$



# 思考题

---

- 8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取一支射击，结果中靶。求该枪是已校准过的概率。





# 分布

---

- 复习各种常见分布本身的统计量
- 在复习各种分布的同时，重温积分、Taylor 展式等前序知识
- 常见分布是可以完美统一为**一类分布**



# 两点分布

---

## 0-1分布

已知随机变量  $X$  的分布律为

$X$	1	0
$p$	$p$	$1-p$

则有  $E(X) = 1 \cdot p + 0 \cdot q = p,$

$$\begin{aligned} D(X) &= E(X^2) - [E(X)]^2 \\ &= 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = pq. \end{aligned}$$



# 二项分布 Bernoulli distribution

---

设随机变量  $X$  服从参数为  $n, p$  二项分布,

**(法一)** 设  $X_i$  为第  $i$  次试验中事件  $A$  发生的次数,  $i=1, 2, \dots, n$

则

$$X = \sum_{i=1}^n X_i$$

显然,  $X_i$  相互独立均服从参数为  $p$  的0-1分布,

$$\text{所以 } E(X) = \sum_{i=1}^n E(X_i) = np.$$

$$D(X) = \sum_{i=1}^n D(X_i) = np(1-p).$$



# 二项分布

(法二)  $X$  的分布律为

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, (k = 0, 1, 2, \dots, n),$$

$$\text{则有 } E(X) = \sum_{k=0}^n k \cdot P\{X = k\} = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \sum_{k=0}^n \frac{kn!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$= \sum_{k=1}^n \frac{np(n-1)!}{(k-1)![(n-1)-(k-1)]!} p^{k-1} (1-p)^{(n-1)-(k-1)}$$

$$= np \sum_{k=1}^n \frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!} p^{k-1} (1-p)^{(n-1)-(k-1)}$$

$$= np[p + (1-p)]^{n-1} = np$$



# 二项分布

$$E(X^2) = E[X(X-1) + X] = E[X(X-1)] + E(X)$$

$$= \sum_{k=0}^n k(k-1) \binom{k}{n} p^k (1-p)^{n-k} + np$$

$$= \sum_{k=0}^n \frac{k(k-1)n!}{k!(n-k)!} p^k (1-p)^{n-k} + np$$

$$= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{(n-k)!(k-2)!} p^{k-2} (1-p)^{(n-2)-(k-2)} + np$$

$$= n(n-1)p^2 [p + (1-p)]^{n-2} + np = (n^2 - n)p^2 + np.$$

$$D(X) = E(X^2) - [E(X)]^2 = (n^2 - n)p^2 + np - (np)^2$$

$$= np(1-p)$$



# 考察Taylor展式

---

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + R_k$$

$$1 = 1 \cdot e^{-x} + x \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \frac{x^3}{3!} \cdot e^{-x} + \cdots + \frac{x^k}{k!} \cdot e^{-x} + R_n \cdot e^{-x}$$

$$\frac{x^k}{k!} \cdot e^{-x} \longrightarrow \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$



# 泊松分布

---

设  $X \sim \pi(\lambda)$ , 且分布律为

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots, \quad \lambda > 0.$$

则有

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \cdot \lambda \\ &= \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda \end{aligned}$$



# 泊松分布Poisson distribution

---

- 在实际事例中，当一个随机事件，以固定的平均瞬时速率 $\lambda$  (或称密度)随机且独立地出现时，那么这个事件在单位时间(面积或体积)内出现的次数或个数就近似地服从泊松分布 $P(\lambda)$ 。
  - 某一服务设施在一定时间内到达的人数
  - 电话交换机接到呼叫的次数
  - 汽车站台的候客人数
  - 机器出现的故障数
  - 自然灾害发生的次数
  - 一块产品上的缺陷数
  - 显微镜下单位分区内的细菌分布数
  - 某放射性物质单位时间发射出的粒子数





# 泊松分布

---

$$\begin{aligned} E(X^2) &= E[X(X-1) + X] \\ &= E[X(X-1)] + E(X) \\ &= \sum_{k=0}^{+\infty} k(k-1) \cdot \frac{\lambda^k}{k!} e^{-\lambda} + \lambda \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{+\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda. \end{aligned}$$

所以  $D(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

泊松分布的期望和方差都等于参数  $\lambda$ .



# 均匀分布

---

设  $X \sim U(a, b)$ , 其概率密度为

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & \text{其他.} \end{cases}$$

$$\text{则有 } E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b \frac{1}{b-a} x dx = \frac{1}{2}(a+b).$$

$$\begin{aligned} D(X) &= E(X^2) - [E(X)]^2 \\ &= \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12} \end{aligned}$$



# 指数分布

设随机变量  $X$  服从指数分布, 其概率密度为

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x > 0, \\ 0, & x \leq 0. \end{cases} \quad \text{其中 } \theta > 0.$$

则有

$$E(X) = \int_{-\infty}^{+\infty} xf(x) dx = \int_0^{+\infty} x \cdot \frac{1}{\theta} e^{-x/\theta} dx$$

$$= -xe^{-x/\theta} \Big|_0^{+\infty} + \int_0^{+\infty} e^{-x/\theta} dx = \theta$$

$$D(X) = E(X^2) - [E(X)]^2 = \int_0^{+\infty} x^2 \cdot \frac{1}{\theta} e^{-x/\theta} dx - \theta^2$$

$$= 2\theta^2 - \theta^2 = \theta^2$$



# 指数分布

---

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

- 其中  $\lambda > 0$  是分布的一个参数，常被称为率参数(rate parameter)。即每单位时间内发生某事件的次数。指数分布的区间是  $[0, \infty)$ 。如果一个随机变量  $X$  呈指数分布，则可以写作： $X \sim \text{Exponential}(\lambda)$ 。
- 指数分布可以用来表示独立随机事件发生的时间间隔，比如旅客进机场的时间间隔、软件更新的时间间隔等等。
- 许多电子产品的寿命分布一般服从指数分布。有的系统的寿命分布也可用指数分布来近似。它在可靠性研究中最常用的一种分布形式。



# 指数分布的无记忆性

---

□ 指数函数的一个重要特征是无记忆性(遗失记忆性, Memoryless Property)。

■ 如果一个随机变量呈指数分布, 当 $s, t \geq 0$ 时有:

$$P(x > s + t | x > s) = P(x > t)$$

■ 即, 如果 $x$ 是某一元件的寿命, 已知元件使用了 $s$ 小时, 它总共使用至少 $s+t$ 小时的条件概率, 与从开始使用时算起它使用至少 $t$ 小时的概率相等。



# 正态分布

---

设  $X \sim N(\mu, \sigma^2)$ , 其概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \sigma > 0, \quad -\infty < x < +\infty.$$

则有  $E(X) = \int_{-\infty}^{+\infty} xf(x) dx$

$$= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

$$\text{令 } \frac{x-\mu}{\sigma} = t \Rightarrow x = \mu + \sigma t,$$



# 正态分布

---

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\mu + \sigma t) e^{-\frac{t^2}{2}} dt \\ &= \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt \\ &= \mu. \end{aligned}$$



# 正态分布

---

$$\begin{aligned} D(X) &= \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \\ &= \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx. \end{aligned}$$

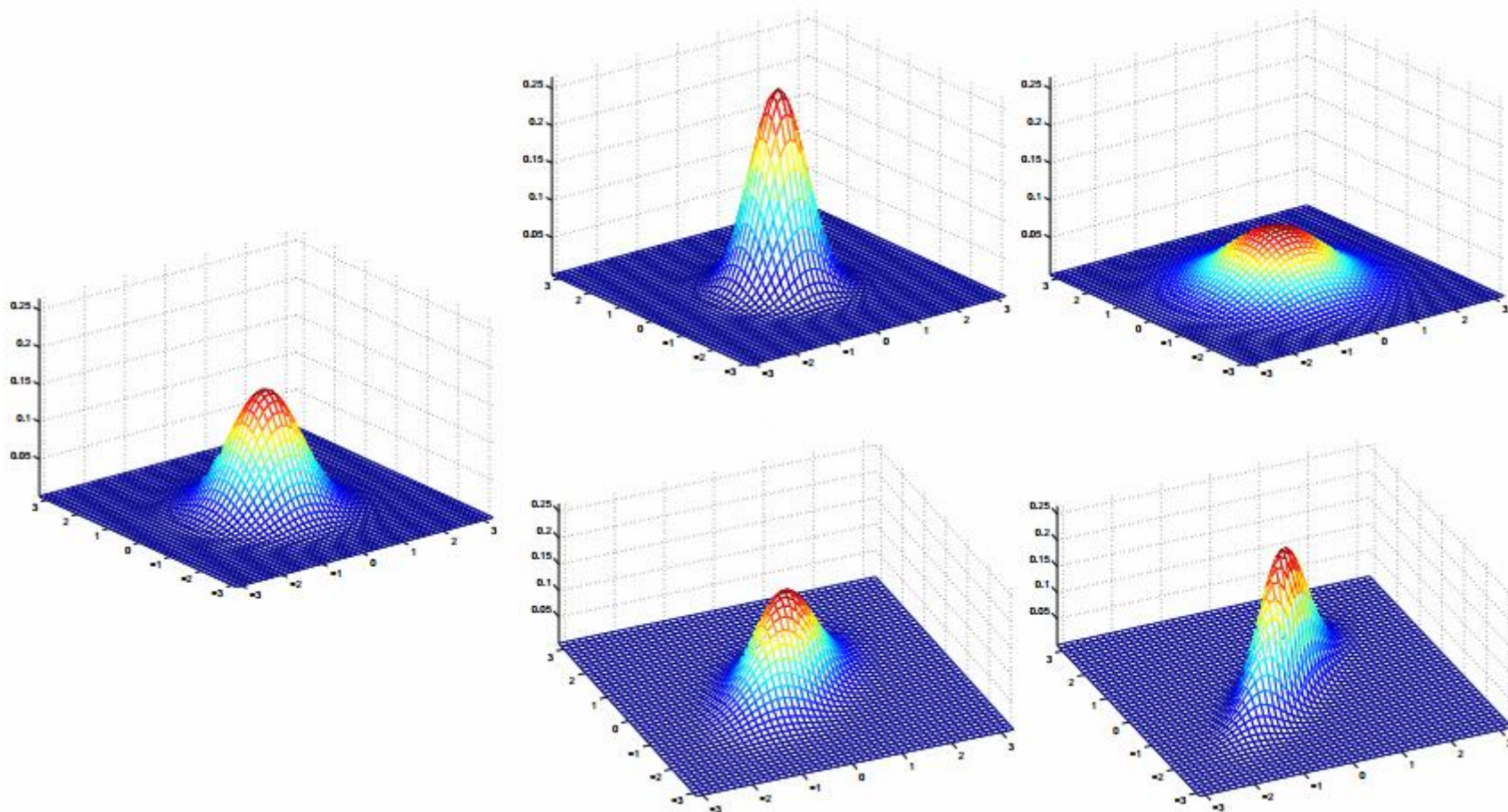
令  $\frac{x - \mu}{\sigma} = t$ , 得

$$\begin{aligned} D(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left( -te^{-\frac{t^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt \right) \\ &= 0 + \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \sigma^2. \end{aligned}$$





# 二元正态分布



# 总结

分 布	参 数	数学期望	方 差
两点分布	$0 < p < 1$	$p$	$p(1-p)$
二项分布	$n \geq 1,$ $0 < p < 1$	$np$	$np(1-p)$
泊松分布	$\lambda > 0$	$\lambda$	$\lambda$
均匀分布	$a < b$	$(a+b)/2$	$(b-a)^2/12$
指数分布	$\theta > 0$	$\theta$	$\theta^2$
正态分布	$\mu, \sigma > 0$	$\mu$	$\sigma^2$



# 指数族

---

## The exponential family

To work our way up to GLMs, we will begin by defining exponential family distributions. We say that a class of distributions is in the exponential family if it can be written in the form

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta)) \quad (6)$$

Here,  $\eta$  is called the **natural parameter** (also called the **canonical parameter**) of the distribution;  $T(y)$  is the **sufficient statistic** (for the distributions we consider, it will often be the case that  $T(y) = y$ ); and  $a(\eta)$  is the **log partition function**. The quantity  $e^{-a(\eta)}$  essentially plays the role of a normalization constant, that makes sure the distribution  $p(y; \eta)$  sums/integrates over  $y$  to 1.

A fixed choice of  $T$ ,  $a$  and  $b$  defines a *family* (or set) of distributions that is parameterized by  $\eta$ ; as we vary  $\eta$ , we then get different distributions within this family.



# 指数族

---

- 指数族概念的目的，是为了说明广义线性模型 Generalized Linear Models
  - 凡是符合指数族分布的随机变量，都可以用 GLM 回归分析



## 如：Bernoulli分布和高斯分布

---

We now show that the Bernoulli and the Gaussian distributions are examples of exponential family distributions. The Bernoulli distribution with mean  $\phi$ , written  $\text{Bernoulli}(\phi)$ , specifies a distribution over  $y \in \{0, 1\}$ , so that  $p(y = 1; \phi) = \phi$ ;  $p(y = 0; \phi) = 1 - \phi$ . As we vary  $\phi$ , we obtain Bernoulli distributions with different means. We now show that this class of Bernoulli distributions, ones obtained by varying  $\phi$ , is in the exponential family; i.e., that there is a choice of  $T$ ,  $a$  and  $b$  so that Equation (6) becomes exactly the class of Bernoulli distributions.



# Bernoulli分布属于指数族

---

We write the Bernoulli distribution as:

$$\begin{aligned} p(y; \phi) &= \phi^y (1 - \phi)^{1-y} \\ &= \exp(y \log \phi + (1 - y) \log(1 - \phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1 - \phi}\right)\right) y + \log(1 - \phi)\right). \end{aligned}$$

Thus, the natural parameter is given by  $\eta = \log(\phi/(1 - \phi))$ . Interestingly, if we invert this definition for  $\eta$  by solving for  $\phi$  in terms of  $\eta$ , we obtain  $\phi = 1/(1 + e^{-\eta})$ . This is the familiar sigmoid function! This will come up again when we derive logistic regression as a GLM. To complete the formulation of the Bernoulli distribution as an exponential family distribution, we also have

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1 - \phi) \\ &= \log(1 + e^\eta) \\ b(y) &= 1 \end{aligned}$$



## 考察参数 $\Phi$

---

□ 注意在推导过程中，出现了Logistic方程。

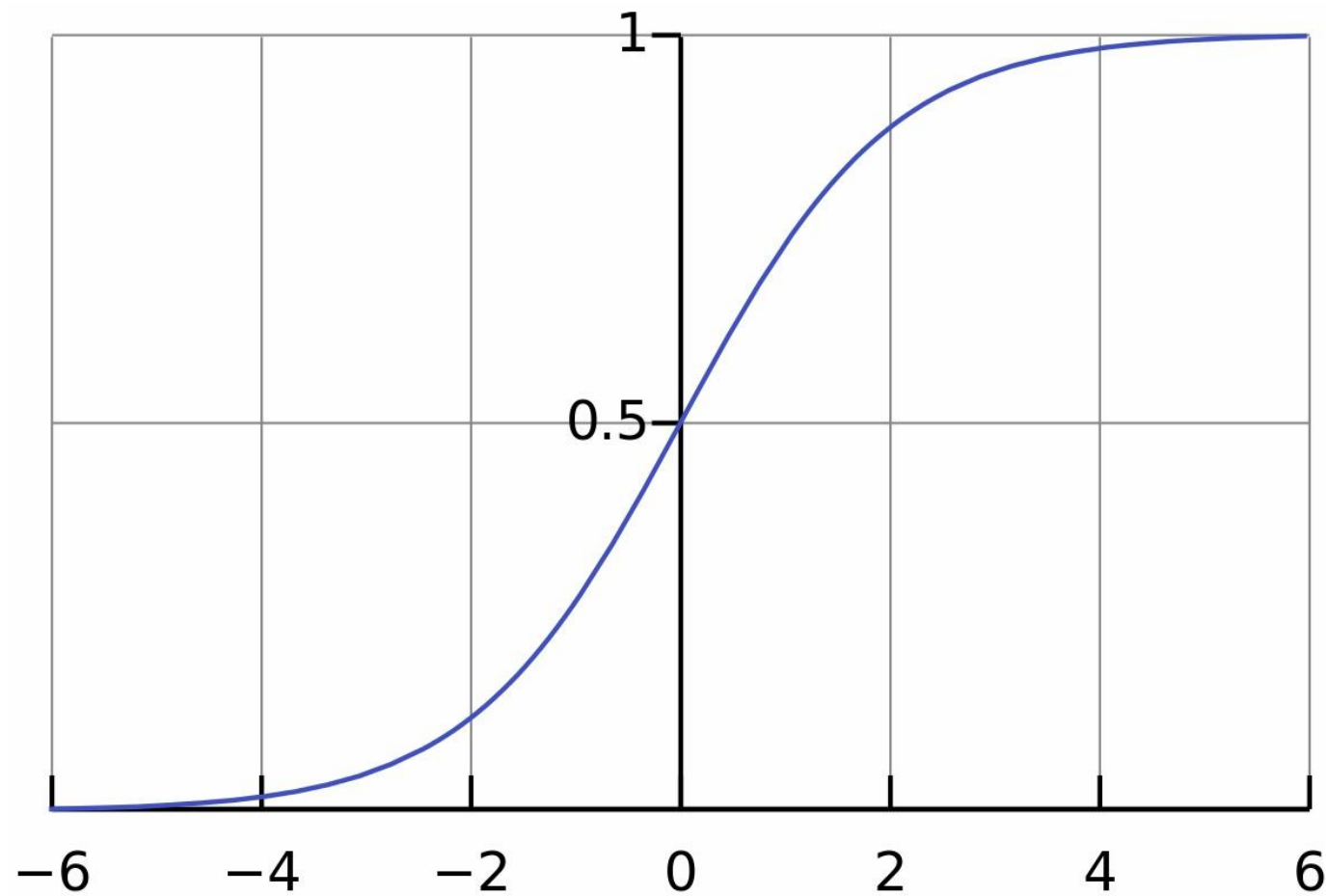
$$\Phi = \frac{1}{1 + e^{-\eta}}$$

$$f(x) = \frac{1}{1 + e^{-x}}$$



# Logistic函数

---





# Logistic函数的导数 $f(x) = \frac{1}{1+e^{-x}}$

$$\begin{aligned} f'(x) &= \left( \frac{1}{1+e^{-x}} \right)' \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \frac{1}{1+e^{-x}} \cdot \left( 1 - \frac{1}{1+e^{-x}} \right) \\ &= f(x) \cdot (1 - f(x)) \end{aligned}$$

□ 该结论后面会用到



# Gaussian分布也属于指数族分布

Lets now move on to consider the Gaussian distribution. Recall that, when deriving linear regression, the value of  $\sigma^2$  had no effect on our final choice of  $\theta$  and  $h_\theta(x)$ . Thus, we can choose an arbitrary value for  $\sigma^2$  without changing anything. To simplify the derivation below, lets set  $\sigma^2 = 1$ . We then have:

$$\begin{aligned} p(y; \mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \cdot \exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

$$\eta = \mu$$

$$T(y) = y$$

$$a(\eta) = \mu^2/2$$

$$= \eta^2/2$$

$$b(y) = (1/\sqrt{2\pi}) \exp(-y^2/2)$$



# 参考文献

---

- Prof. Andrew Ng, Machine Learning, Stanford University
- 同济大学数学教研室 主编，高等数学，高等教育出版社，1996
- 王松桂，程维虎，高旅端编，概率论与数理统计，科学出版社，2000



---

感谢大家!

恳请大家批评指正!

