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

历史遗留问题

- 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 $\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x$ 的极限存在，定义为e。

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

极限存在的证明

- 根据前文中 $a_n=\left(1+\frac{1}{n}\right)^n$ 的二项展开式，已经证明数组 $\{a_n\}$ 单增有上界，因此，必有极限。
- 同时：

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

$$\lim_{n\to\infty}\left(1+\frac{1}{n+1}\right)^n=\lim_{n\to\infty}\frac{\left(1+\frac{1}{n+1}\right)^{n+1}}{1+\frac{1}{n+1}}=\frac{\lim\limits_{n\to\infty}\left(1+\frac{1}{n+1}\right)^{n+1}}{\lim\limits_{n\to\infty}\left(1+\frac{1}{n+1}\right)}=\frac{e}{1+0}=e$$

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n+1}=\lim_{n\to\infty}\left(\left(1+\frac{1}{n}\right)^n\left(1+\frac{1}{n}\right)\right)=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n\cdot\lim_{n\to\infty}\left(1+\frac{1}{n}\right)=e\cdot(1+0)=e$$

- 根据两边夹定理，函数 $f(x)=\left(1+\frac{1}{x}\right)^x$ 的极限存在，为e.

期望

- 离散型 $$E(X)=\sum_i x_i p_i$$
- 连续型 $$E(X)=\int_{-\infty}^{\infty} xf(x)dx$$
- 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

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

  - 反之不成立。事实上，若E(XY)=E(X)E(Y)，只能说明X和Y不相关。
  - 关于不相关和独立的区别，稍后马上给出。

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
$$Var(X + Y) = Var(X) + Var(Y)$$
  - 此外，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$Cov(X,Y)>0$，它们的变化趋势相同，若$Cov(X,Y)<0$，它们的变化趋势相反；若$Cov(X,Y)=0$，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 $Var(Y)=\sigma_2^2 \quad Var(X)=\sigma_1^2$
- 则 $|Cov(X,Y)| \le \sigma_1\sigma_2$
- 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 $\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
- 由协方差上界定理可知，$|\rho| \le 1$
- 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...X_n)，任意两个元素Xi和Xj都可以得到一个协方差，从而形成n*n的矩阵；显然，协方差矩阵是对称阵。

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

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

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

$$E\left(X^k\right)$$

- X的k阶中心矩为

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

统计参数的总结

- 均值(期望，一阶)
- 方差(标准差，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·V
- 偏度Skew(三阶)
- 峰度Kurtosis(四阶)

偏度

- 偏度衡量随机变量概率分布的不对称性。
- 偏度的值可以为正，可以为负或者无定义。
- 偏度为负(负偏态)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)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，但不一定意味着一定是对称分布。

偏度公式

- 其中μ_3是三阶中心矩，σ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Skew[X]来表示。

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

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

峰度 $\frac{\mu_4}{\sigma^4}$

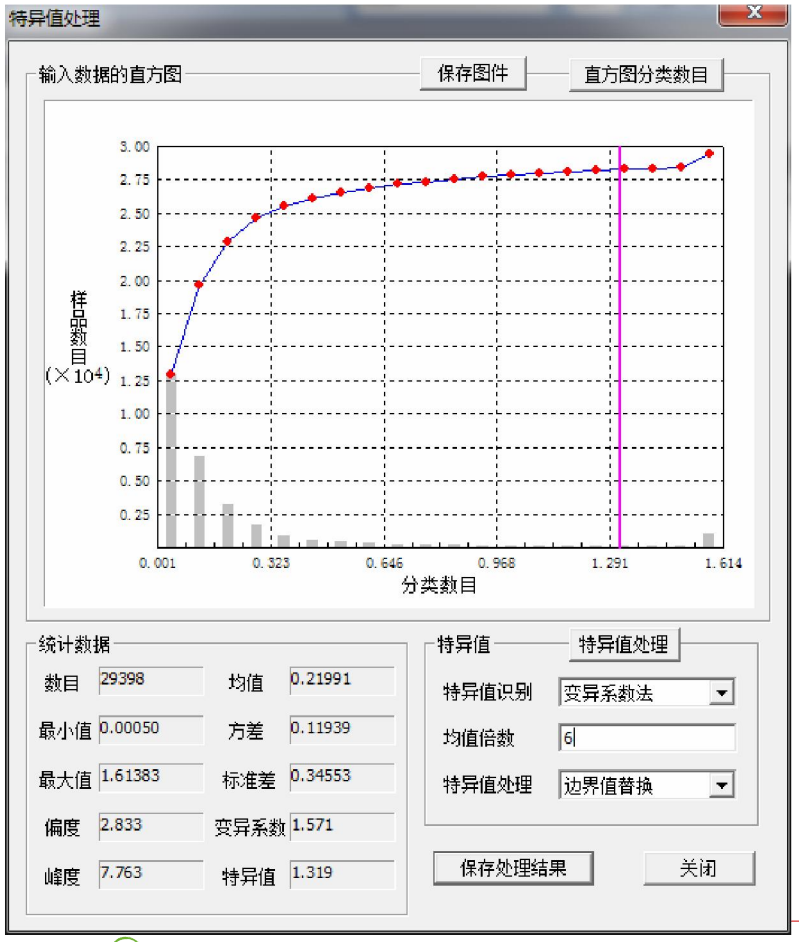
- 峰度通常被定义四阶中心矩除以方差的平方再减去3：

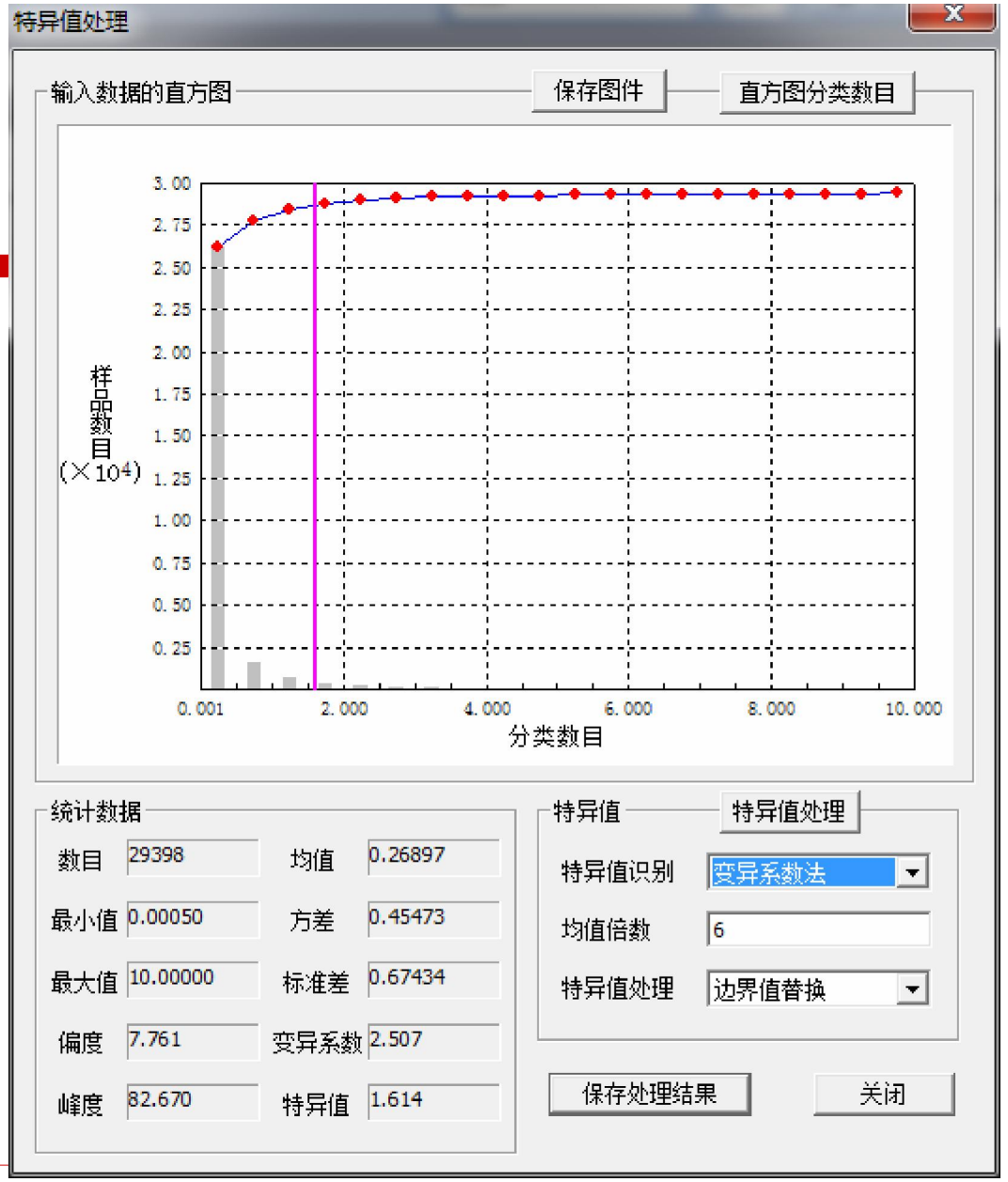
$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

- $\frac{\mu_4}{\sigma^4}$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)。

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

实践中的例子


特异值处理
输入数据的直方图
保存图件
直方图分类数目
样品数目 (×10⁴)
分类数目
统计数据
数目 29398
均值 0.21991
最小值 0.00050
方差 0.11939
最大值 1.61383
标准差 0.34553
偏度 2.833
变异系数 1.571
峰度 7.763
特异值 1.319
特异值
特异值处理
特异值识别 变异系数法
均值倍数 6
特异值处理 边界值替换
保存处理结果
关闭


特异值处理
输入数据的直方图
保存图件
直方图分类数目
样品数目 (×10⁴)
分类数目
统计数据
数目 29398
均值 0.26897
最小值 0.00050
方差 0.45473
最大值 10.00000
标准差 0.67434
偏度 7.761
变异系数 2.507
峰度 82.670
特异值 1.614
特异值
特异值处理
特异值识别 变异系数法
均值倍数 6
特异值处理 边界值替换
保存处理结果
关闭

思考

- 1、给定两个随机变量X和Y，如何度量这两个随机变量的“距离”？

- 2、设随机变量X的期望为 μ，方差为 σ^2，对于任意整数 ε，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)

$$P\{|X-\mu|\geq\varepsilon\}$$

$$=\int_{|X-\mu|\geq\varepsilon} f(x)dx$$

$$\leq\int_{|X-\mu|\geq\varepsilon}\frac{|X-\mu|^2}{\varepsilon^2}f(x)dx$$

$$=\frac{1}{\varepsilon^2}\int_{|X-\mu|\geq\varepsilon}(X-\mu)^2 f(x)dx$$

$$\leq\frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}(X-\mu)^2 f(x)dx$$

$$=\frac{\sigma^2}{\varepsilon^2}$$

$$P\{|X-\mu|<\varepsilon\}$$

$$=1-P\{|X-\mu|\geq\varepsilon\}$$

$$\geq 1-\frac{\sigma^2}{\varepsilon^2}$$

切比雪夫不等式

- 设随机变量X的期望为μ，方差为σ^2，对于任意整数ε，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- μ|＜ε}发生的概率越大。即：X取的值基本上集中在期望μ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$X_1,X_2...X_n...$互相独立，并且具有相同的期望μ和方差σ^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

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

大数定理的意义

- 当n很大时，随机变量$X_1,X_2...X_n$的平均值Y_n在概率意义下无限接近期望 μ 。
  - 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ε，

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

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$X_1, X_2 \ldots X_n \ldots$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，…，X_n为一组样本，则
- 样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$
- 样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$
  - 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

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

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θ(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...，X_n，如何估计参数θ？
  - 样本是独立同分布的
  - 可以通过X_1，X_2...，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 μ，方差 σ^2，(μ 和 σ 未知，待求)则有中心距表达式：

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

- 根据该总体的一组样本，求得中心距：

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

矩估计的结论

- 根据各自阶的中心矩相等，计算得到：

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

- 由于是根据样本求得的估计结果，根据记号习惯，写作：

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

例：正态分布的矩估计

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

- 解：直接使用矩估计的结论

$$\begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \end{cases}$$

例：均匀分布的矩估计

- 设X1,X2,…,Xn为定义在[a,b]上的均匀分布的总体采样得到的样本，求a，b。
- 解：

已知均匀分布的均值和方差为 $\begin{cases} E(X)=\dfrac{a+b}{2} \\ Var(X)=\dfrac{(b-a)^2}{12} \end{cases}$

矩估计要求满足 $\begin{cases} \hat{\mu}=\overline{X} \\ \hat{\sigma}^2=\dfrac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \end{cases}$

从而： $\begin{cases} \dfrac{a+b}{2}=\hat{\mu} \\ \dfrac{(b-a)^2}{12}=\hat{\sigma}^2 \end{cases} \Rightarrow \begin{cases} a=\hat{\mu}-\sqrt{3}\hat{\sigma} \\ b=\hat{\mu}+\sqrt{3}\hat{\sigma} \end{cases}$

极大似然估计

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

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

- 这里，θ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$x_1, x_2 \ldots x_n$是固定的，$L(x, \theta)$是关于θ的函数，即似然函数。
- 求参数θ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

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

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

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：$\max P = \max_{0 \le p \le 1} p^7 (1-p)^3$
- 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$X_1,X_2...X_n$，已知它们来自于高斯分布$N(\mu,\sigma)$，试估计参数μ,σ。

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

$$l(x)=\log\prod_i \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\sum_i \log\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\left(\sum_i \log\frac{1}{\sqrt{2\pi}\sigma}\right)+\left(\sum_i -\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

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

参数估计的结论

- 目标函数 $$l(x)=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_i\left(x_i-\mu\right)^2$$
- 将目标函数对参数 μ，σ 分别求偏导，很容易得到 μ，σ 的式子：

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

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

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μ和σ。
- 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×s阶的矩阵，B为s×n阶的矩阵，那么，C=A×B是m×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×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×B)×C和A×(B×C)的乘法次数是$a_0a_1a_2+a_0a_2a_3$、$a_1a_2a_3+a_0a_1a_3$，二者一般情况是不相等的。
  - 问：给定n个矩阵的连乘积：$A_1 \times A_2 \times A_3 \ldots\ldots \times A_n$，如何添加括号来改变计算次序，使得乘法的计算量最小？

解

- 矩阵乘法C=A×B优化问题
  - 分治法
- 矩阵连乘的加括号最优策略
  - 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×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×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×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

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

- 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:$a_1,a_2,...,a_m$线性表示的充要条件是矩阵$A=(a_1,a_2,...a_m)$的秩等于矩阵$B=(a_1,a_2,...a_m,b)$的秩。
- 设有两个向量组A:$a_1,a_2,...,a_m$及B:$b_1,b_2,...,b_n$，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,...,a_m)和B=(b_1,b_2,...,b_n)，B组能由A组线性表示，即对每个向量b_j，存在 $k_{1j},k_{2j},\cdots k_{mj}$
- 使得

$$b_j = k_{1j}a_1 + k_{2j}a_2 + \cdots + k_{mj}a_m = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \end{pmatrix} \begin{pmatrix} k_{1j} \\ k_{2j} \\ \vdots \\ k_{mj} \end{pmatrix}$$

- 从而得到系数矩阵K

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

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: $b_1,b_2,...,b_n$能由向量组A: $a_1,a_2,...,a_m$线性表示的充要条件是矩阵$A=(a_1,a_2,...,a_m)$的秩等于矩阵$(A,B)=(a_1,a_2,...,a_m,b_1,b_2,...,b_n)$的秩，即：$R(A)=R(A,B)$

正交阵

- 若n阶矩阵A满足$A^TA=I$，成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·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×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λ和n维非0列向量x满足$Ax=\lambda x$，那么，数λ称为A的特征值，x称为A的对应于特征值λ的特征向量。
  - 根据定义，立刻得到$(A-\lambda I)x=0$，令关于λ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λ_0带入方程组$(A-\lambda I)x=0$，求得到的非零解，即λ_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$\lambda_1, \lambda_2, \ldots \lambda_n$,则
- $\lambda_1+\lambda_2+\ldots+\lambda_n=a11+a22+\ldots+ann$
- $\lambda_1 \lambda_2 \ldots \lambda_n=|A|$
  - 矩阵A主行列式的元素和，称作矩阵A的迹。

思考

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

不同特征值对应的特征向量

- 设$\lambda_1, \lambda_2, \ldots, \lambda_m$是方阵A的m个特征值，$p_1, p_2, \ldots, p_m$是依次与之对应的特征向量，若$\lambda_1, \lambda_2, \ldots, \lambda_m$各不相等，则$p_1, p_2, \ldots, p_m$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λ_1 λ_2对应的特征向量分别是μ_1 μ_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：$\mu_1^TA=\lambda_1\mu_1^T$
- 所以：$\mu_1^TA\mu_2=\lambda_1\mu_1^T\mu_2$
- 同时，$\mu_1^TA\mu_2=\mu_1^T(A\mu_2)=\mu_1^T\lambda_2\mu_2=\lambda_2\mu_1^T\mu_2$
- 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：μ_1，μ_2正交。

实对称阵的特征值是实数

- 设复数λ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$\overline{\lambda}$表示λ的共轭复数，$\overline{x}$表示x的共轭复向量，而A是实矩阵，有$\overline{A}=A$
- 下面给出证明过程。

证明

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

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

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

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

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

利用上述结论很快得到

- 将实数λ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

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

  - Λ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$x^TAx>0$，则称A是正定阵。
  - 若条件变成$x^TAx\geqslant 0$，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！