

# 参数估计与矩阵运算基础

---

3月机器学习在线班 邹博

2015年3月8日

# 期望

---

□ 离散型  $E(X) = \sum_i x_i p_i$

□ 连续型  $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

□ 即：概率加权下的“平均值”



# 期望的性质

---

□ 无条件成立

$$E(kX) = kE(X)$$

$$E(X + Y) = E(X) + E(Y)$$

□ 若 $X$ 和 $Y$ 相互独立

$$E(XY) = E(X)E(Y)$$



# 方差

---

□ 定义  $Var(X) = E\{[X - E(X)]^2\}$

□ 无条件成立  $Var(c) = 0$

$$Var(X + c) = Var(X)$$

$$Var(kX) = k^2 Var(X)$$

□ X和Y独立

$$Var(X + Y) = Var(X) + Var(Y)$$

■ 方差的平方根，称为标准差



# 协方差

---

□ 定义  $Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$

□ 性质:

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(aX + b, cY + d) = acCov(X, Y)$$

$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$



# 协方差和独立、不相关

---

- $X$ 和 $Y$ 独立时,  $E(XY) = E(X)E(Y)$
- 而  $Cov(X, Y) = E(XY) - E(X)E(Y)$
- 从而, 当 $X$ 和 $Y$ 独立时,  $Cov(X, Y) = 0$
  
- 但 $X$ 和 $Y$ 独立这个前提太强, 我们定义: 若  $Cov(X, Y) = 0$ , 称 $X$ 和 $Y$ 不相关。



# 协方差的意义

---

- 协方差是两个随机变量具有相同方向变化趋势的度量；若  $\text{Cov}(X, Y) > 0$ ，它们的变化趋势相同，若  $\text{Cov}(X, Y) < 0$ ，它们的变化趋势相反；若  $\text{Cov}(X, Y) = 0$ ，称  $X$  和  $Y$  不相关。
- 思考：两个随机变量的协方差，是否有上界？



# 协方差的上界

---

- 若  $Var(Y) = \sigma_2^2$      $Var(X) = \sigma_1^2$
- 则  $|Cov(X, Y)| \leq \sigma_1 \sigma_2$
- 当且仅当  $X$  和  $Y$  之间有线性关系时，等号成立。





# 再谈独立与不相关

---

- 因为上述定理的保证，使得“不相关”事实上即“**线性独立**”。
- 即：若 $X$ 与 $Y$ 不相关，说明 $X$ 与 $Y$ 之间没有线性关系(但有可能存在其他函数关系)，不能保证 $X$ 和 $Y$ 相互独立。
- 但对于**二维正态随机变量**， $X$ 与 $Y$ 不相关等价于 $X$ 与 $Y$ 相互独立。



# 相关系数

---

□ 定义 
$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

□ 由协方差上界定理可知,  $|\rho| \leq 1$

□ 当且仅当X与Y有线性关系时, 等号成立

□ 容易看到, 相关系数是标准尺度下的协方差。上面关于协方差与XY相互关系的结论, 完全适用于相关系数和XY的相互关系。



# 协方差矩阵

- 对于n维随机向量 $(X_1, X_2, \dots, X_n)$ ，任意两个元素 $X_i$ 和 $X_j$ 都可以得到一个协方差，从而形成 $n \times n$ 的矩阵；显然，协方差矩阵是对称阵。

$$c_{ij} = E\{[X_i - E(X_i)][X_j - E(X_j)]\} = \text{Cov}(X_i, X_j)$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$



# 思考题

---

- 对称阵的不同特征值对应的特征向量，是否一定正交？



# 矩

---

□ 对于随机变量 $X$ ,  $X$ 的 $k$ 阶原点矩为

$$E(X^k)$$

□  $X$ 的 $k$ 阶中心矩为

$$E\{[X - E(X)]^k\}$$



# 统计参数的总结

---

- 均值(期望, 一阶)
- 方差(二阶)
- 变异系数(Coefficient of Variation)
  - 标准差与平均数的比值称为变异系数, 记为 $C \cdot V$
- 偏度Skew(三阶)
- 峰度Kurtosis(四阶)



# 偏度

---

- 在机率论和统计学中，偏度衡量实数随机变量概率分布的不对称性。偏度的值可以为正，可以为负或者甚至是无法定义。在数量上，偏度为负(负偏态)就意味着在概率密度函数左侧的尾部比右侧的长，绝大多数的值(包括中位数在内)位于平均值的右侧。偏度为正(正偏态)就意味着在概率密度函数右侧的尾部比左侧的长，绝大多数的值(包括中位数在内)位于平均值的左侧。偏度为零就表示数值相对均匀地分布在平均值的两侧，但不一定意味着其为对称分布。



# 偏度公式

□ 其中  $\mu_3$  是三阶中心矩， $\sigma$  是标准差。E 是期望算子。等式的最后以三阶累积量与二阶累积量的 1.5 次方的比率来表示偏度。这和用四阶累积量除去二阶累积量的平方来表示峰度的方法向类似。

□ 偏度有时用  $\text{Skew}[X]$  来表示。

$$\gamma_1 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = \frac{\mu_3}{\sigma^3} = \frac{E[(X - \mu)^3]}{(E[(X - \mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

$$\gamma_1 = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = \frac{E[X^3] - 3\mu E[X^2] + 2\mu^3}{\sigma^3} = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$



# 峰度 $\frac{\mu_4}{\sigma^4}$

- 在更通常的情况下，峰度被定义为四阶累积量除以二阶累积量的平方，它等于四阶中心矩除以概率分布方差的平方再减去3：

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

- $\frac{\mu_4}{\sigma^4}$  也被称为超值峰度(excess kurtosis)。

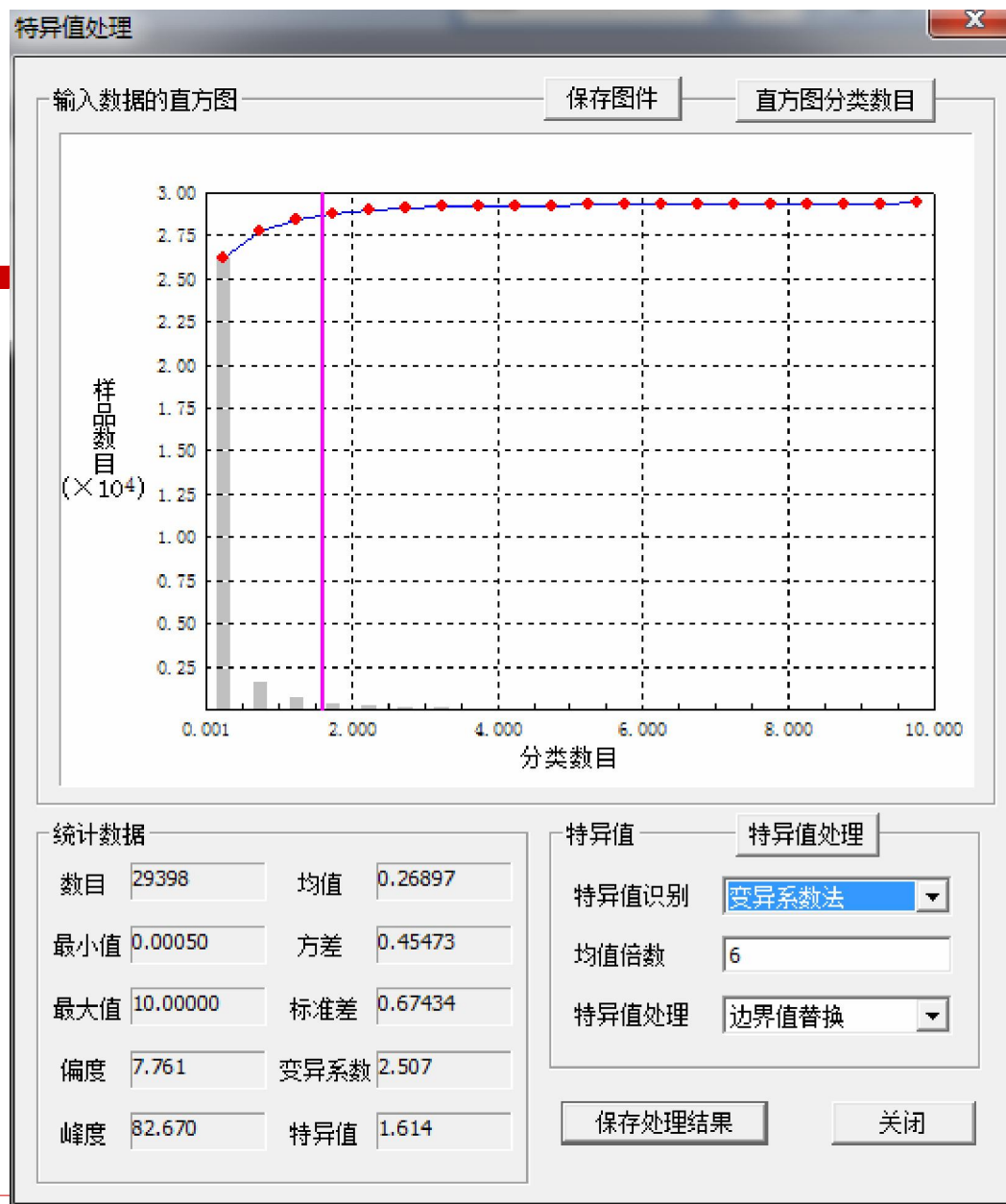
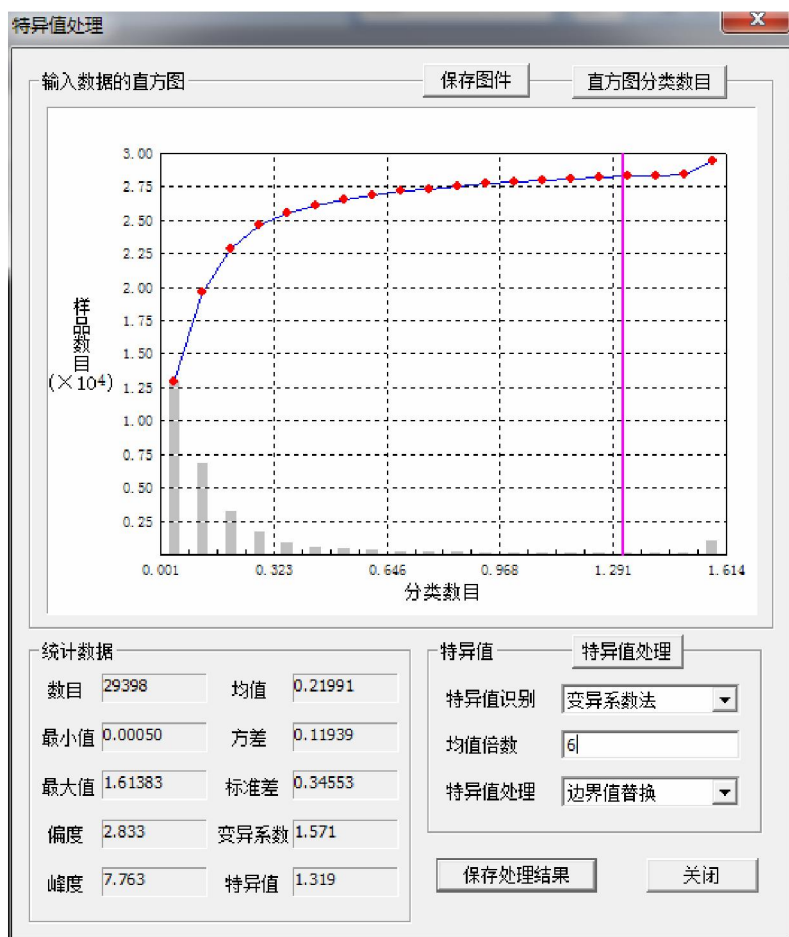
- “减3”是为了让正态分布的峰度为0。

- 如果超值峰度为正，称为尖峰态(leptokurtic)，超值峰度为负，称为低峰态(platykurtic)。



$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2} - 3$$

# 实践中的例子



# 思考

---

□ 设随机变量 $X$ 的期望为 $\mu$ ，方差为 $\sigma^2$ ，对于任意整数 $\varepsilon$ ，试估计概率 $P\{|X-\mu|<\varepsilon\}$ 的上限。

■ 即：随机变量的变化值落在期望值附近的概率



# 解(以连续型随机变量为例)

---

$$\begin{aligned} & P\{|X - \mu| \geq \varepsilon\} \\ &= \int_{|X - \mu| \geq \varepsilon} f(x) dx \\ &\leq \int_{|X - \mu| \geq \varepsilon} \frac{|X - \mu|^2}{\varepsilon^2} f(x) dx \\ &= \frac{1}{\varepsilon^2} \int_{|X - \mu| \geq \varepsilon} (X - \mu)^2 f(x) dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} (X - \mu)^2 f(x) dx \\ &= \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

$$\begin{aligned} & P\{|X - \mu| < \varepsilon\} \\ &= 1 - P\{|X - \mu| \geq \varepsilon\} \\ &\geq 1 - \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$



# 切比雪夫不等式

---

- 设随机变量 $X$ 的期望为 $\mu$ ，方差为 $\sigma^2$ ，对于任意整数 $\varepsilon$ ，有：

$$P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

- 切比雪夫不等式说明， $X$ 的方差越小，事件 $\{|X - \mu| < \varepsilon\}$ 发生的概率越大。即： $X$ 取的值基本上集中在期望 $\mu$ 附近。
- 该不等式进一步说明了方差的含义
  - 该不等式可证明大数定理。



# 大数定理

---

□ 设随机变量  $X_1, X_2, \dots, X_n, \dots$  互相独立, 并且具有相同的期望  $\mu$  和方差  $\sigma^2$ 。作前  $n$  个随机变量的平均  $Y_n = \frac{1}{n} \sum_{i=1}^n X_i$ , 则对于任意整数  $\varepsilon$ , 有

$$\lim_{n \rightarrow \infty} P\{|Y_n - \mu| < \varepsilon\} = 1$$



# 大数定理的意义

---

- 当 $n$ 很大时，随机变量 $X_1, X_2, \dots, X_n$ 的平均值 $Y_n$ 在概率意义下无限接近期望 $\mu$ 。
- 出现偏离是可能的，但这种可能性很小，当 $n$ 无限大时，这种可能性的概率为0。



# 思考题

---

□ 如何证明大数定理？

- 提示：根据 $Y$ 的定义，求出它的期望和方差，带入切比雪夫不等式即可。





# 重要推论

---

□ 一次试验中事件A发生的概率为p；重复n次独立试验中，事件A发生了 $n_A$ 次，则p、n、 $n_A$ 的关系满足：

对于任意整数  $\varepsilon$  ，

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{n_A}{n} - p \right| < \varepsilon \right\} = 1$$



# 伯努利定理

---

- 上述推论是最早的大数定理的形式，称为伯努利定理。该定理表明事件A发生的频率  $n_A/n$  以概率收敛于事件A的概率  $p$ ，以严格的数学形式表达了频率的稳定性。
- 上述事实为我们在实际应用中用频率来估计概率提供了一个理论依据。
  - 回忆一下朴素贝叶斯做垃圾邮件分类的例子，就是用的频率估计的概率。



# 中心极限定理

---

- 设随机变量  $X_1, X_2, \dots, X_n, \dots$  互相独立，服从同一分布，并且具有相同的期望  $\mu$  和方差  $\sigma^2$ ，则随机变量

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma}}$$

的分布收敛到标准正态分布。

- 容易得到： $\sum_{i=1}^n X_i$  收敛到正态分布  $N(n\mu, n\sigma^2)$



# 中心极限定理的意义

---

- 实际问题中，很多随机现象可以看做许多因素的独立影响的综合反应，往往近似服从正态分布。
  - 城市耗电量：大量用户的耗电量总和
  - 测量误差：许多观察不到的、微小误差的总和
    - 注意：是多个随机变量的和才可以，有些问题是乘性误差，则需要鉴别或者取对数后再使用。
  - 线性回归中，将使用该定理论证最小二乘法的合理性



# 样本的统计量

---

□ 设  $X_1, X_2, \dots, X_n$  为一组样本, 则

□ 样本均值  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

□ 样本方差  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

■ 样本方差的分母使用  $n-1$  而非  $n$ , 是为了无偏。



# 样本的矩

---

□ k阶样本原点矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

□ k阶样本中心矩

$$M_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$



# 思考

---

- 随机变量的矩和样本的矩，有什么关系？
- 换个提法：
  - 假设总体服从某参数为  $\theta$  (存在且未知，有可能是值或者向量) 的分布，从总体中抽出一组样本  $X_1, X_2, \dots, X_n$ ，如何估计参数  $\theta$ ？
  - 样本是独立同分布的
  - 可以通过  $X_1, X_2, \dots, X_n$  方便的计算出样本的  $k$  阶矩
  - 假设样本的  $k$  阶矩等于总体的  $k$  阶矩，可估计出总体的参数。



# 矩估计

---

- 设总体的均值为  $\mu$ ，方差  $\sigma^2$ ，（ $\mu$  和  $\sigma$  未知，待求）则有中心距表达式：

$$\begin{cases} E(X) = \mu \\ E(X^2) = \text{Var}(X) + [E(X)]^2 = \sigma^2 + \mu^2 \end{cases}$$

- 根据该总体的一组样本，求得中心距：

$$\begin{cases} A_1 = \frac{1}{n} \sum_{i=1}^n X_i \\ A_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$





# 矩估计的结论

---

□ 根据各自阶的中心矩相等，计算得到：

$$\begin{cases} \mu = \bar{X} \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

□ 由于是根据样本求得的估计结果，根据记号习惯，写作：

$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$



# 矩估计举例

---

□ 在正态分布的总体中采样得到n个样本：  
 $X_1, X_2, \dots, X_n$ ，估计该总体的均值和方差。

□ 解：直接使用矩估计的结论 
$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$



# 极大似然估计

- 设总体分布为  $f(x, \theta)$ ， $X_1, X_2, \dots, X_n$  为该总体采样得到的样本。因为  $X_1, X_2, \dots, X_n$  独立同分布，于是，它们的联合密度函数为：

$$L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k)$$

- 这里， $\theta$  被看做固定但未知的参数；反过来，因为样本已经存在，可以看成  $x_1, x_2, \dots, x_n$  是固定的， $L(x, \theta)$  是关于  $\theta$  的函数，即似然函数。
- 求参数  $\theta$  的值，使得似然函数取极大值，这种方法就是极大似然估计。



# 极大似然估计的具体实践操作

- 在实践中，由于求导数的需要，往往将似然函数取对数，得到对数似然函数；若对数似然函数可导，可通过求导的方式，解下列方程组，得到驻点，然后分析该驻点是极大值点

$$\log L(\theta_1, \theta_2, \dots, \theta_k) = \sum_{i=1}^n \log f(x_i; \theta_1, \theta_2, \dots, \theta_k)$$

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k$$



# 极大似然估计

---

- 找出与样本的分布最接近的概率分布模型。
- 简单的例子
  - 10次抛硬币的结果是：正正反正正正反反正正
- 假设 $p$ 是每次抛硬币结果为正的概率。则：
- 得到这样的实验结果的概率是：

$$\begin{aligned} P &= pp(1-p)ppp(1-p)(1-p)pp \\ &= p^7(1-p)^3 \end{aligned}$$



# 极大似然估计MLE

---

- 目标函数:  $\max P = \max_{0 \leq p \leq 1} p^7 (1-p)^3$
- 最优解是:  $p=0.7$ 
  - 思考: 如何求解?

- 一般形式:  $L_{\bar{p}} = \prod_x p(x)^{\bar{p}(x)}$

$p(x)$ 模型是估计的概率分布

$\bar{p}(x)$ 是实验结果的分布



# 正态分布的极大似然估计

---

- 若给定一组样本 $X_1, X_2, \dots, X_n$ ，已知它们来自于高斯分布 $N(\mu, \sigma)$ ，试估计参数 $\mu, \sigma$ 。



# 按照MLE的过程分析

---

□ 高斯分布的概率密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

□ 将 $X_i$ 的样本值 $x_i$ 带入，得到：

$$L(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$





# 化简对数似然函数

---

$$\begin{aligned}l(x) &= \log \prod_i \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\&= \sum_i \log \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\&= \left( \sum_i \log \frac{1}{\sqrt{2\pi\sigma}} \right) + \left( \sum_i -\frac{(x_i-\mu)^2}{2\sigma^2} \right) \\&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\end{aligned}$$



# 参数估计的结论

---

□ 目标函数  $l(x) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$

□ 将目标函数对参数  $\mu, \sigma$  分别求偏导，很容易得到  $\mu, \sigma$  的式子：

$$\mu = \frac{1}{n} \sum_i x_i$$

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$



# 符合直观想象

---

$$\mu = \frac{1}{n} \sum_i x_i$$

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

- 上述结论和矩估计的结果是一致的，并且意义非常直观：样本的均值即高斯分布的均值，样本的方差即高斯分布的方差。
  - 注：经典意义下的方差，分母是n-1；在似然估计的方法中，求的方差是n
- 该结论将在EM(期望最大化算法)、高斯混合模型中将继续使用。



# 思考

---

- 在西单商场随机挑选100位顾客，测量这100位顾客的身高：
- 若这100个样本服从正态分布 $N(\mu, \sigma)$ ，试估计参数 $\mu$ 和 $\sigma$ 。
- 若样本中存在男性和女性顾客，它们服从 $N(\mu_1, \sigma_1)$ 和 $N(\mu_2, \sigma_2)$ 的分布，试估计 $\mu_1, \sigma_1, \mu_2, \sigma_2$ 。



# 线性代数

---

## □ 方阵的行列式(递归定义)

- 1阶方阵的行列式为该元素本身
- $n$ 阶方阵的行列式等于它的任一行(或列)的各元素与其对应的代数余子式乘积之和。



# 范德蒙行列式 Vandermonde

□ 证明范德蒙行列式 Vandermonde:

$$D_n = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{n \geq i \geq j \geq 1} (x_i - x_j)$$

■ 提示：数学归纳法



# 矩阵的乘法

---

□ A为 $m \times s$ 阶的矩阵，B为 $s \times n$ 阶的矩阵，那么， $C=A \times B$ 是 $m \times n$ 阶的矩阵，其中，

$$c_{ij} = \sum_{k=1}^s a_{ik} b_{kj}$$



# 思考

---

- 根据定义来计算  $C=A \times B$ ，需要  $m*n*s$  次乘法。
  - 即：若  $A$ 、 $B$  都是  $n$  阶方阵， $C$  的计算时间复杂度为  $O(n^3)$
  - 问：可否设计更快的算法？
- 三个矩阵  $A$ 、 $B$ 、 $C$  的阶分别是  $a_0 \times a_1$ ， $a_1 \times a_2$ ， $a_2 \times a_3$ ，从而  $(A \times B) \times C$  和  $A \times (B \times C)$  的乘法次数是  $a_0 a_1 a_2 + a_0 a_2 a_3$ 、 $a_1 a_2 a_3 + a_0 a_1 a_3$ ，二者一般情况是不相等的。
  - 问：给定  $n$  个矩阵的连乘积： $A_1 \times A_2 \times A_3 \dots \times A_n$ ，如何添加括号来改变计算次序，使得乘法的计算量最小？





# 解

---

- 分治法
- 动态规划
- 属于算法的经典问题，将在姊妹班“算法班”中做进一步探讨。



# 矩阵的秩

- 在  $m \times n$  矩阵  $A$  中，任取  $k$  行  $k$  列，不改变这  $k^2$  个元素在  $A$  中的次序，得到  $k$  阶方阵，称为矩阵  $A$  的  $k$  阶子式。
  - 显然， $m \times n$  矩阵  $A$  的  $k$  阶子式有  $C_m^k C_n^k$  个。
- 设在矩阵  $A$  中有一个不等于 0 的  $r$  阶子式  $D$ ，且所有  $r+1$  阶子式(如果存在的话)全等于 0，那么， $D$  称为矩阵  $A$  的最高阶非零子式， $r$  称为矩阵  $A$  的秩，记做  $R(A)=r$ 。
  - $n \times n$  的可逆矩阵，秩为  $n$
  - 可逆矩阵又称满秩矩阵
  - 矩阵的秩等于它行(列)向量组的秩



# 秩与线性方程组的解的关系

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \dots\dots\dots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases} \quad Ax = b$$

□ 对于n元线性方程组 $Ax=b$ ,

- 无解的充要条件是 $R(A) < R(A, b)$
- 有唯一解的充要条件是 $R(A) = R(A, b) = n$
- 有无限多解的充要条件是 $R(A) = R(A, b) < n$



# 推论

---

- $Ax=0$  有非零解的充要条件是  $R(A) < n$
- $Ax=b$  有解的充要条件是  $R(A) = R(A, b)$



# 向量组等价

---

- 向量 $b$ 能由向量组 $A:a_1, a_2, \dots, a_m$ 线性表示的充要条件是矩阵 $A=(a_1, a_2, \dots, a_m)$ 的秩等于矩阵 $B=(a_1, a_2, \dots, a_m, b)$ 的秩。
- 设有两个向量组 $A:a_1, a_2, \dots, a_m$ 及 $B:b_1, b_2, \dots, b_n$ , 若 $B$ 组的向量都能由向量组 $A$ 线性表示, 则称向量组 $B$ 能由向量组 $A$ 线性表示。若向量组 $A$ 与向量组 $B$ 能相互线性表示, 则称两个向量组等价。



# 系数矩阵

□ 把向量组A和B所构成的矩阵依次记做

$A=(a_1, a_2, \dots, a_m)$ 和 $B=(b_1, b_2, \dots, b_n)$ , B组能由A组  
线性表示, 即对每个向量 $b_j$ , 存在  $k_{1j}, k_{2j}, \dots, k_{mj}$

□ 使得

$$b_j = k_{1j}a_1 + k_{2j}a_2 + \dots + k_{mj}a_m = (a_1 \quad a_2 \quad \dots \quad a_m) \begin{pmatrix} k_{1j} \\ k_{2j} \\ \vdots \\ k_{mj} \end{pmatrix}$$

□ 从而得到系数矩阵K

$$(b_1 \quad b_2 \quad \dots \quad b_n) = (a_1 \quad a_2 \quad \dots \quad a_m) \begin{pmatrix} k_{11} & k_{12} & \dots & k_{1n} \\ k_{21} & k_{22} & \dots & k_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m1} & k_{m2} & \dots & k_{mn} \end{pmatrix}$$



# 对 $C=AB$ 的重认识

---

- 由此可知，若 $C=AB$ ，则矩阵 $C$ 的列向量能由 $A$ 的列向量线性表示， $B$ 即为这一表示的系数矩阵。
- 向量组 $B: b_1, b_2, \dots, b_n$ 能由向量组 $A: a_1, a_2, \dots, a_m$ 线性表示的充要条件是矩阵 $A=(a_1, a_2, \dots, a_m)$ 的秩等于矩阵 $(A, B)=(a_1, a_2, \dots, a_m, b_1, b_2, \dots, b_n)$ 的秩，即： $R(A)=R(A, B)$



# 正交阵

---

- 若 $n$ 阶矩阵 $A$ 满足 $A^T A = I$ ，称 $A$ 为正交矩阵，简称正交阵。
  - $A$ 是正交阵的充要条件： $A$ 的列向量都是单位向量，且两两正交。
- $A$ 是正交阵， $x$ 为向量，则 $A \cdot x$ 称作正交变换。
  - 正交变换不改变向量长度





# 思考

---

- 若 $A$ 、 $B$ 都是 $n$ 阶正交阵，那么， $A \times B$ 是正交阵吗？
- 正交阵和对称阵，能够通过何种操作获得一定意义下的联系？



# 特征值和特征向量

---

- $A$  是  $n$  阶矩阵，若数  $\lambda$  和  $n$  维非 0 列向量  $x$  满足  $Ax = \lambda x$ ，那么，数  $\lambda$  称为  $A$  的特征值， $x$  称为  $A$  的对应于特征值  $\lambda$  的特征向量。
- 根据定义，立刻得到  $(A - \lambda I)x = 0$ ，令关于  $\lambda$  的多项式  $|A - \lambda I|$  为 0，方程  $|A - \lambda I| = 0$  的根为  $A$  的特征值；将根  $\lambda_0$  带入方程组  $(A - \lambda I)x = 0$ ，求得到的非零解，即  $\lambda_0$  对应的特征向量。



# 特征值的性质

---

- 设 $n$ 阶矩阵 $A=(a_{ij})$ 的特征值为 $\lambda_1, \lambda_2, \dots, \lambda_n$ , 则
- $\lambda_1 + \lambda_2 + \dots + \lambda_n = a_{11} + a_{22} + \dots + a_{nn}$
- $\lambda_1 \lambda_2 \dots \lambda_n = |A|$

■ 矩阵 $A$ 主行列式的元素和, 称作矩阵 $A$ 的迹。



# 试证明

---

- 已知  $\lambda$  是方阵  $A$  的特征值,
- 则
  - $\lambda^2$  是  $A^2$  的特征值
  - $A$  可逆时,  $\lambda^{-1}$  是  $A^{-1}$  的特征值。



# 不同特征值对应的特征向量

□ 设  $\lambda_1, \lambda_2, \dots, \lambda_m$  是方阵  $A$  的  $m$  个特征值,  $p_1, p_2, \dots, p_m$  是依次与之对应的特征向量, 若  $\lambda_1, \lambda_2, \dots, \lambda_m$  各不相同, 则  $p_1, p_2, \dots, p_m$  线性无关。

## □ 总结

■ 不同特征值对应的特征向量, 线性无关。

■ 若方阵  $A$  是对称阵呢? 结论是否会加强?

□ 协方差矩阵、二次型矩阵、无向图的邻接矩阵等都是对称阵

□ 在谱聚类中将会有所涉及



# 实对称阵不同特征值的特征向量正交

- 令实对称矩阵为 $A$ ，它的两个不同的特征值 $\lambda_1, \lambda_2$ 对应的特征向量分别是 $\mu_1, \mu_2$
- 则有： $A\mu_1 = \lambda_1\mu_1$ ， $A\mu_2 = \lambda_2\mu_2$
- $(A\mu_1)^T = (\lambda_1\mu_1)^T$ ，从而： $\mu_1^T A = \lambda_1\mu_1^T$
- 所以： $\mu_1^T A\mu_2 = \lambda_1\mu_1^T\mu_2$
- 同时， $\mu_1^T A\mu_2 = \mu_1^T (A\mu_2) = \mu_1^T \lambda_2\mu_2 = \lambda_2\mu_1^T\mu_2$
- 所以， $\lambda_1\mu_1^T\mu_2 = \lambda_2\mu_1^T\mu_2$
- 故： $(\lambda_1 - \lambda_2)\mu_1^T\mu_2 = 0$
- 而 $\lambda_1 \neq \lambda_2$ ，所以 $\mu_1^T\mu_2 = 0$ ，即： $\mu_1, \mu_2$ 正交。



# 实对称阵的特征值是实数

---

- 设复数  $\lambda$  为对称阵  $A$  的特征值，复向量  $x$  为对应的特征向量，即  $Ax = \lambda x (x \neq 0)$
- 用  $\bar{\lambda}$  表示  $\lambda$  的共轭复数， $\bar{x}$  表示  $x$  的共轭复向量，而  $A$  是实矩阵，有  $\bar{A} = A$
- 下面给出证明过程。



# 证明

□ 首先  $A\bar{x} = \overline{A\bar{x}} = \overline{Ax} = \overline{\lambda x} = \overline{\lambda}\bar{x}$

□ 因为  $\bar{x}^T (Ax) = \overline{\bar{x}^T (Ax)} = \overline{\bar{x}^T \lambda x} = \overline{\lambda \bar{x}^T x}$

$$\bar{x}^T (Ax) = (\bar{x}^T A^T)x = (A\bar{x})^T x = (\overline{\lambda}\bar{x})^T x = \overline{\lambda}\bar{x}^T x$$

□ 从而

$$\lambda \bar{x}^T x = \overline{\lambda} \bar{x}^T x \Rightarrow (\lambda - \overline{\lambda}) \bar{x}^T x = 0$$

□ 而

$$\bar{x}^T x = \sum_{i=1}^n \overline{x_i} x_i = \sum_{i=1}^n |x_i|^2 \neq 0$$

□ 所以

$$\lambda - \overline{\lambda} = 0 \Rightarrow \lambda = \overline{\lambda}$$





# 利用结论很快得到

---

- 将实数  $\lambda$  带入方程组  $(A - \lambda I)x = 0$ ，该方程组为实系数方程组，因此，实对称阵的特征向量可以取实向量。



# 最终结论

---

□ 设A为n阶对称阵，则必有正交阵P，使得

$$P^{-1}AP = P^T AP = \Lambda$$

■  $\Lambda$  是以A的n个特征值为对角元的对角阵。



# 二次型

---

- 含有 $n$ 个变量的二次齐次函数，称为二次型；
- 一个二次型对应一个对称阵；
- 而对称阵可以由正交阵对角化，
- 从而二次型可以化成只有 $n$ 个变量平方项的标准型，而这个正交阵，对应着坐标系的旋转变化。



# 正定阵

---

- 对于 $n$ 阶方阵 $A$ ，若任意 $n$ 阶向量 $x$ ，都有 $x^T Ax > 0$ ，则称 $A$ 是正定阵。
  - 若条件变成 $x^T Ax \geq 0$ ，则 $A$ 称作半正定阵
  - 类似还有负定阵，半负定阵。



# 正定阵的判定

- 对称阵A为正定阵;
- A的特征值都为正;
- A的顺序主子式大于0;
- 以上三个命题等价。

$$(a_{11}) \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$



# 参考文献

---

- 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
- 同济大学数学系 编，工程数学线性代数(第五版)，高等教育出版社，2007
- Ulrike von Luxburg, A tutorial on spectral clustering, 2007



---

感谢大家!

恳请大家批评指正!

