第九章 概率模型

第 24 讲 参数估计

黄定江

DaSE @ ECNU
djhuang@dase.ecnu.edu.cn

24.1.1 矩估计

定义 1

θ 的矩估计定义为 $\hat{\theta}_n$，使得

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1, \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2, \\ &\vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k. \end{aligned} \tag{1}$$

24.1.1 矩估计举例

例 1

令 $X_1, \cdots, X_n \sim Bernoulli(p)$. 则 $\alpha_1 = E_p(X) = p$ 且 $\hat{\alpha}_1 = n^{-1}\sum_{i=1}^n X_i$. 让它们相等可以得到估计值

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

24.1.1 矩估计举例

例 2

令 $X_1, \cdots, X_n \sim Normal(\mu, \sigma^2)$. 则 $\alpha_1 = E_\theta(X) = \mu$ 且 $\alpha_2 = E_\theta(X_1^2) = V_\theta(X_1) + (E_\theta(X))^2 = \sigma^2 + \mu^2$. 现在需要解下述方程：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

这是由两个方程组成含有两个未知参数的方程组。它的解为

$$\hat{\mu} = \bar{X},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

24.2.1 极大似然估计

若总体 X 属离散型，其分布律 $PX=x=p(x;\theta),\theta\in\Theta$ 的形式为已知，θ 为待估参数，Θ 是 θ 可能取值的范围，设 $X_1,X_2,\cdots,X_n$ 是来自 X 的样本，则 $X_1,X_2,\cdots,X_n$ 的联合分布律为

$$\prod_{i=1}^{n}p(x_i;\theta)$$

又设 $x_1,x_2,\cdots,x_n$ 是相应于样本 $X_1,X_2,\cdots,X_n$ 值。易知样本 $X_1,X_2,\cdots,X_n$ 取到观察值 $x_1,x_2,\cdots,x_n$ 的概率，亦即事件 $\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$ 发生的概率为

$$L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}p(x_i;\theta),\theta\in\Theta$$

这一概率随 θ 的取值而变化，它是 θ 的函数，$L(\theta)$ 称为样本的似然函数 (注意，这里 $x_1,x_2,\cdots,x_n$ 是已知的样本值，它们都是常数)。

24.2.1 似然函数和极大似然估计的定义

定义 2

令 $\mathbf{X}_1, \cdots, \mathbf{X}_n$ 独立同分布于概率密度函数 $p(x|\theta)$。似然函数定义为

$$\mathcal{L}(\theta) = p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\theta) \tag{2}$$

有时也记为 $\mathcal{L}(\theta|\mathcal{D})$，表示似然函数为在给定数据 $\mathcal{D}$ 的情况下，参数 θ 的函数。

定义 3

极大似然估计 *MLE*，记为 $\hat{\theta}_n$，是使得 $\mathcal{L}(\theta)$ 最大的 θ 值。

24.2.1 极大似然估计举例：高斯分布

首先给出以下高斯分布的概率密度函数。其中 μ 为均值，σ^2 为方差。

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

令 $x_1,\cdots,x_N\sim\mathcal{N}(\mu,\sigma^2)$，参数为 μ,σ^2，似然函数为

$$\begin{aligned}\ell(\mu,\sigma^2)&=\sum_{i=1}^{N}\log p(x_i|\mu,\sigma)\\&=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2-\frac{N}{2}\log\sigma^2-\frac{N}{2}\log 2\pi\\&=-\frac{NS^2}{2\sigma^2}-\frac{N(\overline{x}-\mu)^2}{2\sigma^2}-N\log\sigma-\frac{N}{2}\log 2\pi\end{aligned}$$

24.2.1 极大似然估计举例：高斯分布

其中 $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ 为样本均值，$S^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ 为样本方差

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{N}(x_i - \bar{x} + \bar{x} - \mu)^2 = NS^2 + N(\bar{x}-\mu)^2$$

对 log 似然函数求极值点，即分别对 μ 和 σ 求一阶导数为 0。

24.2.1 极大似然估计举例：高斯分布

解方程

$$\begin{cases} \dfrac{\partial \ell(\mu,\sigma)}{\partial \mu} = \dfrac{N(x-\overline{x})}{\sigma^2} = 0 \\ \dfrac{\partial \ell(\mu,\sigma)}{\partial \sigma} = -\dfrac{N}{\sigma} + \dfrac{NS^3}{\sigma^3} = 0 \end{cases}$$

得到

$$\begin{cases} \hat{\mu} = \overline{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i \\ \hat{\sigma}^2 = S^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i-\overline{x})^2 = (\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i^2) - \overline{x}^2 \end{cases}$$

可以证明，这是似然函数的全局最大值。

24.2.1 极大似然估计举例：Bernoulli 分布

首先给出 Bernoulli 分布的概率密度函数

$$\mathrm{Ber}(x|\theta) = \theta^x (1-\theta)^{1-x}$$

假设我们投掷硬币 N 次，并记录每次投掷结果的序列，用 $\mathcal{D} = x_1, \cdots, x_N$ 表示，则概率函数为 $\mathrm{Ber}(x_i|\theta)$。似然函数为

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^{N} \log \mathrm{Ber}(x_i|\theta) \\ &= \sum_{i=1}^{N} \log(\theta^{x_i}(1-\theta)^{1-x_i}) = N_1 \log\theta + N_2 \log(1-\theta)\end{aligned}$$

其中

$$\begin{cases} N_1 = \sum_{i=1}^{N} x_i & \text{实验中结果为 1 的次数} \\ N_2 = \sum_{i=1}^{N} (1-x_i) & \text{实验中结果为 0 的次数} \end{cases}$$

24.2.1 极大似然估计举例：Bernoulli 分布

所以

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_2}{1-\theta} = 0 \implies \hat{\theta} = \frac{N_1}{N_1 + N_2} = \frac{N_1}{N}$$

$$\mathrm{Bin}(x|n;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

共进行 N 次实验，第 i 次实验中抛掷了 n_i 次硬币，其中 x_i 枚硬币正面朝上。则似然函数为

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^{N} \mathrm{Bin}(x_i|n_i;\theta) \\ &= \prod_{i=1}^{N} \binom{n_i}{x_i} \theta^{x_i} (1-\theta)^{n_i - x_i} \propto \theta^{N_1} (1-\theta)^{N_2} \end{aligned}$$

其中

$$\begin{cases} N_1 = \sum_{i=1}^{N} x_i \\ N_2 = \sum_{i=1}^{N} (n_i - x_i) \end{cases}$$

24.2.1 极大似然估计举例：Bernoulli 分布

对似然函数取对数:

$$\log \mathcal{L} \propto N_1 \log \theta + N_2 \log(1-\theta)$$

求导数为 0 的点:

$$\frac{N_1}{\theta} - \frac{N_2}{1-\theta} = 0$$

解得:

$$\hat{\theta} = \frac{N_1}{N_1 + N_2}$$

参数估计值与 Bernoulli 分布的估计一样。

24.2.1 极大似然估计举例：Multinoulli 分布

$$Mu(x|N,\theta) = \binom{N}{x_1 \cdots x_K} \prod_{k=1}^{K} \theta_k^{x_k}$$

$$\binom{N}{x_1 \cdots x_K} = \frac{N!}{x_1! \cdots x_K!}$$

假设我们投掷一个有 K 面的骰子，共进行了 N 次试验，并记每次投掷结果的序列，用 $\mathcal{D} = x_1, \cdots, x_N$ 表示，$x_i \in 1, \cdots, K$, 则似然函数为：

$$l(\theta) = \log p(\mathcal{D}|\theta) = \sum_{k=1}^{K} N_k \log \theta_k$$

其中 $N_k = \sum_{i=1}^{N} \mathbf{1}(x_i = k)$ 表示 N 次此试验中出现 k 的次数，这是带有约束 $\sum_{k=1}^{K} \theta_k = 1$ 的优化问题，采用拉格朗日乘子法，得到

$$l(\theta,\lambda) = \sum_{k=1}^{K} N_k \log \theta_k + \lambda(1 - \sum_{k=1}^{K} \theta_k)$$

24.2.1 极大似然估计举例：Multinoulli 分布

$$\begin{cases} \dfrac{\partial l(\theta,\lambda)}{\partial \lambda} = 1 - \sum_{k=1}^{K} \theta_k = 0 \\ \dfrac{\partial l(\theta_k,\lambda)}{\partial \theta} = \dfrac{N_k}{\theta_k} - \lambda = 0 \end{cases}$$

因此:

$$\theta_k \propto N_k$$

解得:

$$\hat{\theta}_k = \frac{N_k}{N}$$

24.2.1 极大似然估计举例：线性回归

最简单的回归模型是线性模型，我们假设

$$
\begin{aligned}
y &= f(x) + \varepsilon \\
&= w^T x + \varepsilon
\end{aligned}
$$

其中 w 称为权重向量，ε 为线性预测和真值之间的残差。
由于 $y|x \sim \mathcal{N}(f(x), \sigma^2)$ ，则 $p(y|x;\theta) \sim \mathcal{N}(y;\ w^T x, \sigma^2)$ 其中模型的参数为 $\theta = (w, \sigma^2)$
极大似然估计定义为

$$
\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}|\theta)
$$

其中似然函数

$$
\ell(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(y_i|x_i;\theta)
$$

24.2.1 极大似然估计举例：线性回归

极大似然可等价地写成极小负 log 似然损失 (negative log likelihood，NLL)

$$\mathsf{NLL}(\theta)=\sum_{i=1}^{N}-\log p(y_i|x_i;\theta)$$

将概率模型 $p(y_i|x_i,w,\sigma^2)=\mathcal{N}(y_i|w^Tx_i,\sigma^2)$ 代入，似然函数为

$$\begin{aligned}\ell(\theta)&=\sum_{i=1}^{N}\log\left[(\frac{1}{2\pi\sigma^2})^{1/2}\exp(-\frac{1}{2\sigma^2}(y_i-w^Tx_i)^2)\right]\\&=-\frac{N}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{N}(y_i-w^Tx_i)^2}_{\mathsf{RSS}(w)}\end{aligned}$$

其中 RSS 表示残差平方和 (residual sum of squares)，RSS/N 为平均平方误差 (MSE)，也可以写成残差向量的 L2 模，即

$$\mathsf{RSS}(w)=|\varepsilon|_2^2=\sum_{i=1}^{N}\varepsilon^2,\qquad \varepsilon_i=y_i-w^Tx_i$$

24.2.1 极大似然估计举例：线性回归

将 NLL 写成矩阵形式

$$\begin{aligned}\mathrm{NLL}(w,\sigma) &= \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - w^T x_i)^2 \\ &= \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)\end{aligned}$$

只取与 w 有关的项，得到

$$\mathrm{NLL}(W) = w^T(X^T X)w - 2w^T(X^T y)$$

求梯度为 0 的点

$$\frac{\partial}{\partial w}\mathrm{NLL}(w) = wX^T Xw - 2X^T y = 0 \implies X^T Xw = X^T y$$

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T y$$

其中 OLS 指的是普通最小二乘 (Ordinary least squares)

24.2.1 极大似然估计举例：线性回归

对参数 σ

$$\mathsf{NLL}(\hat{w},\sigma)=\frac{N}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}(y-X\hat{w})^T(y-X\hat{W})$$

$$\frac{\partial}{\partial\sigma}\mathsf{NLL}(\hat{w},\sigma^2)=\frac{N}{\sigma}-\frac{1}{\sigma^3}(y-X\hat{w})^T(y-X\hat{W})=0$$

得到

$$\hat{\sigma^2}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{w}^Tx_i)^2=\frac{1}{N}(y-X\hat{w})^T(y-X\hat{W})$$

当样本数目 N 较小时，可采用 OLS 结论，用矩阵 QR 分解分解得到优化解。
当样本数目 N 较大时，可采用随机梯度下降方法优化求解 (略)。

24.2.1 极大似然估计的性质

极大似然估计的特征有：

- 极大似然估计是相合估计：$\widehat{\theta}_n \xrightarrow{P} \theta_*$，其中，$\theta_*$ 表示参数 θ 的真实值。
- 极大似然估计是同变估计：如果 $\widehat{\theta}_n$ 是 θ 的极大似然估计，则 $g(\widehat{\theta}_n)$ 是 $g(\theta)$ 的极大似然估计.
- 极大似然估计是渐近正态的：$\left(\widehat{\theta}_n - \theta_*\right) / \widehat{\mathrm{se}} \rightsquigarrow N(0,1)$。同时，估计的标准差 $\hat{se}$ 可以解出来。
- 极大似然估计是渐近最优或有效的：这表示，在所有表现优异的估计中，极大似然估计的方差最小，至少对大样本这肯定成立。
- 极大似然估计接近于贝叶斯估计。

24.3.1 极大后验估计前言

若有足够多的样本，则极大似然估计可以准确的估计参数。但是，当只有少量训练样本时，极大似然估计会获得不准确的结果。极大后验估计因此提出。

24.3.1 极大后验估计理论知识

先验概率 (Prior probability): 在贝叶斯统计中，先验概率分布，即关于某个变量 X 的概率分布，是在获得某些信息或者依据前，对 X 之不确定性所进行的猜测。这是对不确定性 (而不是随机性) 赋予一个量化的数值的表征，这个量化数值可以是一个参数，或者是一个潜在的变量。先验概率仅仅依赖于主观上的经验估计，也就是事先根据已有的知识的推断。例如，X 可以是投一枚硬币，正面朝上的概率，显然在我们未获得任何其他信息的条件下，我们会认为 $P(X) = 0.5$；再比如上面例子中的，$P(G) = 0.4$。
似然函数 (Likelihood Function): 似然函数也称作似然，是一个关于统计模型参数的函数。也就是这个函数中自变量是统计模型的参数。对于观测结果 x，在参数集合 θ 上的似然，就是在给定这些参数值的基础上，观察到的结果的概率 $L(\theta) = P(x|\theta)$。也就是说，似然是关于参数的函数，在参数给定的条件下，对于观察到的 x 的值的条件分布。似然函数在统计推断中发挥重要的作用，因为它是关于统计参数的函数，所以可以用来对一组统计参数进行评估，也就是说在一组统计方案的参数中，可以用似然函数做筛选。

24.3.1 极大后验估计理论知识

后验概率 (Posterior probability)：后验概率是关于随机事件或者不确定性断言的条件概率，是在相关证据或者背景给定并纳入考虑之后的条件概率。后验概率分布就是未知量作为随机变量的概率分布，并且是在基于实验或者调查所获得的信息上的条件分布。后验概率是关于参数 θ 在给定的证据信息 X 下的概率，即 $P(\theta|X)$ 。若对比后验概率和似然函数，似然函数是在给定参数下的证据信息 X 的概率分布，即 $P(X|\theta)$ 。

后验概率与似然函数关系二者有如下关系：我们用 $P(\theta)$ 表示概率分布函数，用 $P(X|\theta)$ 表示观测值 X 的似然函数。后验概率定义为 $P(\theta|X)=\dfrac{P(X|\theta)P(\theta)}{P(X)}$，注意这也是贝叶斯定理所揭示的内容。鉴于分母是一个常数，上式可以表达成如下比例关系 (而且这也是我们更多采用的形式)：后验概率 $\propto$ 似然 $\times$ 先验概率。

24.3.1 极大后验估计举例：高斯先验

考虑偏向小的系数值，从而得到比较平滑的曲线的 0 均值高斯先验 $w_j \sim N(0, \tau^2)$

$$p(\mathbf{w}) = \prod_{j=1}^{D} N(w_j|0, \tau^2) \propto \exp(-\frac{1}{2\tau^2}\sum_{j=1}^{D}\mathbf{w}_j^2) = \exp(-\frac{1}{2\tau^2}[\mathbf{w}^T\mathbf{w}])$$

其中 $1/\tau^2$ 控制先验的强度。
此时针对一个样本的似然函数为：$p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma^2) = N((y_i|\mathbf{w}^{\mathbf{T}}\mathbf{x}_i, \sigma^2)$
针对整个数据集的似然函数为：

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, w_0, \sigma^2) = N(\mathbf{y}|\mathbf{X}\mathbf{w} + w_0\mathbf{1}_N, \sigma^2\mathbf{1}_N)$$
$$\propto \exp(-\frac{1}{2\sigma^2}[(y - (\mathbf{X}\mathbf{w} + w_0\mathbf{1}_N))^T(y - (\mathbf{X}\mathbf{w} + w_0\mathbf{1}_N))])$$

24.3.1 极大后验估计举例：高斯先验

则由贝叶斯公式知后验概率为：

$$p(\mathbf{w}, w_0|\mathbf{X}, \mathbf{y}, \sigma^2) \propto \exp(-\frac{1}{2\sigma^2}[(y - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_N)^T(y - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_N)] - \frac{1}{2\tau^2}[\mathbf{w}^T\mathbf{w}])$$

则极大后验估计等价于最小化的目标函数如下：

$$\begin{aligned} J(\mathbf{w}) &= \sum_{i=1}^{N}(y_i - (\mathbf{w}^T\mathbf{x}_i) + w_0))^2 + \lambda \|\mathbf{w}\|_2^2 \\ &= (\mathbf{y} - (\mathbf{X}\mathbf{w} + w_0 + \mathbf{1}_N))^T(\mathbf{y} - (\mathbf{X}\mathbf{w} + w_0 + \mathbf{1}_N)) + \lambda\mathbf{w}^T\mathbf{w} \end{aligned}$$

其中 $\lambda = \sigma^2/\tau^2$

这种形式称为岭回归，或正则化的最小二乘。注意 w_0 没有被正则 (w_0 只影响函数的高度，不影响复杂性)。

24.3.1 极大后验估计举例：Laplace 先验

Laplace 分布：

$$\mathrm{Lap}(x|\mu,b)=\frac{1}{2b}\exp\{-\frac{|x-\mu|}{b}\}$$

假设线性回归中参数的先验为 Laplace 先验：

$$p(\mathbf{W}|\lambda)=\prod_{j=1}^{D}\mathrm{Lap}(w_j|0,\frac{1}{\lambda})\propto\prod_{j=1}^{D}\exp(-\lambda\,|w_j|)=\exp(-\lambda\sum_{j=1}^{D}|w_j|)$$

似然为：

$$p(y_i|\mathbf{x}_i,\mathbf{W},\sigma^2)=\mathcal{N}(y_i|\mathbf{w}^T\mathbf{x}_i,\sigma^2)\propto\exp(-\frac{1}{2\sigma^2}(y_i-(\mathbf{w}^T\mathbf{x}_i+w_0))^2)$$

24.3.1 极大后验估计举例：Laplace 先验

后验为：

$$p(\mathbf{w}, w_0|\mathbf{X}, \mathbf{y}, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - (\mathbf{w}^T\mathbf{x}_i + w_0))^2 - \sum_{j=1}^{D}\lambda\,|w_j|\right)$$

极大后验估计 MAP 等价于 L1 正则的线性回归（Lasso）：

$$J(\mathbf{w}) = \underbrace{\sum_{i=1}^{N}(y_i - (\mathbf{w}^T\mathbf{x}_i + w_0))^2}_{RSS(\mathbf{w})} + \lambda \underbrace{|\mathbf{w}|}_{\text{正则项复杂性惩罚}}$$

当 λ 取合适值时，$\mathbf{w}$ 变得稀疏 (有些系数为 0)，但是相比岭回归，优化计算更复杂。

24.4.1 贝叶斯推断引言

- 回顾之前讲过的贝叶斯公式：

$$P(\theta \mid x) = \frac{P(x \mid \theta)\pi(\theta)}{P(x)} = \frac{P(x \mid \theta)\pi(\theta)}{\int P(x \mid \theta)\pi(\theta)\,d\theta}$$

- 若要求一个未知概率分布，该分布由参数 θ 决定，根据经验，若能估计 θ 可能的取值，即 θ 的概率分布，我们就能解决该问题。θ 的概率分布称之为先验分布 $\pi(\theta)$，另外，$P(\theta \mid x)$ 称为后验分布。
- 当这个后验分布和先验分布是同一个分布时，我们称先验分布和似然函数为共轭分布，也就是我们先验分布假设的比较准确。

在介绍共轭分布前我们先介绍一下 Gamma 函数和 Beta 函数。

24.4.1 Gamma 函数

Gamma 函数 $\Gamma(x)$ 定义为

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

通过分部积分法，可以很容易证明 Gamma 函数具有如下之递归性质

$$\Gamma(x+1) = x\Gamma(x)$$

也是便很容易发现，它还可以看做是阶乘在实数集上的延拓，即

$$\Gamma(x) = (x-1)!$$

24.4.1 Beta 函数

定义 Beta 函数如下

$$\mathbf{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Beta 函数的另外一种定义形式为 (注意这两种定义是等价的)

$$\mathbf{B}(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$$

24.4.1 Beta 分布

Beta 分布的概率密度函数 (PDF) 定义为:

$$Beta(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

或

$$Beta(\theta|a,b) = \frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

Beta 分布的均值和方差分别有下面两式给出

$$E[\theta] = \frac{a}{a+b}$$

$$var[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$$

$Beta(\theta|a,b) = \frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$ 可见，Beta 分布有两个控制参数 a 和 b，而且当这两个参数取不同值时，Beta 分布的 PDF 图形可能会呈现出相当大的差异。如下图??所示。

24.4.1 Beta 分布

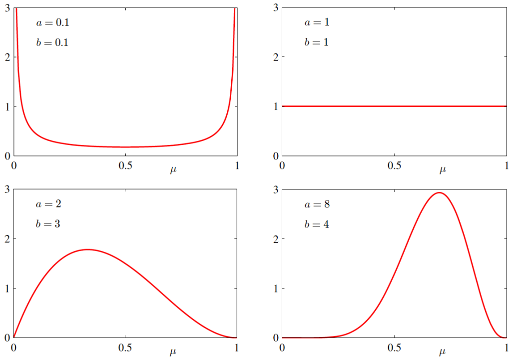


图 1: 参数 a 和 b 变化时 Beta 分布的 PDF 图

24.4.1 共轭分布

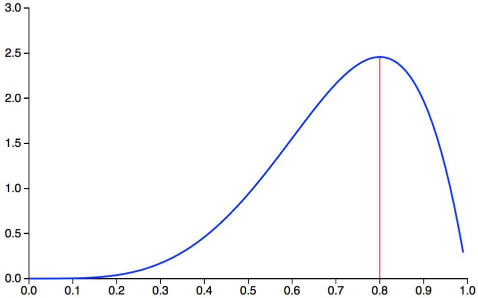
- 假如你有一个硬币，它有可能是不均匀的，所以投这个硬币有 θ 的概率抛出正面，有 $(1-\theta)$ 的概率抛出背面。如果抛了五次这个硬币，有三次是正面，有两次是背面，完全根据目前观测的结果来估计 θ，那么显然你会得出结论 $\theta = \frac{3}{5}$。
- 点估计的方法有漏洞。实验次数较少，估计结果可能有较大偏差。
- 如果抛了五次都是正面，以后永远都抛出正面么?

24.4.1 共轭分布

- 在贝叶斯学派看来，参数 θ 不是一个固定的值，而满足一定的概率分布。
- 在估计 θ 时，我们心中可能有一个根据经验的估计，即先验概率，$P(\theta)$。而给定一系列实验观察结果 X 的条件下，我们可以得到后验概率为 $P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$
- 在上面的贝叶斯公式中，$P(\theta)$ 就是个概率分布。

24.4.1 共轭分布

- 这个概率分布可以是任何概率分布，比如高斯分布，或者刚刚提过的 Beta 分布。
- 下图是 Beta(5,2) 的概率分布图。如果我们将这个概率分布作为 $P(\theta)$，那么我们在还未抛硬币前，便认为 θ 很可能接近于 0.8，



24.4.1 共轭分布

使用 Beta 分布的原因：

- 虽然 $P(\theta)$ 可以是任何种类的概率分布，但是如果使用 Beta 分布，会让之后的计算更加方便。(稍后说明)
- 通过调节 Beta 分布中的 a 和 b，你可以让这个概率分布变成各种你想要的形状！

24.4.1 共轭分布

- $P(X|\theta)$ 是个二项 (Binomial) 分布。
- 继续以前面抛 5 次硬币抛出 3 次正面的观察结果为例，X= 抛 5 次硬币 3 次结果为正面的事件，则 $P(X|\theta) = C_5^2\theta^3(1-\theta)^2$。
- 而如果我们采用 Beta 分布，θ 的概率分布在 [0,1] 之间是连续的，用积分，即 $P(X) = \int_0^1 P(X|\theta)P(\theta)\,d\theta$

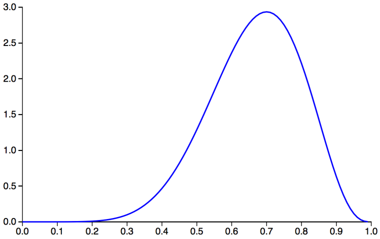
24.4.1 共轭分布

$P(\theta)$ 是个 Beta 分布，那么在观测到 “X= 抛 5 次硬币中出现 3 个正面” 的事件后，$P(\theta|X)$ 依旧是个 Beta 分布！这是使用 Beta 分布方便计算的原因。

$$
\begin{aligned}
P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int_0^1 P(X|\theta)P(\theta)\,d\theta} \\
&= \frac{C_5^2\theta^3(1-\theta)^2\frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 C_5^2\theta^3(1-\theta)^2\frac{1}{\mathbf{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}\,d\theta} \\
&= \frac{\theta^{(a+3-1)}(1-\theta)^{(b+2-1)}}{\int_0^1 \theta^{(a+3-1)}(1-\theta)^{(b+2-1)}\,d\theta} \\
&= \frac{\theta^{(a+3-1)}(1-\theta)^{(b+2-1)}}{\mathbf{B}(a+3,b+2)} \\
&= Beta(\theta|a+3,b+2)
\end{aligned}
$$

24.4.1 共轭分布

- 当我们得知 $P(\theta|X) = Beta(\theta|a+3, b+2)$ 后，我们就只要根据 Beta 分布的特性，得出 θ 最有可能等于多少了。(即 θ 等于多少时，观测后得到的 Beta 分布有最大的概率密度)。
- 例如下图，仔细观察新得到的 Beta 分布，和上一图中的概率分布对比，发现峰值从 0.8 左右的位置移向了 0.7 左右的位置。Bayesian 方法和普通的统计方法不同的地方：我们结合自己的先验概率和观测结果来给出预测。



24.4.1 共轭分布

共轭性

- 共轭性：后验概率分布 (正比于先验和似然函数的乘积) 拥有与先验分布相同的函数形式。这个性质被叫做共轭性 (Conjugacy)。

- 共轭先验使得后验概率分布的函数形式与先验概率相同，因此使得贝叶斯分析得到了极大的简化。例如，二项分布的参数之共轭先验就是我们前面介绍的 Beta 分布。多项式分布的参数之共轭先验则是 Dirichlet 分布，高斯分布的均值之共轭先验是另一个高斯分布。