数据科学导论

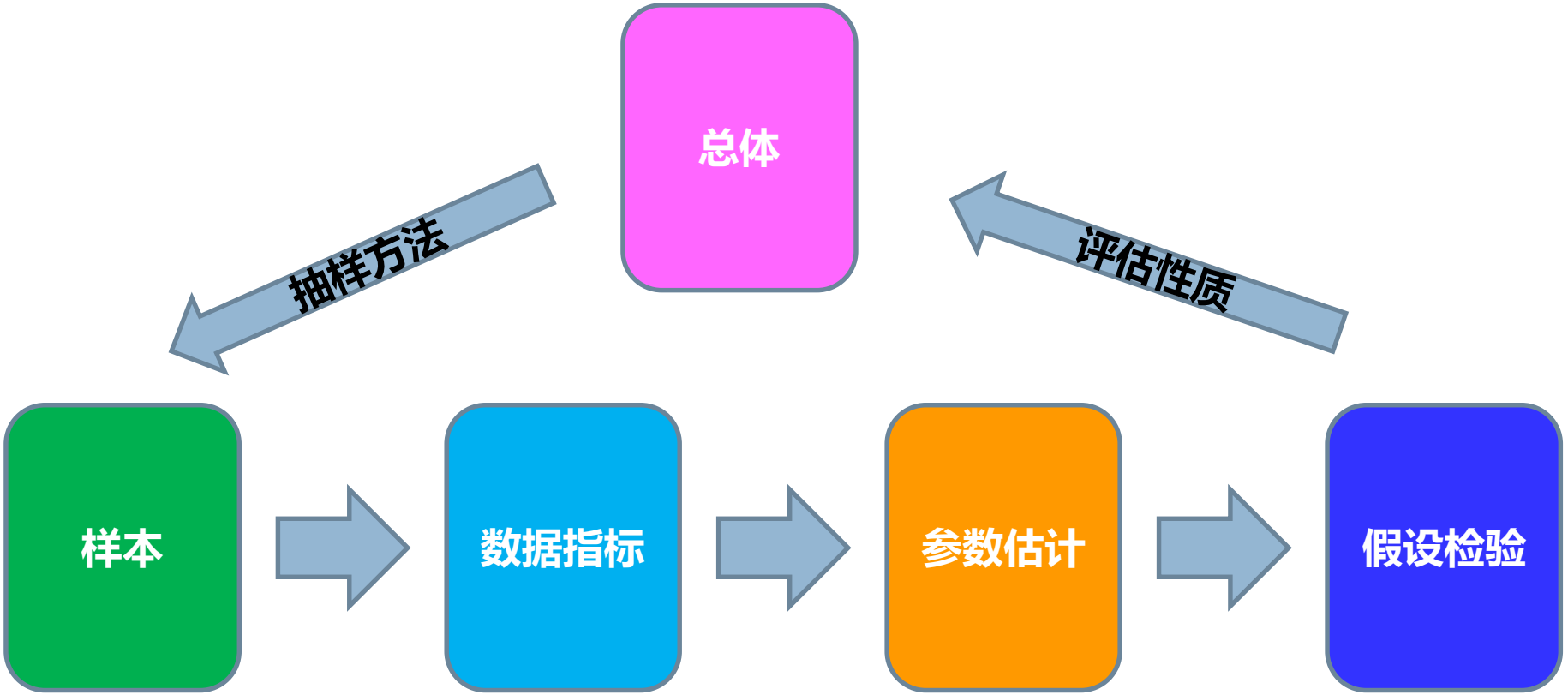
Introduction to Data Science

第三章 数据统计基础

黄振亚，陈恩红，刘淇
Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html

回顾：数据统计



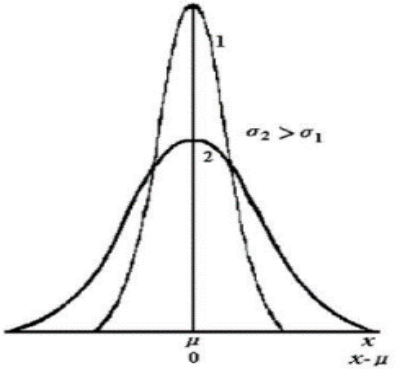
回顾：数据统计

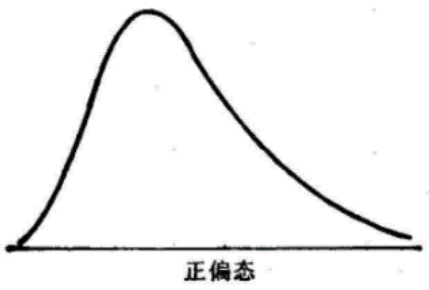
- 数据分布
- 参数估计
- 假设检验
- 抽样方法

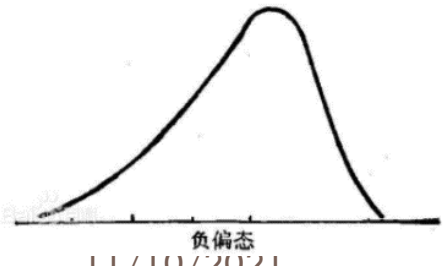
回顾：数据分布

- 在对大数据进行研究时，首先希望知道所获得的数据的**基本分布特征**

- 数据分布的特征可以从三个方面进行测度和描述：
  - 描述数据分布的**集中趋势**：反映数据向其中心靠拢或聚集程度
  - 描述数据分布的**离散程度**：反映数据远离中心的趋势或程度
  - 描述数据分布的**形状变化**：反应数据分布的形状特征







回顾：参数估计

- 参数(parameter)
  - 参数 是用来描述**总体数据特征**的度量
- 统计量(statistic)
  - 统计量 是用来描述**样本数据特征**的度量
    - 由试验计算得出，不依赖于任何其他未知的量（特别是不能依赖于总体分布中所包含的未知参数）
- 参数估计(parameter estimation)
  - 是统计推断的基本问题之一：用**样本统计量**估计总体的**参数**
    - 参数未知的真实
    - 统计量已知的估计
  - 例：掷骰子例子

回顾：参数估计

- 点估计
  - 用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值
- 点估计的常用方法
  - 矩估计
  - **最小二乘估计 LSE**
  - 极大似然估计 MLE
  - 最大后验估计 MAP
  - 贝叶斯估计

回顾：矩估计

- 举例：黑白球（矩估计）
  - 例：假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。每次任意从已经摇匀的罐中拿1个球出来，记录球的颜色，然后把拿出来的球再放回罐中。假如在前面的100次重复记录中，有70次是白球。请问罐中白球所占的比例是多少？

解：用样本中白球比例的均值作为估计代替总体均值。

即估计结果为罐中白球所占的比例$70\% = \frac{7}{10}$

符合直观

(独立同分布，无偏估计)

参数估计—矩估计 vs 最小二乘估计

- 举例：黑白球（最小二乘估计）—课堂练习

目标：最小化估计值θ与观测值$\hat{\theta}$之差的平方和

$$\min L(\theta) = \sum_{i=1}^{N}(\theta - \hat{\theta}_i)^2$$

解：假设：白球占比为θ

已知100次观测量：出现白球概率$\frac{7}{10}$，黑球概率$\frac{3}{10}$

则，利用最小二乘估计的目标为：

$$\min L(\theta) = \sum_{i=1}^{N}\left(\theta - \hat{\theta}_i\right)^2$$
$$= [(\theta - 1)^2 \times 70 + (\theta - 0)^2 \times 30]$$

令：$\frac{\partial L(\theta)}{\partial \theta} = 0$

则：$\theta = \frac{7}{10}$

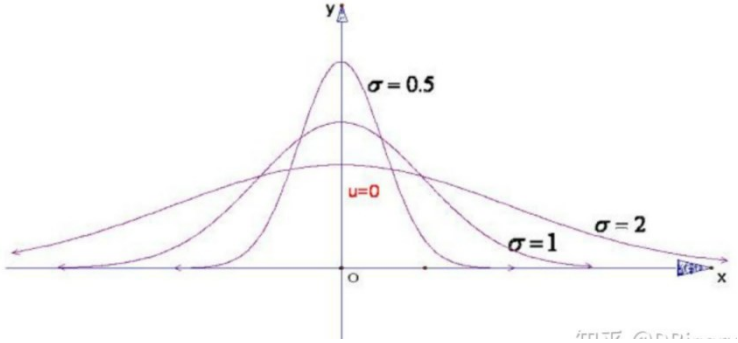
参数估计

- 点估计
  - 用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值
- 点估计的常用方法
  - 矩估计
  - 最小二乘估计 LSE
  - **极大似然估计 MLE**
  - **最大后验估计 MAP**
  - **贝叶斯估计**

参数估计—极大似然估计

- 极大**似然**估计(Maximum Likelihood Estimate，MLE)
  - 思想：利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值
  - **模型已定，参数未知**
  - 目标：概率分布函数或者似然函数最大
    - 用似然函数取到最大值时的参数值作为估计值
  - 概率分布模型
    - 伯努利分布
    - 二项分布
    - 高斯分布
    - 泊松分布

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



参数估计—极大似然估计

- 极大似然估计(MLE)
  - MLE目标：用似然函数取到最大值时的参数值作为估计值
  - 设总体分布为$f(X|\theta)$，$x_1, x_2, x_3, \cdots, x_N$为样本。样本满足独立同分布，则他们的联合密度函数为：

$$\mathrm{L}(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

  - 其中，θ为未知参数。样本已经存在(观测)，即，$x_1, x_2, x_3, \cdots, x_n$是固定的。$\mathrm{L}(X|\theta)$是关于$\theta$的函数，称为似然函数
  - 目标：求参数θ，**使似然函数取极大值，称为极大似然估计**
  - 实践中，通常对似然函数取对数(log或ln)(连乘运算变为连加运算)，即**对数似然函数**。所以，极大似然估计问题可以写成

$$\ln \mathrm{L}(x_1, x_2, \cdots, x_n|\theta) = \sum_{i=1}^{\mathrm{n}} ln\big(f(x_i|\theta)\big)$$

参数估计—极大似然估计

- 例子: 扔硬币，X每次实验X_i服从伯努利分布
  - 参数为$\boldsymbol{\theta}$，假设为事件(正面向上)发生的概率

$$P(X_i \mid \theta)=\begin{cases}\theta, & X_i\text{为正面}\\ 1-\theta, & X_i\text{为反面}\end{cases}$$

- n次实验，共k次正面向上，采用MLE估计参数$\boldsymbol{\theta}$：

样本观测

正 反

k次 N-k次
5次 5次

产生观测样本的概率不同

目标：找到发生样本最大概率的参数

总体参数

	正	反
θ	0.5	0.5

	正	反
θ	0.2	0.8

	正	反
θ	0.9	0.1

参数估计—极大似然估计

- 例子: 扔硬币，X每次实验X_i服从伯努利分布
  - 参数为$\boldsymbol{\theta}$，假设为事件(正面向上)发生的概率

$$\boldsymbol{P(X_i \mid \theta)}=\begin{cases}\boldsymbol{\theta}, & \boldsymbol{X_i}\text{为正面}\\ \boldsymbol{1-\theta}, & \boldsymbol{X_i}\text{为反面}\end{cases}$$

- n次实验，共k次正面向上，采用极大似然估计估计参数$\boldsymbol{\theta}$：
  - 似然函数： $\mathrm{L}(x_1, x_2, \cdots, x_n|\theta)=\boldsymbol{C_n^k\theta^k(1-\theta)^{n-k}}$
  - 对数似然函数：$\ln \mathrm{L}(X|\theta) =\ln \boldsymbol{C_n^k + k\ln\theta + (n-k)\ln(1-\theta)}$
  - 求极值：$\frac{\partial L(\theta)}{\partial \theta} = 0$，则：$\frac{k}{\theta}-\frac{n-k}{1-\theta}=0$
  - 参数$\boldsymbol{\theta}$的最大似然估计值: $\theta_{MLE} = \frac{k}{k+n-k} = \frac{k}{n}$

✓ 二项分布中每次事件发生的概率$\boldsymbol{\theta}$ = 做N次独立重复随机试验中事件发生的概率
✓ 例如：如果做20次实验，出现正面14次，反面6次:
  ✓ MLE得到参数值p为14/20 = 0.7

参数估计—极大似然估计

- 极大似然估计—高斯分布的参数
  - 例：给定$x_1, x_2, x_3, \cdots, x_N$为样本，已知样本来自于高斯分布$N(\mu, \sigma)$,估计参数$\mu, \sigma$

解：

- 高斯分布的概率密度函数：$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
- 带入样本，似然函数：$$L(X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
- 对数似然：$$\ln L(X) = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
$$= -\frac{n}{2}\ln(2\pi\sigma^2) + -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$
- 求偏导估计参数：$$\mu = \frac{1}{n}\sum_i x_i \qquad \sigma^2 = \frac{1}{n}\sum_i (x_i-\mu)^2$$

与矩估计结果相同。

参数估计—矩估计 vs LSE vs MLE

- 举例：黑白球（极大似然估计）
  - 例：假如有一个罐子，里面有黑白球，数目不知，两种颜色比例也不知。每次任意从摇匀的罐中拿1个球出来，记录颜色，然后把拿出来的球再放回罐中。假如在前面的100次重复记录中，有70次是白球。请问罐中白球所占的比例是多少？

解：假设：白球占比为θ，黑球占比为1-θ

已知100次观测量：出现白球70次，黑球概率30次

则，极大似然估计的目标为：

$$\max L(x_1, x_2, \cdots, x_{100}|\theta) = C_{100}^{70}\theta^{70}(1-\theta)^{30}$$

取对数为：$\ln\left(L(x_1, x_2, \cdots, x_{100}|\theta)\right) = 70 \times ln\theta + 30 \times ln(1-\theta)$

令：$\frac{\partial L(\theta)}{\partial \theta} = 0$

则：$\theta = \frac{7}{10}$

矩估计、LSE、MLE结果均相同，但原理不同

参数估计—课后学习与思考

- 矩估计 vs LSE vs MLE—关联与区别
  - LSE可以通过高斯分布+MLE推算出来
  - LSE和MLE对应机器学习中的经验风险最小化
- MLE
  - 似然函数取对数后导数还是不好求：期望最大算法（EM）
  - 高斯混合模型
  - 机器学习中的交叉熵
  - 线性模型的极大似然估计方法
  - 逻辑斯蒂回归的极大似然估计方法

参数估计—最大后验估计

- 回顾扔硬币的例子
  - X每次实验X_i服从伯努利分布
  - 假设为事件(正面向上)发生的概率，参数为$\boldsymbol{\theta}$，
  - n次实验，共k次正面向上，目标为估计参数$\boldsymbol{\theta}$

- MLE的思想
  - $L(X|\theta)$ 似然函数取到最大值时的参数值作为估计值

样本观测

正 反

k次 N-k次
5次 5次

MLE

总体参数

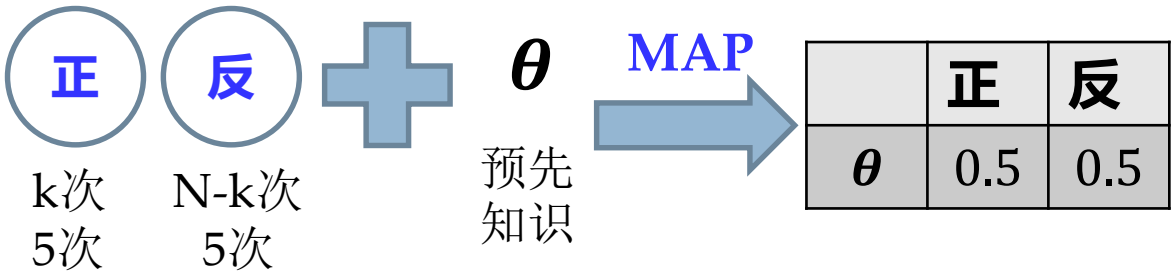
	正	反
$\boldsymbol{\theta}$	0.5	0.5

- 频率学派
  - 完全相信数据
  - 世界是确定的
  - 事件在多次重复实验中趋于稳定

参数估计—最大后验估计

- MLE是否有不足？—实验是否靠谱？
  - 实验对象(硬币)是否均匀？实验次数是否足够？
  - 实验环境是否有影响？。。。
- 最大后验估计(Maximum A Posteriori Estimation, MAP)
  - 目标：最大化在给定数据样本X的情况下模型参数的**后验概率**
    - 模型参数使得模型能够产生该数据样本的概率最大—**似然概率**
    - 但对于模型参数有了一个**假设**，加入了**先验知识**，即模型参数可能满足某种分布，即，估计不止依赖数据样本。

样本观测　　**参数假设**　　**总体参数**



- **贝叶斯学派**
  - 不能完全相信数据
  - 世界是不确定的
  - 数据量的增加，参数向数据靠拢—先验影响越小

参数估计—最大后验估计

- 贝叶斯公式：

$$P(A,B) = P(B) * P(A|B) = P(A) * P(B|A)$$

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

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

$$\sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

- 因果概率

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Effect)}{P(Effect)}$$

参数估计

- 贝叶斯公式的理解—举例
  - 已知：
    - 临床案例发现：患者得 meningitis(脑膜炎) 导致 stiff neck(颈部僵硬) 的概率为 50%
    - 先验知识：患者得 meningitis 的概率为 1/50,000
    - 先验知识：患者得 stiff neck 的概率为 1/20
  - 问：如果患者得 stiff neck, 那么他患有meningitis的概率为?
    - 设 *M* 为患 meningitis（脑膜炎）的概率, *S* 为患 stiff neck 的概率:

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

注: 患有 meningitis 的后验概率 仍然非常小

参数估计—最大后验估计

- 最大后验概率估计(MAP)
  - 已知：$x_1, x_2, x_3, \cdots, x_N$为样本，问：估计总体的参数$\theta$

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

    - p($\theta|X$)是后验概率，估计的目标：已知数据X，求参数θ的值
    - p($X|\theta$)是**似然函数：回顾MLE**
    - p(θ)是**先验概率：指在没有任何实验数据的时候对参数θ的判断**
    - p(X)是边缘概率：指我们的观测(也叫证据，evidence)
  - 理解：对比MLE
    - MLE的目标是：求参数θ，使得似然函数p($X|\theta$)最大
    - MAP的目标是：求参数θ，使得似然函数p($X|\theta$) p(θ)最大
      - 不仅需要似然函数出现的概率大，也需要参数θ的先验概率大

参数估计—最大后验估计

- 最大后验概率估计(MAP)
  - 已知：$x_1, x_2, x_3, \cdots, x_N$为样本，问：估计总体的参数$\theta$

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - 整理MAP的优化目标为

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\boldsymbol{MAP}} &= argmax_\theta \frac{p(\theta|X)p(\theta)}{p(X)} \\ &\propto argmax_\theta p(\theta|X)p(\theta) \\ &= argmax_\theta \log(p(\theta|X)) + \log(p(\theta)) \\ &= argmax_\theta \{\sum_{x_i \in X} \log(p(\theta|x_i)) + \log(p(\theta))\}\end{aligned}$$

注意这里p(X)与参数θ无关，因此等价于要使分子最大

与MLE相比，多加一个先验分布概率的对数

参数估计—最大后验估计

- 最大后验概率估计(MAP)—理解先验p(θ)

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - 先验$p(\theta)$可以用来描述人们已知或者接受的普遍知识和规律。根据发生的事情做判断时，要考虑所有因素。它会影响参数估计过程中我们对观测数据$p(X)$的相信程度。
  - 在实际中，这样的知识和规律非常普遍
    - 扔硬币：通常认为硬币是均匀的
    - 期末考试：通常认为学霸分数高
    - 导师批评：通常认为学生犯了错误

    → 硬币可能是不均匀的
    → 学霸当天发挥不好
    → 导师当天心情不好

    - 一辆汽车（或者电瓶车）的警报响了，大家会想到什么？
    - 有小偷？撞车了？汽车被砸了 → 无事发生

为什么会这么认为？ 如何修正这样的认知？

参数估计—最大后验估计

- 最大后验概率估计(MAP)—理解先验p(θ)

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - **先验$p(\theta)$**可以用来描述人们已知或者接受的普遍知识和规律。
    - 例如：在扔硬币的试验中，每次抛出正面发生的概率应该服从一个概率分布，这个概率在0.5处取得最大值（均匀），这个分布就是先验分布。先验分布的参数(一个或多个)我们称为超参

$$p(\theta) = p(\theta|\alpha)$$

  - 当上述后验概率取得最大值时，我们就得到根据MAP估计出的参数值。给定观测到的样本数据，一个新的值 $\tilde{x}$ 发生的概率可以用以下公式来估计：

$$p(\tilde{x}|X) = \int_{\theta \in \Theta} p(\tilde{x}|\hat{\theta}_{MAP}) p(\theta|X) d\theta = p(\tilde{x}|\hat{\theta}_{MAP})$$

参数估计—最大后验估计

- 最大后验概率估计(MAP) — 理解先验$p(\theta)$
  - 扔硬币的例子：10次实验，其中正面朝上(参数：θ)的次数为7次，反面朝上的次数为3次，结果记为(1,0,1,1,0,1,0,1,1,1)
  - 设定先验分布：通常认为$\theta=0.5$的可能性最大，因此用均值为0.5，方差为0.1的高斯分布来描述θ的先验分布$p(\theta|\mu,\sigma)$

$$p(\theta|\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}=\frac{1}{10\sqrt{2\pi}}e^{-50(\theta-0.5)^2}$$

  - 求解MAP $p(\theta|\boldsymbol{X})\propto p(\boldsymbol{X}|\theta)p(\theta)=\theta^7(1-\theta)^3\times\frac{1}{10\sqrt{2\pi}}e^{-50(\theta-0.5)^2}$
  - 取对数：$\boldsymbol{ln}\,p(\theta|\boldsymbol{X})\propto 7\boldsymbol{ln}(\theta)+3\boldsymbol{ln}(1-\theta)+\boldsymbol{ln}\left(\frac{1}{10\sqrt{2\pi}}\right)-50(\theta-0.5)^2$
  - 求导解得：$\hat{\theta}\approx 0.558$
  - 若用均值为0.7，方差为0.1的高斯分布来描述描述θ的先验分布$p(\theta|\mu,\sigma)$，解得：$\hat{\theta}=0.7$

合理的先验分布很重要

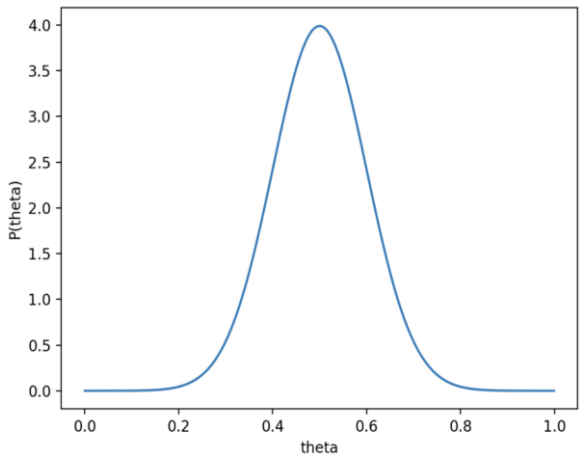
参数估计—最大后验估计

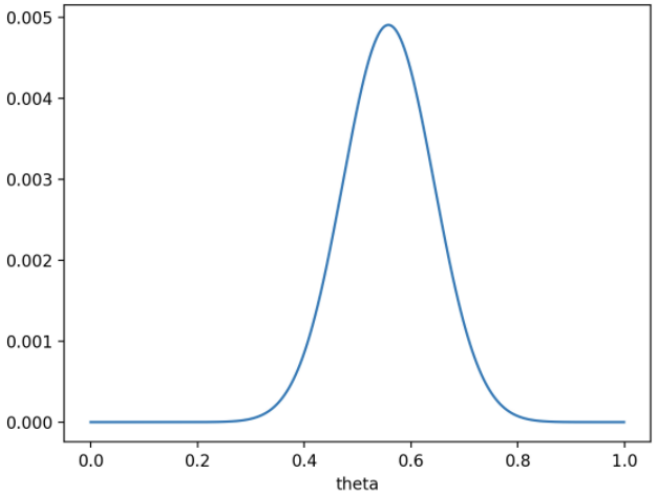
- 最大后验概率估计(MAP) — 理解先验p(θ)
  - 扔硬币的例子：

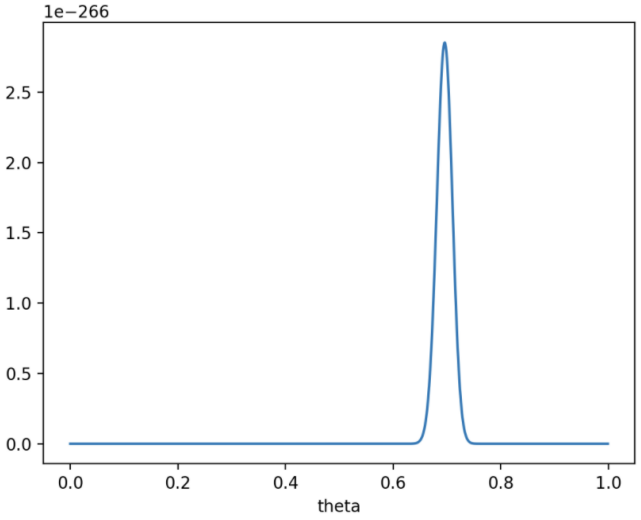
先验 $p(\theta|\mu,\sigma)$
均值0.5，方差0.1

10次实验，其中**正面朝上**(θ)**7次**，反面朝上**3次**
MLE解得：θ=0.7
MAP解得：θ=0.558

1000次实验，其中**正面朝上**(θ)**700次**，反面朝上**300次**
MLE解得：θ=0.7
MAP解得：θ=0.696







数据实验的次数增加，先验分布的影响越小

参数估计——最大后验估计

- 最大后验概率估计(MAP)—理解先验p(θ)
  - 扔硬币的例子：我们期望先验概率（待估计的参数$\boldsymbol{\theta}$）分布在0.5处取得最大值，可以选用**Beta分布**（$\boldsymbol{\theta}$服从Beta分布）即：

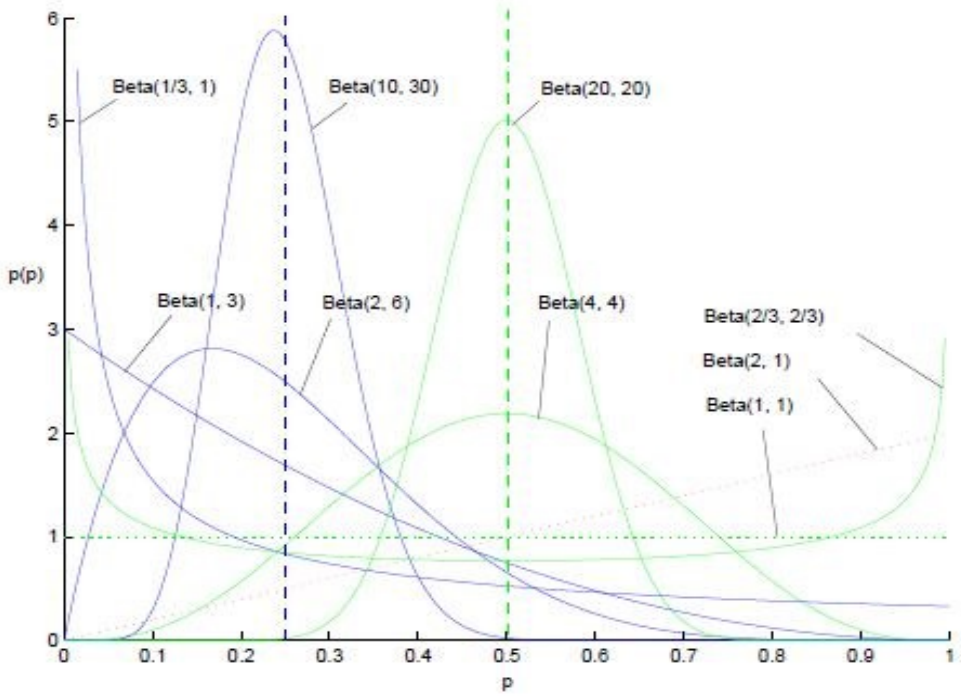
$$p(\theta|\alpha,\beta) \triangleq Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\boldsymbol{\alpha-1}}(1-\theta)^{\beta-1}$$

  - 其中，Beta函数是 $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
  - Gamma函数 $\Gamma(n) = (n-1)!$
  - Beta分布的随机变量范围是[0,1]，不同参数情况下的Beta分布的概率密度函数形式如图

参数估计——最大后验估计

- 最大后验概率估计(MAP)—理解先验p(θ)

在0.5



下的

Fig. 1. Density functions of the beta distribution with different symmetric and asymmetric parametrisations.

参数估计—最大后验估计

- 最大后验概率估计(MAP)—理解先验p(θ)
  - 扔硬币的例子：我们期望待估计的参数θ的先验分布在0.5处取得最大值，可以选用**Beta分布**（θ服从Beta分布）即：

$$p(\theta|\alpha,\beta) \triangleq Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  - 取 $\boldsymbol{\alpha=\beta=5}$，使得先验分布Beta分布在0.5处取得最大值
  - 使用MAP方法求解参数

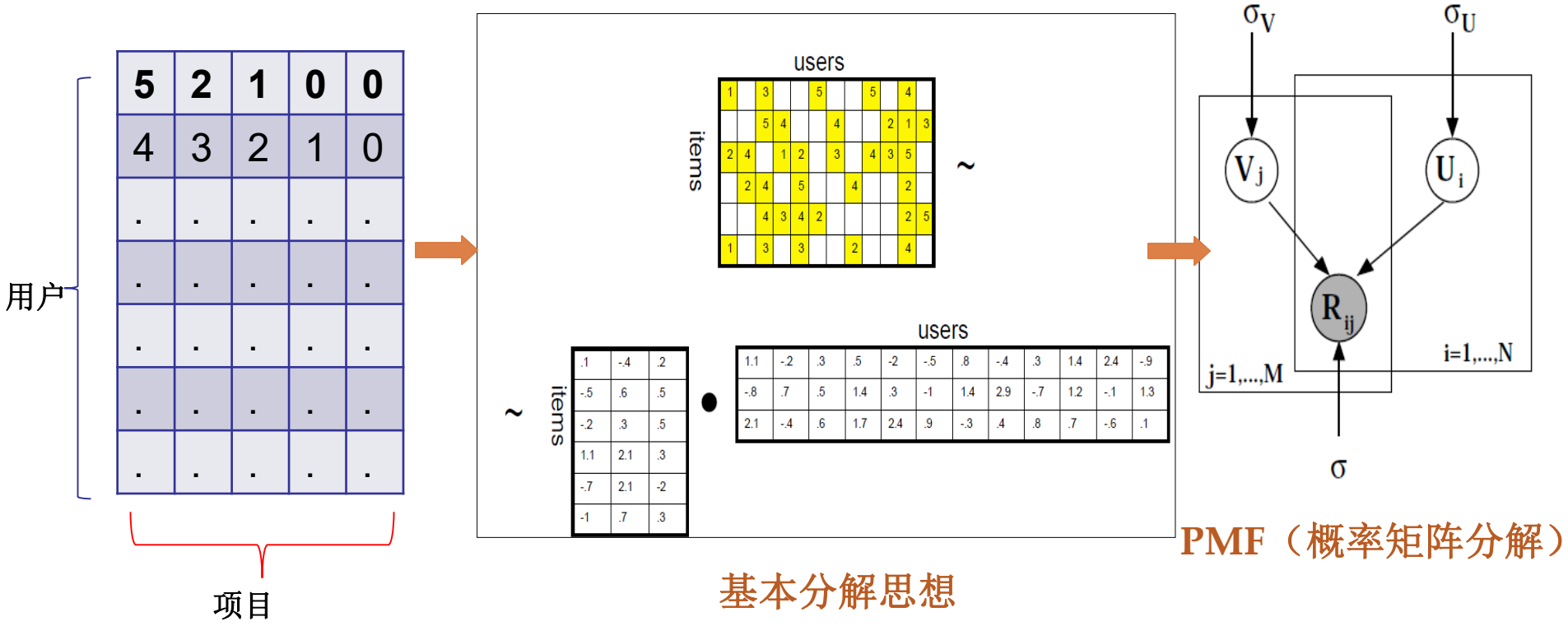
$$\hat{\theta}_{MAP} = \frac{M_1+\alpha-1}{M_1+M_2+\alpha+\beta-2} = \frac{M_1+4}{M_1+M_2+8}$$

  - 与MLE相比，结果中多了α-1和$\alpha+\beta-2$，即先验作用，且超参数越大，为了改变先验分布传递的belief所需要的观察值就越多
  - 同样表明“**硬币一般是均匀的**”这一先验对参数估计的影响

思考：如果先验P(θ=0.5)=1？

参数估计—最大后验估计

- 基于模型的协同过滤—概率矩阵分解
  - 面向评分预测的模型

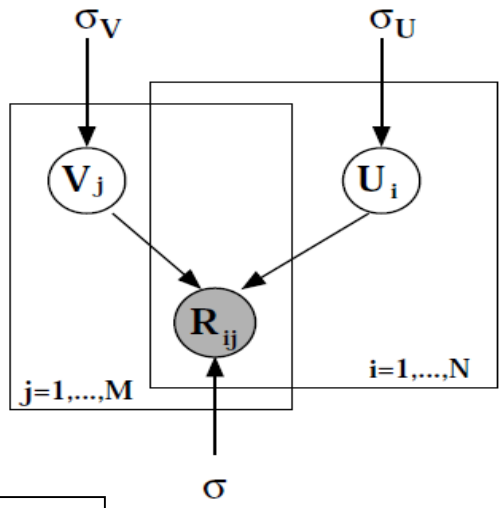


➢ Mnih, Andriy, and Russ R. Salakhutdinov. "Probabilistic matrix factorization." Advances in neural information processing systems. 2008.

参数估计—最大后验估计

- 基于模型的协同过滤—概率矩阵分解
  - 最大后验概率方法估计参数U和V (θ)
    - 目标： $p(U,V \mid R,\sigma^2,\sigma_U^2,\sigma_V^2)$

$$\propto \underbrace{p(R \mid U,V,\sigma^2)}_{\text{似然函数}} * \underbrace{p(U \mid \sigma_U^2) * p(V \mid \sigma_V^2)}_{\text{先验}}$$



    - MLE似然

$$p(R|U,V,\sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}(R_{ij}|U_i^T V_j,\sigma^2)\right]^{I_{ij}}$$

    - 假设先验：

$$p(U|\sigma_U^2) = \prod_{i=1}^{N}\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})$$

$$p(V|\sigma_V^2) = \prod_{j=1}^{M}\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})$$

超参

参数估计—最大后验估计

- 基于模型的协同过滤—概率矩阵分解
  - MAP learning

$$\ln p(U,V|R,\sigma^2,\sigma_V^2,\sigma_U^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij}-U_i^T V_j)^2 - \frac{1}{2\sigma_U^2}\sum_{i=1}^{N} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{M} V_j^T V_j$$
$$-\frac{1}{2}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\right)\ln\sigma^2 + ND\ln\sigma_U^2 + MD\ln\sigma_V^2\right) + C, \qquad (3)$$

  - Equivalent to minimize sum-of-squared-errors with quadratic regularization terms.

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(R_{ij}-U_i^T V_j\right)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2$$

$$\lambda_U = \frac{\sigma^2}{\sigma_U^2}, \lambda_V = \frac{\sigma^2}{\sigma_V^2}$$

机器学习中：正则化项
- 目标：防止过拟合
- 结构风险最小化

参数估计—最大后验估计

- 基于模型的协同过滤—概率矩阵分解

$$E=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(R_{ij}-U_i^TV_j\right)^2+\frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2+\frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2$$

1)Intialize U ,V with small, random values

2)repeat

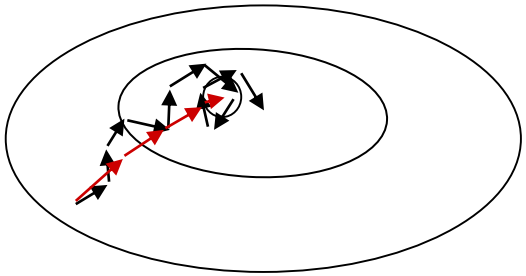
for each record in the training data

2.a) $U_i=U_i-a\frac{\partial E}{\partial U_i}=U_i-a\left(\sum_j I_{ij}(R_{ui}-U_i^TV_j)(-V_j)+\lambda_U U_i\right)$

2.b) $V_j=V_j-a\frac{\partial E}{\partial V_j}=V_j-a\left(\sum_i I_{ij}(R_{ui}-U_i^TV_j)(-U_i)+\lambda_V V_j\right)$

until convergence

优化方法：随机梯度下降(SGD)

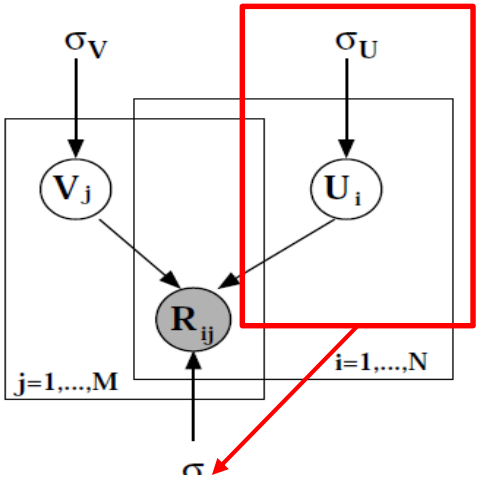


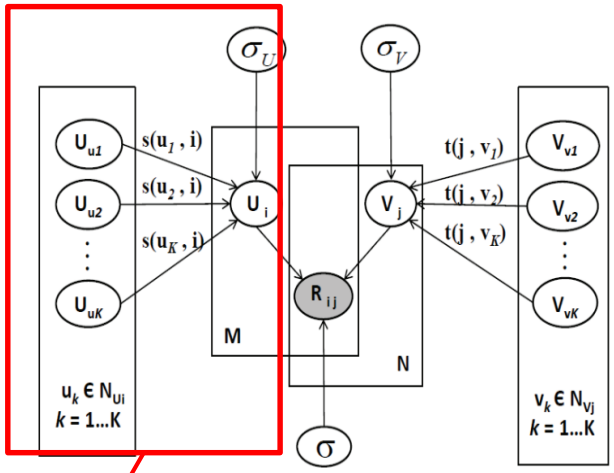
stochastic updates

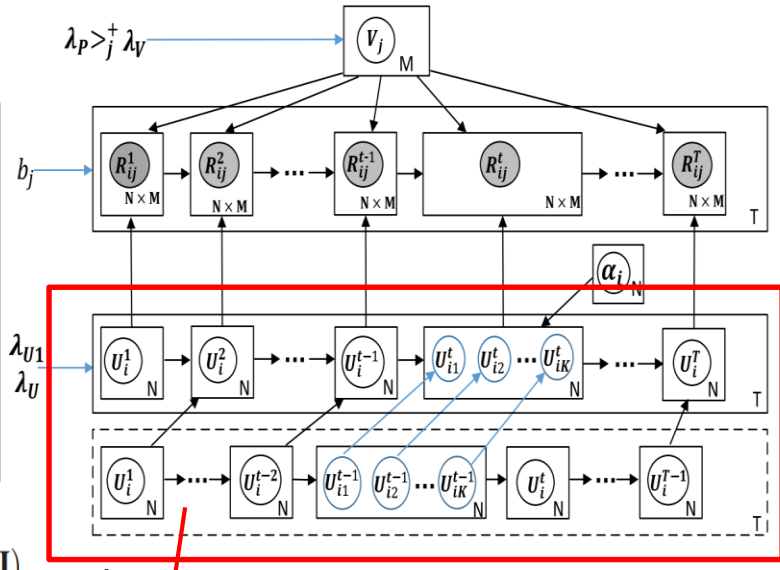
full updates (averaged over all data-items)

参数估计—最大后验估计

- 课后学习：概率矩阵分解







$$p(U|\sigma_U^2) = \prod_{i=1}^{N} \mathcal{N}(U_i|0, \sigma_U^2\mathbf{I})$$

$$p(V|\sigma_V^2) = \prod_{j=1}^{M} \mathcal{N}(V_j|0, \sigma_V^2\mathbf{I})$$

$$U_i = \sum_{l \in N_{U_i}} s(i,l) * U_l + \theta_U, \quad \theta_U \sim N(0, \sigma_U^2\mathbf{I})$$

$$V_j = \sum_{l \in N_{V_j}} t(j,l) * V_l + \theta_V, \quad \theta_V \sim N(0, \sigma_V^2\mathbf{I})$$

$$p\left(U_i^t\right) = \mathcal{N}\left(U_i^t \middle| \bar{U}_i^t, \sigma_U^2\mathbf{I}\right), \quad \text{where } \bar{U}_i^t = \left\{\bar{U}_{i1}^t, \bar{U}_{i2}^t, \ldots, \bar{U}_{iK}^t\right\},$$

$$\bar{U}_{ik}^t = \alpha_i L_{ik}^t(*) + (1-\alpha_i) F_{ik}^t(*), \quad s.t.\ 0 \le \alpha_i \le 1,$$

- Le Wu, Enhong Chen, Qi Liu, et al, Leveraging Tagging for Neighborhood-aware Probabilistic Matrix Factorization. CIKM'2012
- Zhenya Huang, Qi Liu, Le Wu, Keli Xiao, Enhong Chen, Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students, ACM TOIS

参数估计

- MAP的优点
  - 引入了先验知识
  - 在数据量较小时更稳定
- MAP的缺点
  - 和MLE一样，只返回参数的单值估计
    - 导致后验在单值附近有明显尖峰
    - 预测结果不是不确定性上的平均（而是基于单值参数的推断）
  - 当用不同的参数去表示同一分布时，MAP会对超参数很敏感
- 当先验分布均匀时，MAP 估计与 MLE 相等
  - 无信息先验
  - 最大似然方法可被看作一种特殊的 MAP，“让观察数据自己说话”

➢ **Probabilistic graphical models principles and techniques Chap.17.4.4**

参数估计—贝叶斯估计

- 贝叶斯估计—MAP的扩展（MAP没考虑什么？）
  - 已知：$x_1, x_2, x_3, \cdots, x_N$为样本，问：估计总体的参数$\theta$

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

  - 回顾实验的假设
    - 频率角度：样本独立性假设（参数作为固定的值）
    - 贝叶斯角度：条件独立性假设，p(X)也与参数θ有关，每次实验都会告诉我们一些关于θ的信息，因此会改变后面实验的概率
      - p(X) —p($X|\theta$): 即给定参数θ， 样本之间是独立的独立

- 相同：MAP一样**将参数θ视为随机变量**
- 不同：算法不是直接估计参数θ的值，而估计**参数θ的概率分布**
  - MLE和MAP都是只返回了的预估值
  - **p(X)**：MAP忽略(与参数无关)，贝叶斯估计估计整个后验，不能忽略

参数估计—贝叶斯估计

- 贝叶斯估计
  - 为了执行贝叶斯估计，首先需要在参数θ和数据X上描述一个联合分布（记住参数此时也是随机变量）$\mathrm{P}(X,\theta)$，易得:

$$\mathrm{P}(X,\theta)=P(X|\theta)P(\theta)$$

第一项刚好是我们之前描述的似然，后一项为先验

  - 由似然和先验，容易由贝叶斯法则导出后验:

$$\mathrm{P}(\theta|X)=\frac{P(X|\theta)P(\theta)}{p(X)}$$

其中$P(X)=\int_{\Theta}P(X|\theta)P(\theta)d\theta$为似然在所有可能参数赋值上的积分

$$P(\theta|X)=\frac{P(X|\theta)P(\theta)}{\int_{\Theta}P(X|\theta)P(\theta)d\theta}$$

一般来说，计算积分是不可能的

可看出，贝叶斯估计的求解非常复杂，因此选择合适的先验分布就非常重要

参数估计—贝叶斯估计

- 贝叶斯估计 $P(\theta) \sim Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
  - 下面仍然以抛硬币为例，此时选择Beta分布作为先验，类似MAP：

$$P(\theta) = \gamma\theta^{\alpha-1}(1-\theta)^{\beta-1}$$ 其中γ为归一化常数

  - Beta分布在这里作为先验来做参数估计尤为有用
    - 假设我们现在只有先验，没有数据，此时来考虑一次单独的硬币投掷X_1，那么贝叶斯方法预测该硬币朝上的概率为：

$$P(x_1=1) = \int_0^1 P(x_1=1|\theta)P(\theta)d\theta$$

$$= \int_0^1 \theta P(\theta)d\theta = \int_0^1 \theta\gamma\theta^{\alpha-1}(1\text{-}\theta)^{\beta-1}d\theta$$

    - 积分后可得：$P(x_1=1) = \frac{\alpha}{\alpha+\beta}$ (积分过程较复杂，此处省略)

结论：Beta分布作为先验表明（假设）我们已经看到 α 次正面朝上和β次反面朝上

参数估计—贝叶斯估计

- 贝叶斯估计 $P(\theta) \sim Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
  - 现在，让我们在先验的基础上加入更多观测，抛硬币实验X中有正面M_1，反面M_2，则后验估计为：

$$P(\theta|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\theta)P(\theta)}{\int P(\boldsymbol{X}|\theta)P(\theta)d\theta} = \frac{\theta^{M_1}(1-\theta)^{M_2}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{\int \theta^{M_1}(1-\theta)^{M_2}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}d\theta}$$

$$= \frac{\theta^{M_1+\alpha-1}(1-\theta)^{M_2+\beta-1}}{\int \theta^{M_1+\alpha-1}(1-\theta)^{M_2+\beta-1}d\theta} = Beta\,(\theta|\alpha+M_1,\beta+M_2)$$

  - 观察：在抛硬币的实验中(似然$p(X|\theta)$为二项分布), 当先验P(θ)为Beta分布时，后验P(θ|X)也为Beta分布，即更新后的参数服从一个新的Beta(α+M1, β+M2)分布

这种情况我们称之为Beta分布是二项分布似然P(X|θ)的**共轭**

共轭先验的意义：
如果“先验概率”和“后验概率”都服从同样的分布类型（参数不同），则计算先验概率和似然概率的乘积就很方便了，只需要将指数相加即可

参数估计—贝叶斯估计

- 贝叶斯估计—课外学习
  - 在应用中，我们常常使用似然的共轭分布作为参数的先验分布（计算便利）
    - 先验分布叫做似然函数P(X|θ)的共轭先验分布
    - 共轭分布总是针对分布中的某个参数θ而言
    - 采用共轭先验的原因是可以使得先验分布和后验分布的形式相同，但是参数不同

常见共轭先验分布

似然

总体分布	参数	共轭先验分布
二项分布	成功概率p	β分布β(α,β)
泊松分布	均值λ	Γ分布Γ(αβ)
指数分布	均值的倒数λ	Γ分布Γ(αβ)
正态分布（方差已知）	均值μ	正态分布N(μ,σ²)
正态分布（均值已知）	方差σ²	倒Γ分布

Beta(α, β)

[PDF] Latent dirichlet allocation
DM Blei, AY Ng, MI Jordan - the Journal of ma
We describe latent Dirichlet allocation (LDA), a
of discrete data such as text corpora. LDA is a
which each item of a collection is modeled as a
☆ 被引用次数：40012 相关文章

- **LDA算法(文本主题的分布)：** 多项式(Multinomial)分布的似然选取参数服从迪利克雷（Dirichlet）分布作为先验
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

参数估计—贝叶斯估计

- 贝叶斯估计
  - **预测：**由后验我们得到了更新后的参数概率分布的估计 **Beta(α+M_1, β+M_2)**，如何利用已有的数据对新数据进行预测？
  - 假设新的数据为 x^*,则有

$$P(x^*|X) = \int P(x^*|\theta)P(\theta|X)d\theta$$

$$p(\theta|X) = \frac{p(X|\theta)\cdot p(\theta)}{p(X)}$$

  - 计算后可以得到(积分过程较为复杂，此处省略)：

$$P(x^* = 1|X) = \frac{\alpha+M_1}{\alpha+\beta+M_1+M_2}$$

  - 可以观察到这个形式与38页一致：没有数据仅使用先验进行预测，这就是共轭先验的好处：

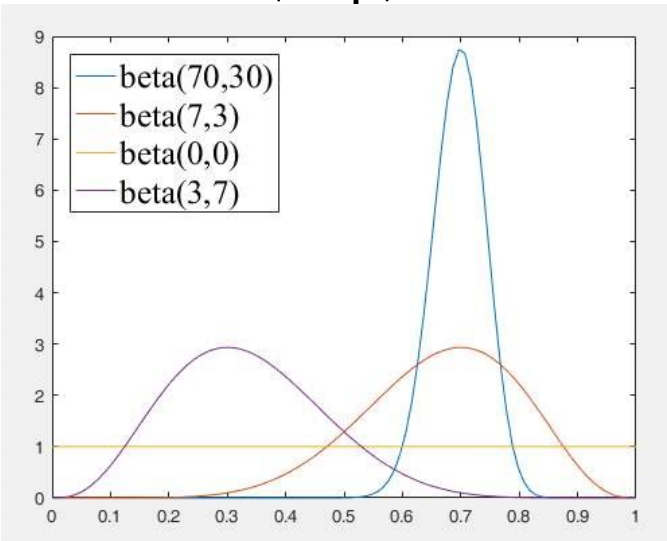
$$P(x_1 = 1) = \int_0^1 P(x_1 = 1|\theta)P(\theta)d\theta \qquad P(x_1 = 1) = \frac{\alpha}{\alpha+\beta}$$

参数估计—贝叶斯估计

- 贝叶斯估计
  - 此时，参数θ的取值（期望）就是这个新的Beta分布Beta(α+M_1, β+M_2)的均值 (M_1次正面，M_2次反面)

$$P(x^*|X) = \int P(x^*|\theta)P(\theta|X)d\theta \qquad \frac{\alpha + M_1}{\alpha + M_1 + \beta + M_2}$$

  - Beta(α, β)的数学期望公式



$$\hat{\theta} = \int_{\Theta} \theta P(\theta|X)d\theta = E(\theta) = \frac{\alpha}{\alpha + \beta}$$

参数估计—贝叶斯估计

- 贝叶斯估计
  - 此时，参数θ的取值（期望）就是这个新的Beta分布Beta(α+M_1, β+M_2)的均值 (M_1次正面，M_2次反面)

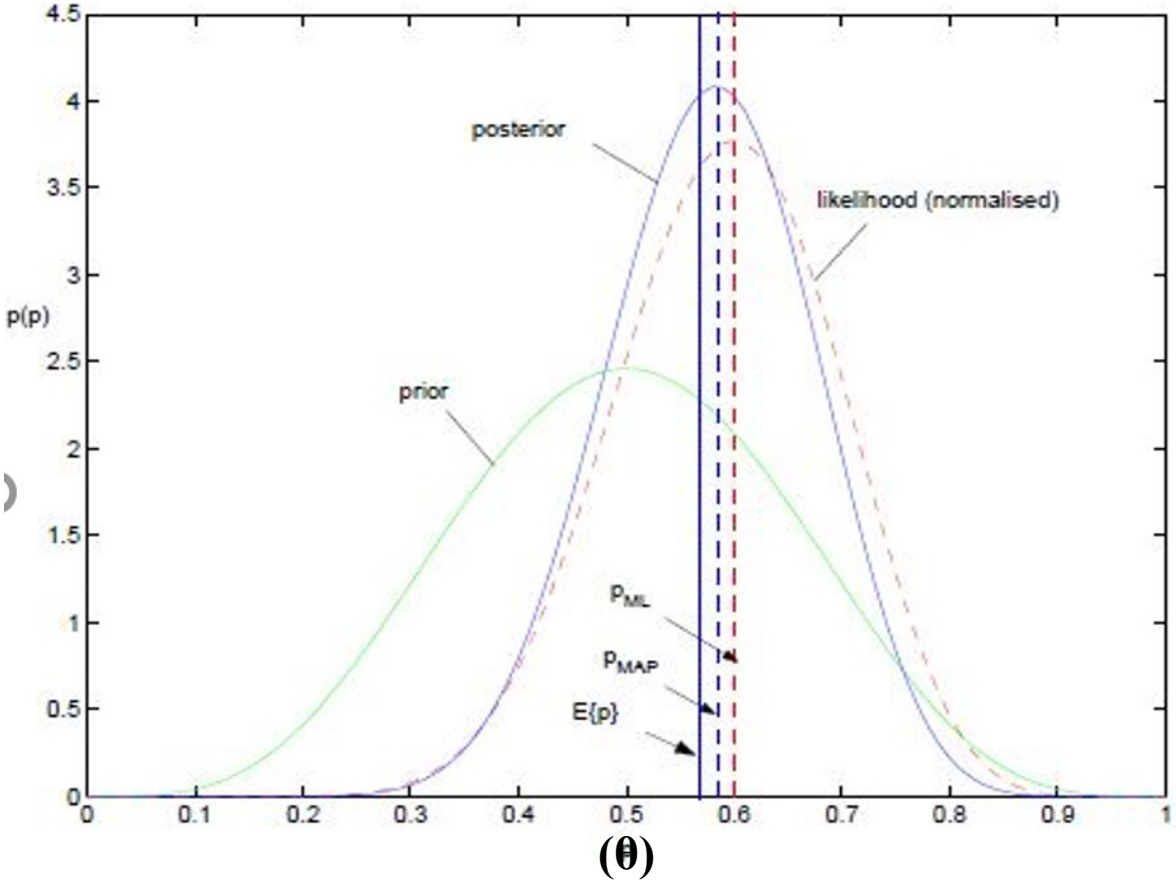
$$\frac{\alpha + M_1}{\alpha + M_1 + \beta + M_2}$$

  - 贝叶斯估计θ的期望和MLE ，MAP中得到的估计值都不同

  - 回顾例子：做20次实验，14次正面，6次反面。
    - ✓ 根据贝叶斯估计得参数θ服从**Beta(14+5, 6+5) 分布**，均值19/30=0. 633
      - MLE: 0.7
      - MAP: 0. 642
      - 贝叶斯估计: 0.633。更加接近先验0.5 （比MLE和MAP小）

参数估计

- MLE、MAP，贝叶斯估计
  - 可视化三个方法对参数的估计结果如下：



结论：

从MLE到MAP再到贝叶斯估计，不断增加先验知识在参数估计过程中的重要性，对参数的表示越来越精确，得到的参数估计结果也越来越接近先验概率0.5。即，越来越能够反映基于样本的真实参数情况。

- **样本数据越少，先验越重要**
- **样本数据越大，三个方法的估计结果差异越小**

参数估计—贝叶斯估计

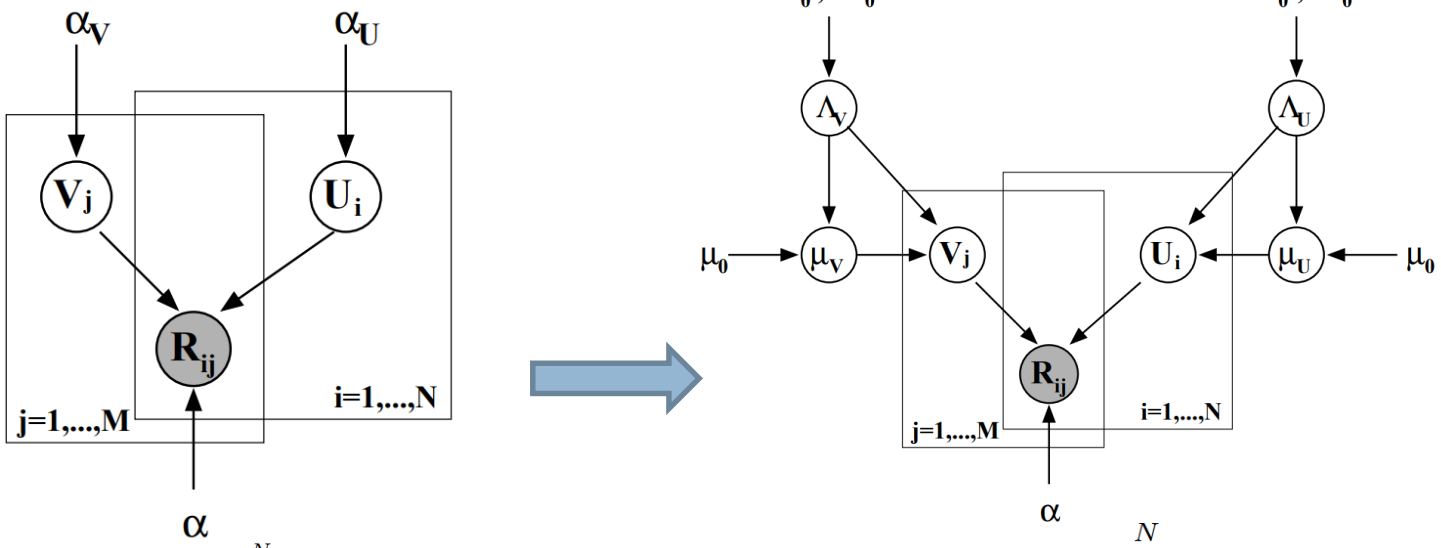
- 贝叶斯估计
  - 抛硬币例子：从理论上来说，贝叶斯估计优于MLE及MAP
  - 存在一些问题：
    - 贝叶斯估计通常需要做积分运算，复杂度较大
    - 有时对积分我们没有一个解析解
    - 有时无法为似然函数likelihood找到合适的共轭先验

因此，人们研究出许多近似方法例如著名的MCMC(Markov chain Monte Carlo) 马尔可夫链蒙特卡洛方法。

课后学习：MCMC方法
http://www.mcmchandbook.net/HandbookChapter1.pdf

参数估计—贝叶斯估计

- 贝叶斯估计—贝叶斯概率矩阵分解BPMF



$$p(U|\alpha_U) = \prod_{i=1}^{N} \mathcal{N}(U_i|0, \alpha_U^{-1} I)$$

$$p(V|\alpha_V) = \prod_{j=1}^{M} \mathcal{N}(V_j|0, \alpha_V^{-1} I),$$

$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \mathcal{N}(U_i|\mu_U, \Lambda_U^{-1}),$$

$$p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(V_i|\mu_V, \Lambda_V^{-1}).$$

课后学习：概率矩阵分解PMF到 贝叶斯概率矩阵分解BPMF

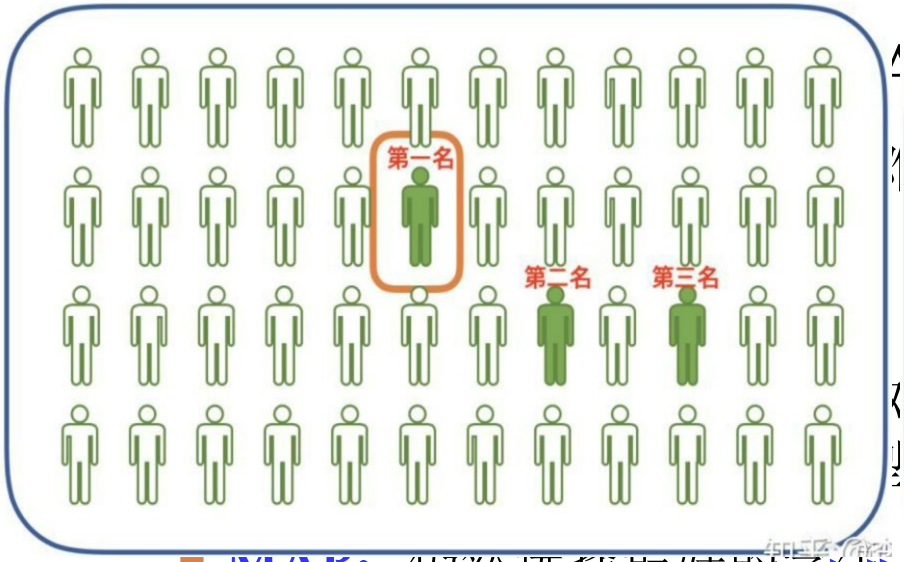
- **Salakhutdinov, Ruslan, and Andriy Mnih. "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo." ICML 2008.**

参数估计

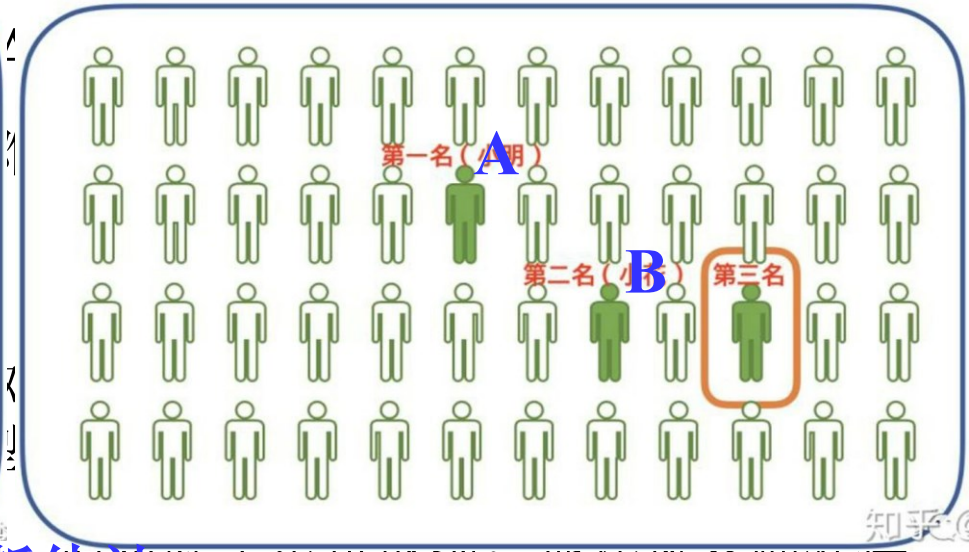
- 应用角度理解：MLE，MAP，贝叶斯估计
  - 举例：假设小江遇到一个计算机难题（数据的预测），碰巧小江有个朋友在大学计算机系当老师，于是他打算找该老师的学生帮忙，那么他该如何寻求帮助呢？
    - **MLE：** 由以往的考试成绩（对应已有数据）排序（A,B,C….）,选出成绩最好的学生A（对应模型中的参数）来解决自己的问题
    - **MAP：** 仍然选择最好的学生，但是除了考试成绩，他还从老师处得知A，B两人考试中有作弊嫌疑（对应先验），结合该知识，小江选择学生C来解决自己的问题
    - **贝叶斯估计：** 此时小江不再寻求单个人的帮助，他会要求每个学生都给出一个答案，并结合考试成绩和老师的提醒给每个学生一个权重（参数的分布），对所有答案加权平均得到最后的解答。

参数估计

MLE



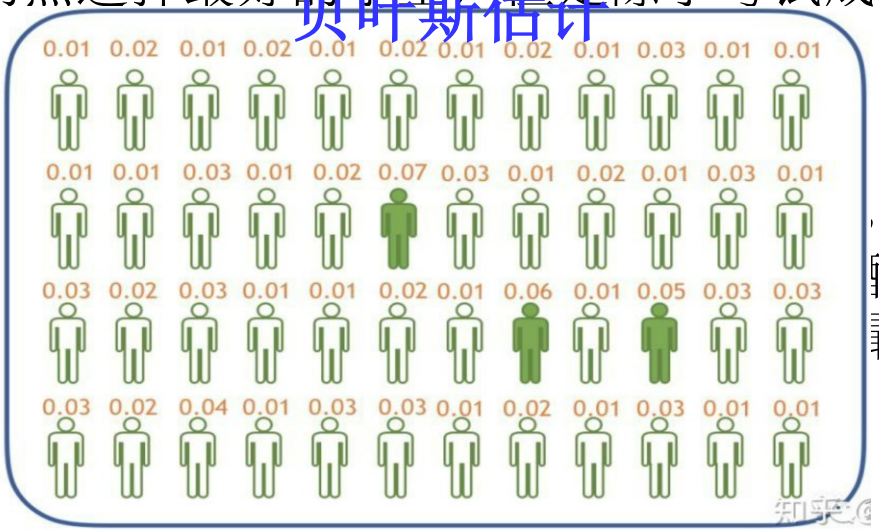
MAP



- MAP：仍然选择最好的学生

知A，B两 结合该知识，小江选择学生C

贝叶斯估计



- 贝叶斯估 他会要求每个学生都给出一 给每个学生一个权重（参数 最后的解答。

参数估计

- 总结： MLE，MAP，贝叶斯估计
  - 从参数估计和模型预测 两个角度—**ML的训练和测试**

    X表示已有数据，θ表示参数，x^* 表示新的未知数据
  - MLE
    - 估计：寻找 $\hat{\theta}$ 使得$P(X|\theta)$最大
    - 预测：$P(x^*|\hat{\theta})$
  - MAP
    - 估计：寻找 $\hat{\theta}$ 使得$P(\theta|X)$最大
    - 预测：$P(x^*|\hat{\theta})$
  - 贝叶斯估计
    - 估计：由数据估计出参数的后验$P(\theta|X)$
    - 预测：$\int_{\Theta} P(x^*|\theta)P(\theta|X)d\theta$

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

参数估计

- 总结：MLE，MAP，贝叶斯估计
  - MAP和贝叶斯估计都考虑了先验，MLE没有
  - MLE和MAP都给出了参数的单值估计，贝叶斯估计给出的是参数的概率分布（后验分布），并通过后验分布做群体决策
  - 样本数无穷时，三种方法都会收敛于同样的结果
  - 贝叶斯估计的计算代价较大，通常选择使用近似算法

参数估计

- 应用场景—课后学习
  - 对一个基础模型，都可以用这三种方法去建模
  - 例如在逻辑回归的模型中：
    - MLE: Logistics Regression
    - MAP: Regularized Logistics Regression
    - 贝叶斯估计: Bayesian Logistic Regression
  - 但是由于它们各自的特性，常用的场景又有所不同：
    - MLE：无先验的回归分类问题，例如 EM算法中的M步
      - EM算法可以看作是含有隐变量情况下MLE的推广
    - MAP：数据量较小时而先验强时，例如变量消元算法中
    - 贝叶斯估计及其近似常用于概率图模型的算法中