第十三章 贝叶斯统计推断

§1 贝叶斯统计与经典统计

关于模型或参数的不同认识

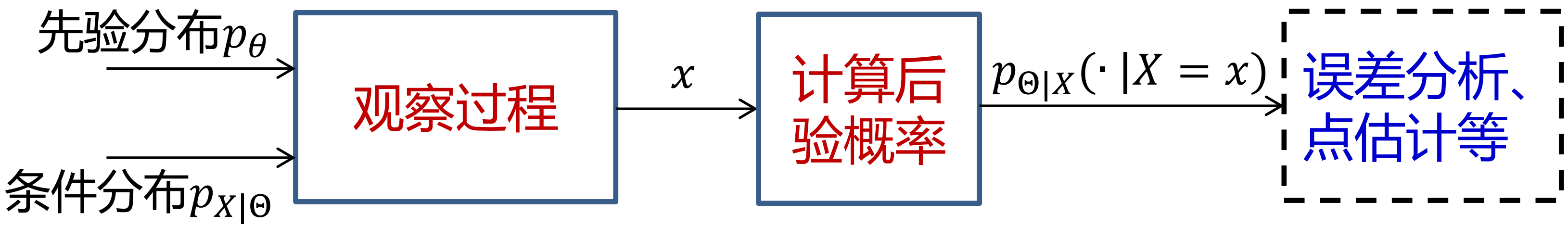
- 经典学派(频率学派)：未知待估计的量(常数)
- 贝叶斯学派：已知分布的随机变量
  - 引入随机变量Θ刻画模型
  - 构造先验分布$p_\Theta(\theta)$
  - 利用数据x能提供的关于θ的所有信息，修正对θ的估计$p_{\Theta|X}(\theta|x)$

贝叶斯推断的主要方法

1. 最大后验(Maximum-a-Posteriori，MAP)估计
   在可能的参数/假设的取值范围内，选择一个在给定数据下，具有最大化条件概率/后验概率的值
2. 最小均方(Least Mean Square, LMS)估计
   选择数据的一个估计量或函数，使得参数与估计之间的均方误差最小
3. 线性最小均方(Linear Minimum Mean Square, LMMS)估计
   选择数据的一个线性函数，使得参数与估计之间的均方误差达到最小

§2 贝叶斯推断与后验分布

贝叶斯推断的目标：



贝叶斯推断：

- 起点是未知随机变量Θ的先验分布$p_\Theta(\theta)$或f_Θ
- 得到观测向量X的$p_{X|\Theta}$或$f_{X|\Theta}$
- 一旦X的一个特定值x观测到后，运用贝叶斯规则计算Θ的后验分布$p_{\Theta|X}(\theta|x)$或$f_{\Theta|X}(\theta|x)$

贝叶斯规则的四种形式：

- Θ离散，X离散

$$p_{\Theta|X}(\theta|x)=\frac{p_{\Theta}(\theta)p_{X|\Theta}(x|\theta)}{\sum_{\theta'}p_{\Theta}(\theta')p_{X|\Theta}(x|\theta')}$$

- Θ离散，X连续

$$p_{\Theta|X}(\theta|x)=\frac{p_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{\sum_{\theta'}p_{\Theta}(\theta')f_{X|\Theta}(x|\theta')}$$

- Θ连续，X离散

$$f_{\Theta|X}(\theta|x)=\frac{f_{\Theta}(\theta)p_{X|\Theta}(x|\theta)}{\int f_{\Theta}(\theta')p_{X|\Theta}(x|\theta')d\theta'}$$

- Θ连续，X连续

$$f_{\Theta|X}(\theta|x)=\frac{f_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{\int f_{\Theta}(\theta')f_{X|\Theta}(x|\theta')d\theta'}$$

例1：A和B约会，假设B在任何约会中都会迟到，迟到时间记为随机变量X，服从$[0,\theta]$上的均匀分布，参数θ未知，是随机变量Θ的一个值， Θ在[0,1]之间均匀分布，假设B在第一次约会中迟到了x，那么A如何利用这个信息更新Θ的分布。

解：先验概率密度：

$$f_\Theta(\theta)=\begin{cases}1\text{，若}0\le\theta\le1\\0\text{，}\quad\text{其他}\end{cases}$$

观测值的条件密度函数：

$$f_{X|\Theta}(x|\theta)=\begin{cases}\dfrac{1}{\theta}\text{，若}0\le x\le\theta\\0\text{，}\quad\text{其他}\end{cases}$$

注意到：

$$f_\Theta(\theta)f_{X|\Theta}(x|\theta)$$

只在$0 \le x \le \theta \le 1$情况下为非零。

由贝叶斯规则有：

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\int_x^1 f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'} = \frac{1/\theta}{\int_x^1 \frac{1}{\theta'}d\theta'}$$

$$= \frac{1}{\theta|\ln x\ |}$$

当$\theta \le x$或$\theta > 1$时$f_{\Theta|X}(\theta|x) = 0$

现在考虑前n次约会所引起的变化，假设B迟到的时间记为$X_1, X_2, \ldots, X_n$，在给定$\Theta = \theta$的条件下，它是区间$[0, \theta]$的均匀分布，且条件独立，记$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$，$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$

类似n=1，有：

$$f_\Theta(\theta) = \begin{cases} 1\text{，若}0 \le \theta \le 1 \\ 0\text{，}\quad\quad \text{其他} \end{cases}$$

观测值的条件密度函数：

$$f_{X|\Theta}(x|\theta) = \begin{cases} \dfrac{1}{\theta^n}\text{，若}\bar{x} \le \theta \le 1 \\ 0\text{，}\quad\quad \text{其他} \end{cases}, \bar{x} = \max\{x_1, \ldots, x_n\}$$

若$\bar{x} \le \theta \le 1$

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\int_{\bar{x}}^1 f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'} = \frac{1/\theta^n}{\int_{\bar{x}}^1 \frac{1}{(\theta')^n} d\theta'}$$

于是

$$f_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{1/\theta^n}{\int_{\bar{x}}^1 \frac{1}{(\theta')^n} d\theta'}, \text{若} \bar{x} \le \theta \le 1 \\ 0, \quad \text{其他} \end{cases},$$

$$\bar{x} = \max\{x_1, \dots, x_n\}$$

例2. (正态随机变量公共均值的推断)设随机变量观测值$X=(X_1\ ,\ \dots\ ,X_n)$具有相同的均值但未知，需要估计。假设给定均值的条件下，X_i是正态的，且相互独立，方差分别为 $\sigma_1^2,\dots,\sigma_n^2$ 。使用贝叶斯方法，对均值进行建模。

解：设X的公共均值为随机变量Θ且已知其先验分布，假设随机变量Θ的分布为正态分布，均值已知为 x_0 ，方差为σ_0^2。

为将来引用，将上述模型等价于：

$$X_i=\Theta+W_i, i=1,\dots,n$$

其中随机变量$\Theta, W_1,\dots W_n$相互独立，且是正态分布的，均值和方差均已知

特别地，对任意的θ,

$$E[W_i] = E[W_i|\Theta = \theta] = 0,$$

$$D[W_i] = D(X\,|\Theta = \theta) = \sigma_i^2$$

这类模型在许多工程应用中非常普遍，工程中一个未知量往往有若干个独立的测量。

根据假设，有：

$$f_\Theta(\theta) = c_1 \exp\left\{-\frac{(\theta - x_0)^2}{2\sigma_0^2}\right\}$$

$$f_{X|\Theta}(x|\theta) = c_2 \exp\left\{-\sum_{i=1}^{n}\frac{(x_i - \theta)^2}{2\sigma_i^2}\right\}$$

这里c_1, c_2是归一化常数，不依赖于θ

由贝叶斯法则：

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'}$$

$$= \frac{c_1 c_2 \exp\left\{-\sum_{i=0}^{n} \frac{(x_i - \theta)^2}{2\sigma_i^2}\right\}}{\int f_\Theta(\theta') f_{X|\Theta}(x|\theta') d\theta'}$$

注意到分母不依赖于θ

$$f_{\Theta|X}(\theta|x)=a\exp\left\{-\sum_{i=0}^{n}\frac{(x_i-\theta)^2}{2\sigma_i^2}\right\}=a\exp\left\{-\sum_{i=0}^{n}\frac{(\theta-m)^2}{2v}\right\}$$

其中：

$$m=\frac{\sum_{i=0}^{n}x_i/\sigma_i^2}{\sum_{i=0}^{n}1/\sigma_i^2},v=\frac{1}{\sum_{i=0}^{n}1/\sigma_i^2}$$

$a=1/\sqrt{2\pi v}$是规一化常数，只依赖于x_i不依赖于θ

当$\sigma_0^2=\sigma_1^2=,\ldots,=\sigma_n^2=\sigma^2$时,

$$m=\frac{\sum_{i=0}^{n}x_i}{n+1},v=\frac{\sigma^2}{n+1}$$

在这种情况下，先验均值扮演着一个观测值的作用，而且对后验均值发挥相同的作用

讨论：

- 注意到后验密度的标准差在观测样本量增大时，趋于0，速度大致是$1/\sqrt{n}$，如果方差对不相同，后验均值仍是每个x_i的加权平均，方差越小，对m的权重就越大
- 后验分布与先验分布是同一个分布族，都是正态分布，二项分布类似
  1. 后验分布的特征只有两个数——均值和方差
  2. 后验分布的解形式可以使用有效的递归推断，假设已经获得观测值$X_1, \ldots, X_n$，且下一个观测值X_{n+1}也得到了，那么可以将$f_{\theta|x_1,\ldots,x_n}$作为先验，然后运用新观测值得到新后验$f_{\theta|x_1,\ldots,x_n,x_{n+1}}$。

例3 (垃圾邮件过滤)一封电子邮件不是垃圾邮件就是正常邮件，我们引入参数Θ，取值为1和2，分别代表垃圾和正常，各自取值的概率分别为$p_\Theta(1)$，$p_\Theta(2)$。设$\{w_1, w_2, \dots, w_m\}$代表一些特殊的词(或者词的组合)形成的集合，它们出现后就表示邮件是垃圾的，对每个i，记X_i是伯努利随机变量，来定义w_i是否出现在信息中，即当w_i 出现时， $X_i = 1$否则$X_i = 0$ ，假设条件概率$p_{X_i|\Theta}(x_i|1)$ 和$p_{X_i|\Theta}(x_i|2)$, $x_i = 0,1$是已知的。简单起见，假设在给定Θ的条件下，随机变量$X_1, \dots, X_n$是相互独立的，计算垃圾邮件和正常邮件的后验概率。

解：

$$P(\Theta = m | X_1 = x_1, \ldots, X_n = x_n) = \frac{p_\Theta(m) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|m)}{\sum_{j=1}^{2} p_\Theta(j) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|j)}, \qquad m = 1,2$$

这两个后验概率可以用于将邮件分类为垃圾还是正常，其计算方法将在后面继续讨论

多参数问题——多个未知参数的估计

例4 (传感器网络的定位) 假设有n个声敏元件，分布在我们关注的一个地理区域内，设第i个声敏元件的坐标是(a_i, b_i)。一辆发送已知声音信号的车辆在这个区域内，坐标为$\Theta = (\Theta_1, \Theta_2)$，但是未知，每个声敏元件探测这个车辆(即捕捉到这个车辆的信号)的概率依赖于它们之间的距离，观测数据是哪些声敏元件探测到车辆，哪些没有探测到，目标就是尽可能地找到车辆所在的位置

先验密度f_Θ是基于历史观测数据对这个车辆的位置的大致认识。

简单起见，假设Θ_1和Θ_2是相互独立的正态随机变量，均值为0、方差为1。于是：

$$f_\Theta(\theta_1,\theta_2)=\frac{1}{2\pi}e^{-(\theta_1^2+\theta_2^2)/2}$$

假定探测到信号的概率和声敏元件与车辆之间的距离呈指数递减，则探测到信号的的条件概率为，

$$P(X_i=1|\Theta=(\theta_1,\theta_2))=p_{X|\Theta}(1|\theta_1,\theta_2)=e^{-d_i(\theta_1,\theta_2)}$$

其中：

$$d_i^2(\theta_1,\theta_2)=(a_i-\theta_1)^2+(b_i-\theta_2)^2$$

定义S为$X_i = 1$的传感器集合，且彼此独立的，计算后验密度$f_{\Theta|X}(\boldsymbol{\theta}|\mathbf{x})$的分子项：

$$f_\Theta(\boldsymbol{\theta}) p_{X|\Theta}(1|\boldsymbol{\theta}) = \frac{1}{2\pi} e^{-(\theta_1^2+\theta_2^2)/2} \prod_{i\in S} e^{-d_i(\theta_1,\theta_2)} \prod_{i\notin S} \left(1 - e^{-d_i(\theta_1,\theta_2)}\right)$$

§3 最大后验准则、点估计与假设检验

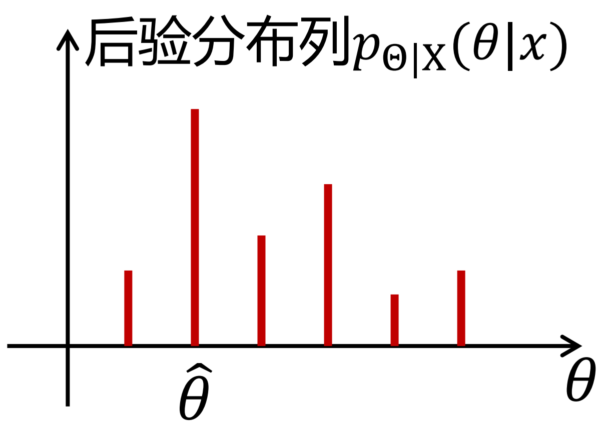
最大后验准则准则：

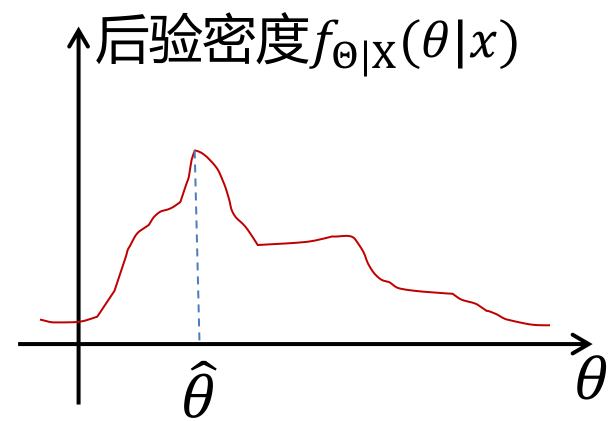
给定观测值x，选择θ的一个取值，记为$\hat{\theta}$，使得后验分布列$p_{\Theta|\mathrm{X}}(\theta|x)$ 达到最大(若Θ连续则为后验分布密度$f_{\Theta|\mathrm{X}}(\theta|x)$)：

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta|\mathrm{X}}(\theta|x), (\Theta\text{为离散})$$

或

$$\hat{\theta} = \arg\max_{\theta} f_{\Theta|\mathrm{X}}(\theta|x), (\Theta\text{为连续})$$





最大后验概率准则：

- 由最大后验概率准则获得的$\hat{\theta}$是Θ最有可能的取值，使对任意给定的x有最大的概率做出正确的决定
- 最大后验概率准则使总体(平均了所有x可能的取值)做出正确决定的概率达到最大(在所有决策准则中)
- 等价地，最大后验概率准则使得做出错误决定的概率达到最小

在贝叶斯准则下由于所有可能估计的分母都是一样的，因此只需要比较分子部分

最大后验概率准则：

- 给定x的观测值，最大后验准则是指在所有的θ中寻找$\hat{\theta}$使得后验分布$p_{\Theta|X}(\theta|x)$ (Θ离散)或$f_{\Theta|X}(\theta|x)$ (Θ连续)达到最大值
- 等价地，最大后验准则是在所有的θ中寻找$\hat{\theta}$使得下面函数达到最大

$$p_{\Theta}(\theta)p_{X|\Theta}(x|\theta), (\Theta\text{和}X\text{均离散})$$

$$p_{\Theta}(\theta)f_{X|\Theta}(x|\theta), (\Theta\text{离散，}X\text{连续})$$

$$f_{\Theta}(\theta)p_{X|\Theta}(x|\theta), (\Theta\text{连续，}X\text{离散})$$

$$f_{\Theta}(\theta)f_{X|\Theta}(x|\theta), (\Theta\text{和}X\text{均连续})$$

例3续（垃圾邮件分类）

$$P(\Theta = m | X_1 = x_1, \ldots, X_n = x_n)$$
$$= \frac{p_\Theta(m) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|m)}{\sum_{j=1}^{2} p_\Theta(j) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|j)}, \qquad m = 1,2$$

参数Θ取值为1和2，分别代表垃圾邮件和正常邮件，各自取值的概率分别为$p_\Theta(1)$，$p_\Theta(2)$，X_i是伯努利随机变量，用于定义词汇w_i是否出现在信息中，即当w_i出现时，$X_i = 1$，否则$X_i = 0$

如果下式成立，则判断一封邮件是垃圾

$$P(\Theta = 1 | X_1 = x_1, \ldots, X_n = x_n) > P(\Theta = 2 | X_1 = x_1, \ldots, X_n = x_n)$$

或等价地：

$$p_\Theta(1) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|1) > p_\Theta(2) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|2)$$

最大后验估计与最大似然估计的区别

最大似然估计：

$$\hat{\theta}_{MLE} = \arg\max_{\theta} f_{X|\Theta}(x|\theta)$$

最大后验估计：

$$\hat{\theta} = \arg\max_{\theta} f_{\Theta|\mathrm{X}}(\theta|x) = \arg\max_{\theta} f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)$$

MAP与MLE区别——是MAP中加入了先验模型，或者说。MLE中认为模型参数本身的概率的是均匀的，即该概率为一个固定值

点估计

在一个估计问题中，给定X的观测值x，后验分布抓住了x提供的所有相关信息，而另一方面，我们对概括了后验性质的某些量很感兴趣、比如，点估计是一个数值、它表达了我们关于Θ取值的最好猜测。

最大后验概率估计量：观测到x在所有的θ中选$\hat{\theta}$使得后验分布达到最大，当有很多这样的取值时，可在备选量中任意选定

条件期望估计量，这里选定的估计量为$\hat{\theta} = E[\Theta|X = x]$。也称最小均方估计

回到前面例1(A和B约会问题)，假设B迟到时间记为随机变量$X \sim U[0,\theta]$，参数θ是随机变量Θ的一个值，$\Theta \sim U[0,1]$， 假设B在第一次约会中迟到了x，那么A对Θ的估计如下。

最大后验估计

$$f_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{1}{\theta|\ln x|}, & x \le \theta \le 1 \\ 0, & 其他 \end{cases}$$

- 对于给定的x，最大后验估计$f_{\Theta|X}(\theta|x) \propto \frac{1}{\theta}$ (Θ取值范围内)➔最大后验概率估计得到对θ的估计就是x
- 这是一个很乐观的估计，若B第一次约会只迟到一小会儿，则未来约会迟到时间的估计是很小的

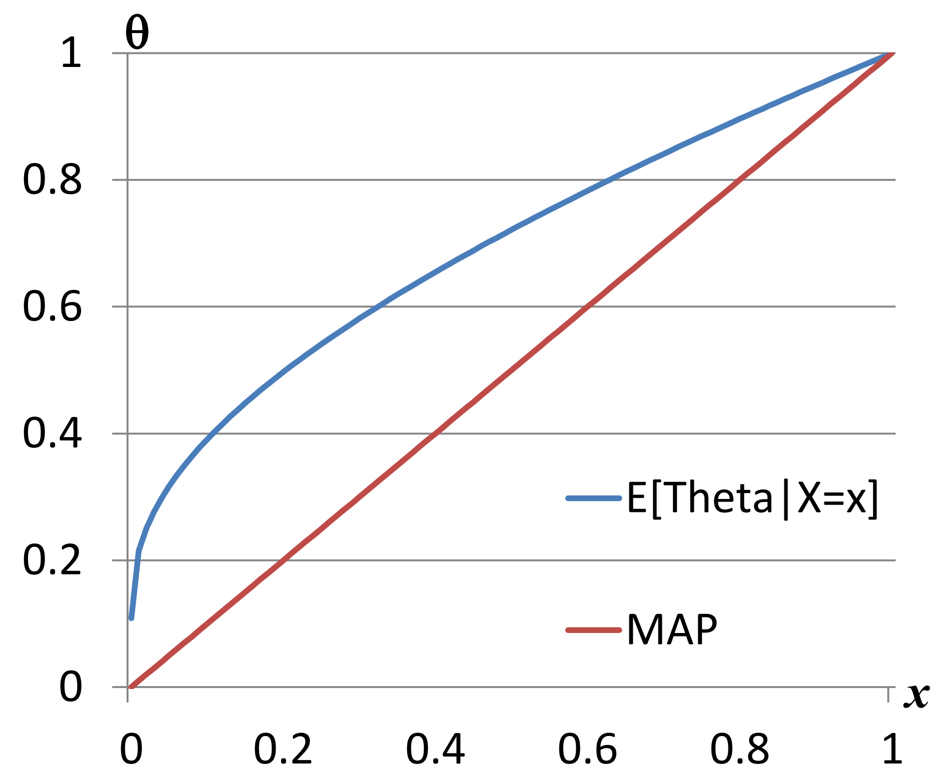
条件期望估计

$$E[\Theta|X=x]=\int_x^1 \theta f_{\Theta|X}(\theta|x)d\theta=\int_x^1 \theta\frac{1}{\theta|\ln x|}d\theta=\frac{1-x}{|\ln x|}$$

- 条件期望估计就则没有这么乐观

注意：如果没有附加的假设条件，点估计的准确性是没有多大保障的

最大后验概率估计可能和后验分布的主体部分相距远



假设检验

假设检验的最大后验概率准则

- 给定观测值x，最大后验概率准则选择使后验概率$P(\Theta = \theta_i | X = x)$最大的假设$H_i$。
- 等价地，也就是使$p_\theta(\theta_i) p_{X|\Theta}(x|\theta)$(X离散)或$p_\theta(\theta_i) f_{X|\Theta}(x|\theta)$(X连续)达到最大的假设$H_i$。
- 与其他决策准则相比，最大后验概率准则对任意观测值x使得选择错误假设的概率，也即犯错的概率达到最小

例5 有两枚不均匀的硬币，记为硬币1和硬币2，正面朝上的概率分别为p_1和p_2。随机选择一枚硬币(每枚有相同的入选概率)，希望在一次抛硬币结果的基础上判断这枚硬币是硬币1还是硬币2。令$\Theta = 1$和$\Theta = 2$分别代表假设”选择硬币1”和”选择硬币2”。记X=1表示硬币正面朝上，X=0表示反面朝上

利用最大后验概率准则，比较$p_\Theta(1)p_{X|\Theta}(x|1)$和$p_\Theta(2)p_{X|\Theta}(x|2)$的大小，并且认为所投硬币就是表达式取值相应较大的那个

由于$p_\Theta(1)=p_\Theta(2)=1/2$，只须比较$p_{X|\Theta}(x|1)$和$p_{X|\Theta}(x|2)$，比如若$p_1=0.46$，$p_2=0.52$，投掷结果是反面，注意到

$$P(\text{反面}|\theta=1)=1-0.46>1-0.52$$
$$=p(\text{反面}|\theta=2)$$

因而认为所抛掷的是硬币1

假设现在将所选的硬币投掷了n次，X是正面朝上的次数，根据最大后验概率准则选择观测结果最有可能发生的假设(建立在假设$p_\Theta(1)=p_\Theta(2)=1/2$的基础上)，因而当X=k时，若

$$p_1^k(1-p_1)^{n-k}>p_2^k(1-p_2)^{n-k}$$

则认为$\Theta=1$，否则，认为$\Theta=2$

Bayes判别：

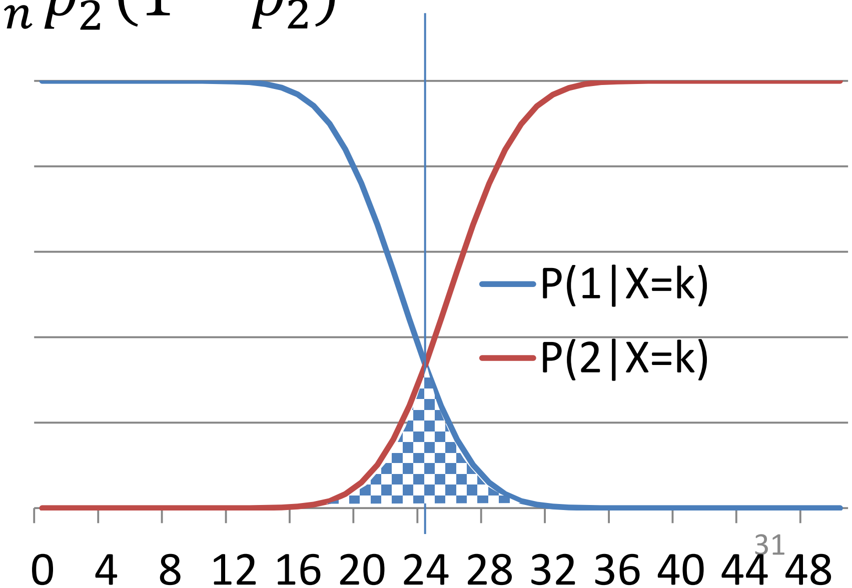
最大后验概率准则可用于典型的两重假设检验问题

在抛硬币的问题中，最大后验概率准则通过设置的划分点分类

$$P(\text{错误}) = P(\Theta = 1; X > k^*) + P(\Theta = 2; X \le k^*)$$
$$= p_\Theta(1) \sum_{k=k^*+1}^{n} C_n^k p_1^k (1-p_1)^{n-k}$$
$$+ p_\Theta(2) \sum_{k=1}^{k^*} C_n^k p_2^k (1-p_2)^{n-k}$$

取$n = 50,\ p_\Theta(1) = p_\Theta(2) = 0.5,$

$p_1 = 0.46,\ p_2 = 0.52$



§4 贝叶斯最小均方估计

重点：条件期望估计量

目的：使可能的均方误差达到最小

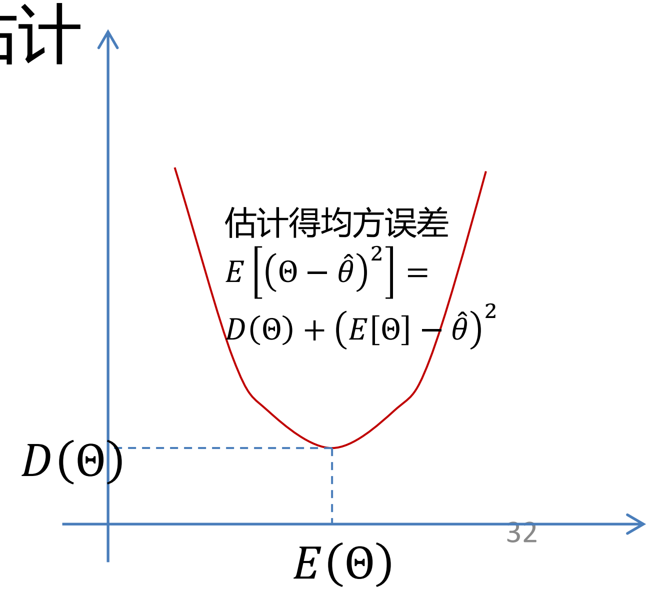
考虑在没有观测值x的情况下用常数$\hat{\theta}$来估计Θ，估计误差$\hat{\theta}-\Theta$是随机的(因为Θ是随机的)，但均方误差

$$E\left[\left(\Theta-\hat{\theta}\right)^2\right]=D\left(\Theta-\hat{\theta}\right)+\left(E\left[\left(\Theta-\hat{\theta}\right)\right]\right)^2$$
$$=D(\Theta)+\left(E[\Theta]-\hat{\theta}\right)^2$$

因此$\hat{\theta}=E[\Theta]$是使得$E\left[\left(\Theta-\hat{\theta}\right)^2\right]$最小的估计

假设现在我们由观测值x来估计Θ，
同时要求均方误差最小

条件期望$E[\Theta|X=x]$在所有常数$\hat{\theta}$中
使得$E\left[\left(\Theta-\hat{\theta}\right)^2|X=x\right]$达到最小



广义上，估计量为$g(X)$的(非条件)均方估计误差定义为

$$E\left[\left(\Theta - g(X)\right)^2\right]$$

如果我们将$E[\Theta|X]$视为X的函数或估计量，则$g(X) = E[\Theta|X]$使得均方误差最小。

关于最小均方估计的重要结论

- 在没有观测值的情况下，当$\hat{\theta} = E[\Theta]$时$E\left[\left(\Theta - \hat{\theta}\right)^2\right]$达到最小，即：
$$E[(\Theta - \mathrm{E}[\Theta])^2] \le E\left[\left(\Theta - \hat{\theta}\right)^2\right],\ \text{对所有}\hat{\theta}\text{成立}$$
- 给定X的取值x，当$\hat{\theta} = E[\Theta|X = x]$时$E\left[\left(\Theta - \hat{\theta}\right)^2 | X = x\right]$达到最小，即
$$E[(\Theta - \mathrm{E}[\Theta|X = x])^2 | X = x] \le E\left[\left(\Theta - \hat{\theta}\right)^2 | X = x\right],\ \text{对所有}\hat{\theta}\text{成立}$$
- 在所有的基于X的Θ的估计量$g(x)$中，当 $g(x) = E[\Theta|x]$时均方估计误差$E\left[\left(\Theta - g(X)\right)^2\right]$达到最小，即
$$E[(\Theta - E[\Theta|X])^2] \le E\left[\left(\Theta - g(X)\right)^2\right],\ \text{对所有}g(\mathrm{X})\text{成立}$$

例6 考虑例1中B第一次约会中迟到时间 $X \sim U[0,\theta]$. 参数θ是随机变量Θ的一个值，$\Theta \sim U[0,1]$，假设B在第一次约会中迟到了x，那么A对Θ的估计如下。

θ的最大后验估计就是x，而最小均方估计

$$E[\Theta|X=x]=\int_x^1 \theta f_{\Theta|X}(\theta|x)d\theta=\frac{1-x}{|\ln x|}$$

考察最大后验概率估计和最小均方估计的条件均方误差

给定$X=x$，对于任意的$\hat{\theta}$有

$$E\left[\left(\hat{\theta}-\Theta\right)^2\middle|X=x\right]=\int_x^1\left(\hat{\theta}-\theta\right)^2 f_{\Theta|X}(\theta|x)d\theta$$

$$=\int_x^1\frac{\left(\hat{\theta}-\theta\right)^2}{\theta|\ln x|}d\theta=\hat{\theta}^2-\hat{\theta}\frac{2(1-x)}{|\ln x|}+\frac{1-x^2}{2|\ln x|}$$

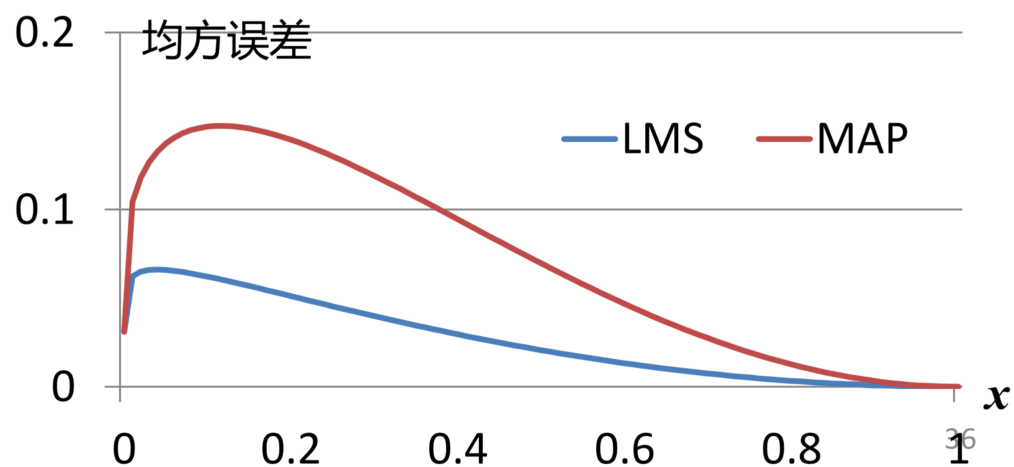
对最大后验概率估计$\hat{\theta}=x$而言，条件均方误差：

$$E\left[\left(\hat{\theta}-\Theta\right)^2 \middle| X=x\right]=x^2+\frac{3x^2-4x+1}{2|\ln x|}$$

对最小均方估计$\hat{\theta}=\frac{1-x}{|\ln x|}$而言，条件均方误差：

$$E\left[\left(\hat{\theta}-\Theta\right)^2 \middle| X=x\right]=\frac{1-x^2}{2|\ln x|}-\frac{(1-x)^2}{(\ln x)^2}$$

最小均方估计有一致的相对较小的均方误差，这是最小均方估计量的总体优良性能的体现



估计误差的一些性质：

$$\hat{\Theta} = E[\Theta|X], \tilde{\Theta} = \hat{\Theta} - \Theta$$

- 估计误差$\tilde{\Theta}$是无偏的，具体说来它的条件期望和非条件期望都是0；
  $E[\tilde{\Theta}] = 0, \qquad E[\tilde{\Theta}|X = x] = 0, \qquad$ 对所有的x
- 估计误差$\tilde{\Theta}$和估计量$\hat{\Theta}$是不相关的：
  $$\mathrm{Cov}(\tilde{\Theta}, \hat{\Theta}) = 0$$
- Θ的方差可以分解为
  $$D(\Theta) = D(\tilde{\Theta}) + D(\hat{\Theta})$$

多元随机变量情形

前面的讨论同样适用于$X=(X_1\text{，}\ldots\text{，}X_n)$是随机向量的情形

均方估计误差在选 $E[\Theta|X_1,\ldots X_n]$ 作为估计量时达到最小，即:

$$E[(\Theta-\mathrm{E}[\Theta|X_1,\ldots X_n])^2]\le E\left[\left(\Theta-\mathrm{g}(X_1,\ldots X_n)\right)^2\right]$$

对于所有的估计量$\mathrm{g}(X_1,\ldots X_n)$都成立

实现上的困难：

- 为计算条件期望$\mathrm{E}[\Theta|X_1,\ldots X_n]$需要建立有联合分布密度函 $f_{\Theta,X_1,\ldots,X_n}$
- 该联合分布函数常常是非常复杂的

实际中常常求助于条件期望的近似，或关注那些并不最优但是简单而易于实现的估计量，如加入了线性估计的约束

多参数估计

自然的准则：

$$\mathrm{E}\left[\left(\Theta_1-\widehat{\Theta}_1\right)^2\right]+\cdots+\mathrm{E}\left[\left(\Theta_m-\widehat{\Theta}_m\right)^2\right]$$

目的是使得上式在一切估计量中达到最小，这与寻找每个$\widehat{\Theta}_i$使得$\mathrm{E}\left[\left(\Theta_i-\widehat{\Theta}_i\right)^2\right]$达到最小是等价的

因此，多参数的估计问题本质上是在处理m个单参数的估计问题

§5 贝叶斯线性最小均方估计

基于观测$X_1,\ldots,X_n$的Θ的线性估计量形式为：

$$\hat{\Theta} = a_0 + a_1X_1 + \cdots + a_nX_n$$

给定$\{a_i\}, i = 0,1,\ldots,n$，相应估计的均方误差是

$$E\left[\left(\Theta - (a_0 + a_1X_1 + \cdots + a_nX_n)\right)^2\right]$$

线性最小均方估计选择$\{a_i\}, i = 0,1,\ldots,n$使得上面的表达式取最小值

考虑n=1，即$\hat{\Theta} = a_0 + a_1X$的情况，使得均方误差

$$E[(\Theta - a_0 - a_1X)^2]$$

达到最小

假设已经确定了a_1，则a_0的选择等价于选择常数a_0来估计随机变量$\Theta - a_1 X$

按照上一节的方法，最好的估计就是

$$a_0 = E[\Theta - a_1 X] = E[\Theta] - a_1 E[X]$$

确定了a_0之后，代回去确定a_1使得下面的表达式取最小值

$$E[(\Theta - a_1 X - E[\Theta] + a_1 E[X])^2]$$

写成等价的方差形式：

$$D[(\Theta - a_1 X)] = \sigma_\Theta^2 + a_1^2 \sigma_X^2 + 2Cov(\Theta, -a_1 X)$$
$$= \sigma_\Theta^2 + a_1^2 \sigma_X^2 - 2a_1 Cov(\Theta, X)$$

于是a_1满足

$$\frac{d}{da_1} D[(\Theta - a_1 X)] = 0$$

$$a_1 = \frac{Cov(\Theta, X)}{\sigma_X^2} = \frac{\rho\sigma_\Theta\sigma_X}{\sigma_X^2} = \rho\frac{\sigma_\Theta}{\sigma_X}$$

其中相关系数

$$\rho = \frac{Cov(\Theta, X)}{\sigma_\Theta\sigma_X}$$

于是相应的均方估计为

$$\hat{\Theta} = E[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

均方估计误差为

$$D\left[\left(\Theta - \hat{\Theta}\right)\right] = \sigma_\Theta^2 + a_1^2\sigma_X^2 - 2a_1 Cov(\Theta, X)$$

$$= \sigma_\Theta^2 + \rho^2\frac{\sigma_\Theta^2}{\sigma_X^2}\sigma_X^2 - 2\rho\frac{\sigma_\Theta}{\sigma_X}\rho\sigma_\Theta\sigma_X = (1 - \rho^2)\sigma_\Theta^2$$

注意：

- 线性最小均方估计的公式只包括均值、方差以及Θ与X 间的协方差
- 直观的解释，为描述准确起见，假设相关系数ρ是正的，估计量以Θ的基本估计$\mathrm{E}[\Theta]$为基础，通过$X-E[X]$的取值来调整

例7. 继续考虑约会迟到问题

B第一次约会中迟到时间 $X\sim U[0,\theta]$. 参数θ是随机变量Θ的一个值， $\Theta\sim U[0,1]$， 假设B在第一次约会中迟到了x。

下面求基于X的Θ的线性最小均方估计

- 利用事实$E[X|\Theta]=\Theta/2$和重期望法则，X的期望值

$$\mathrm{E}[X]=\mathrm{E}\big[E[X|\Theta]\big]=\mathrm{E}\left[\frac{\Theta}{2}\right]=\frac{1}{4}$$

$$D(X)=\int_0^1\int_0^\theta\frac{1}{\theta}(x-E[X])^2dx\,d\theta=\frac{7}{144}$$

$$\mathrm{Cov}(\Theta, X) = \mathrm{E}[\Theta X] - \mathrm{E}[\Theta]\mathrm{E}[\mathrm{X}]$$

由于：

$$\mathrm{E}[\Theta X] = E\big[E[\Theta X|\Theta]\big] = E\big[\Theta E[X|\Theta]\big] = E\left[\frac{\Theta^2}{2}\right] = \frac{1}{6}$$

于是：

$$\mathrm{Cov}(\Theta, X) = \frac{1}{6} - \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{24}$$

线性最小均方估计量为：

$$\widehat{\Theta} = E[\Theta] + \frac{Cov(\Theta, X)}{D(X)}(X - E[X])$$

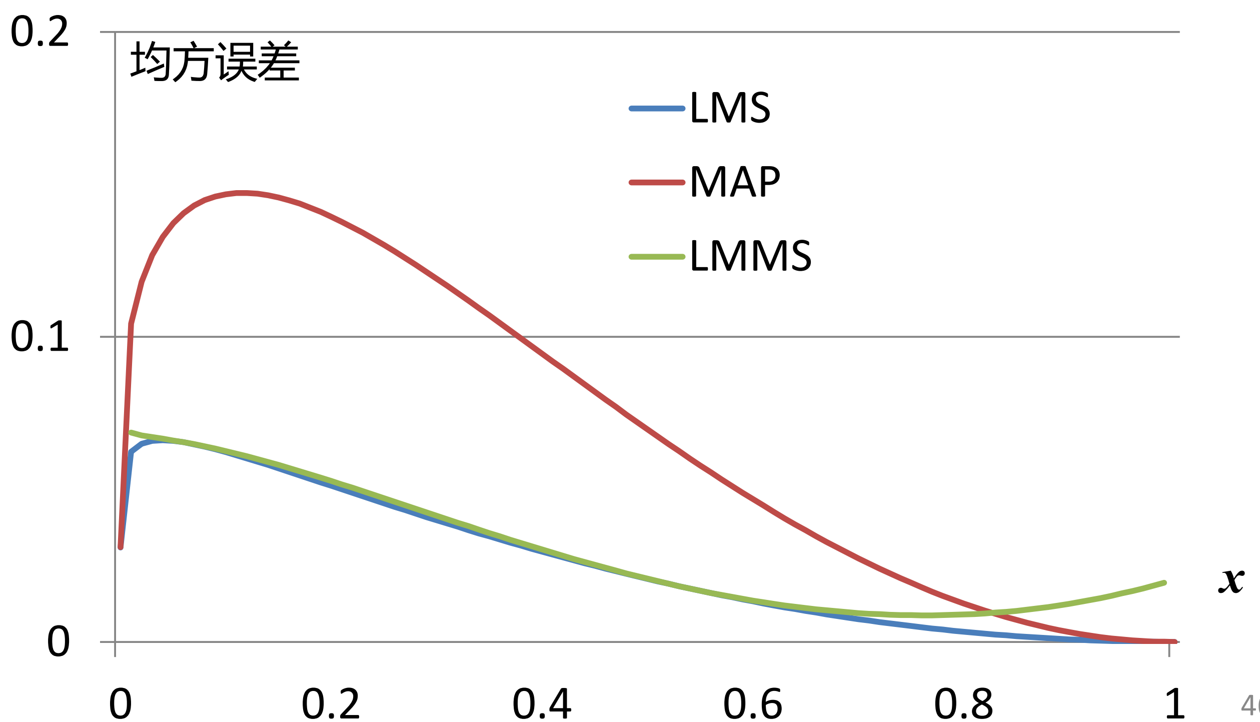
$$= \frac{1}{2} + \frac{1/24}{7/144}\left(X - \frac{1}{4}\right) = \frac{6}{7}X + \frac{2}{7}$$

相应的条件均方误差

由前例

$$E\left[\left(\hat{\theta}-\Theta\right)^2 \middle| X=x\right]=\hat{\theta}^2-\hat{\theta}\frac{2(1-x)}{|\ln x|}+\frac{1-x^2}{2|\ln x|}$$

代入$\hat{\theta}=\frac{6}{7}x+\frac{2}{7}$即可



§6小结和讨论

- 贝叶斯和经典统计推断
- 贝叶斯方法——将参数/模型看作具有先验分布的随机变量Θ，最感兴趣的目标是给定观测时Θ的后验分布
- 原则上后验分布可以通过贝叶斯准则计算
- 贝叶斯提供了一种自然的嵌入先验模型的方法
- 参数估计方法：
  - 最大后验概率准则(使Θ的后验概率达到最大)是用途广泛的推断方法，可以用于估计和假设检验问题
  - 基于使Θ和它的估计量的均方误差最小化原则的估计
    - 最小均方(或条件期望)估计
    - 线性最小均方估计量——有时会导致较大的均方误差，但是计算简单，且只与相关变量的均值、方差和Θ与观测之间的协方差有关