本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

数据科学导论

Introduction to Data Science

参数估计

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

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html

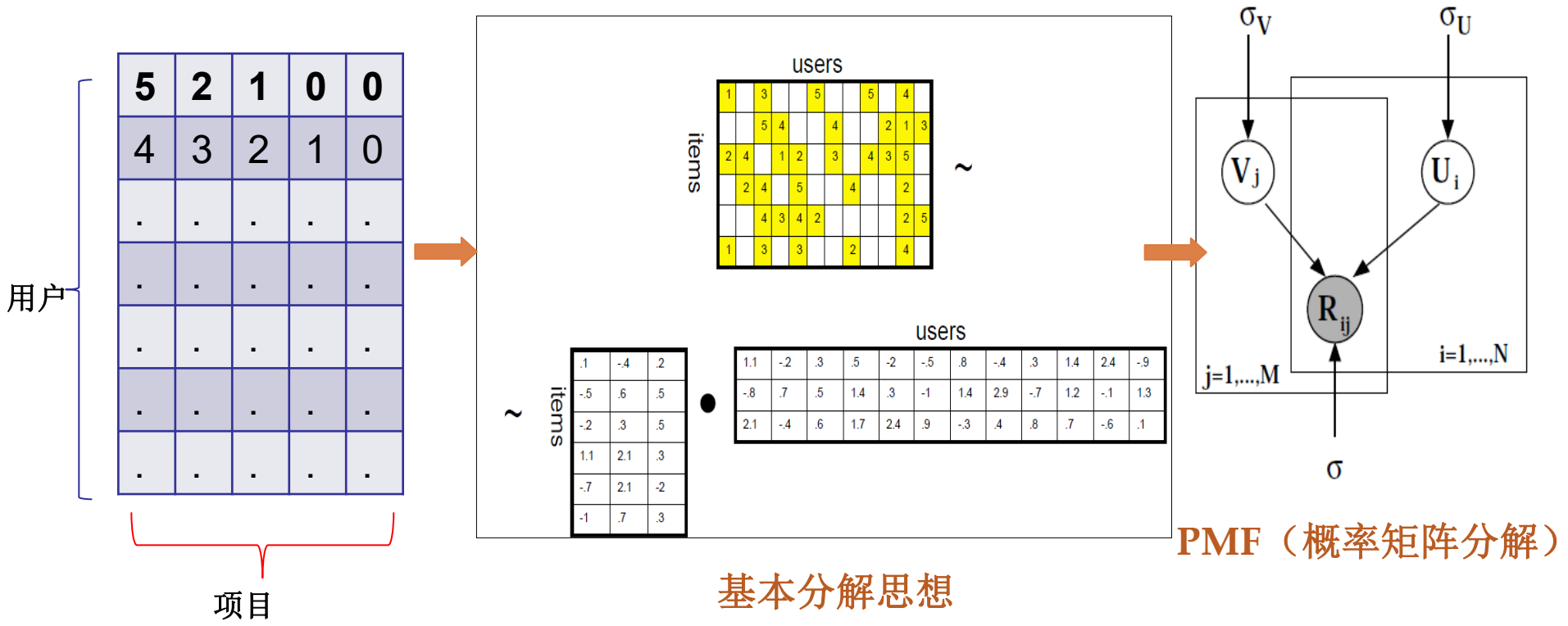


参数估计

- 点估计(point estimate)
  - 点估计是用样本统计量$\hat{\theta}$的某个取值直接作为总体参数θ的估计值
    - 用样本均值x直接作为总体均值μ的估计值
    - 用样本方差s^2直接作为总体方差σ^2的估计值
- 点估计的常用方法
  - 矩估计
  - 最小二乘估计
  - 极大似然估计
  - 最大后验概率
  - 贝叶斯估计

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



基本分解思想

PMF（概率矩阵分解）

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

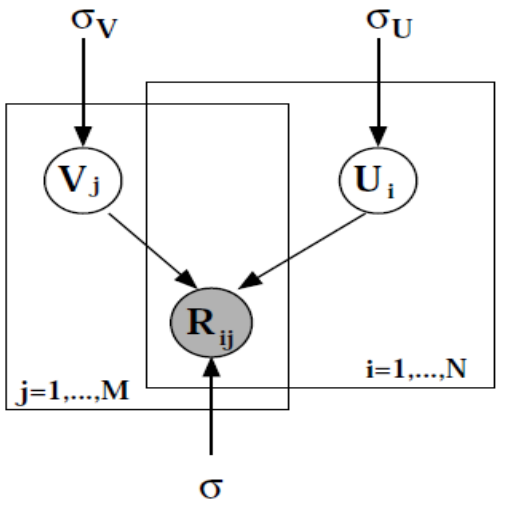
参数估计—最大后验估计

- 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：

$$p(U,V \mid R,\sigma^2,\sigma_U^2,\sigma_V^2)$$

$$\propto p(R \mid U,V,\sigma^2) * p(U \mid \sigma_U^2) * p(V \mid \sigma_V^2)$$

似然函数　　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
    - Max

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

  - 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

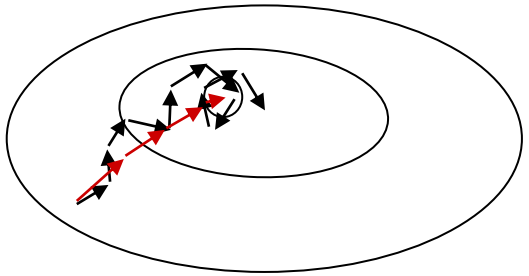
for each record in the training data

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)

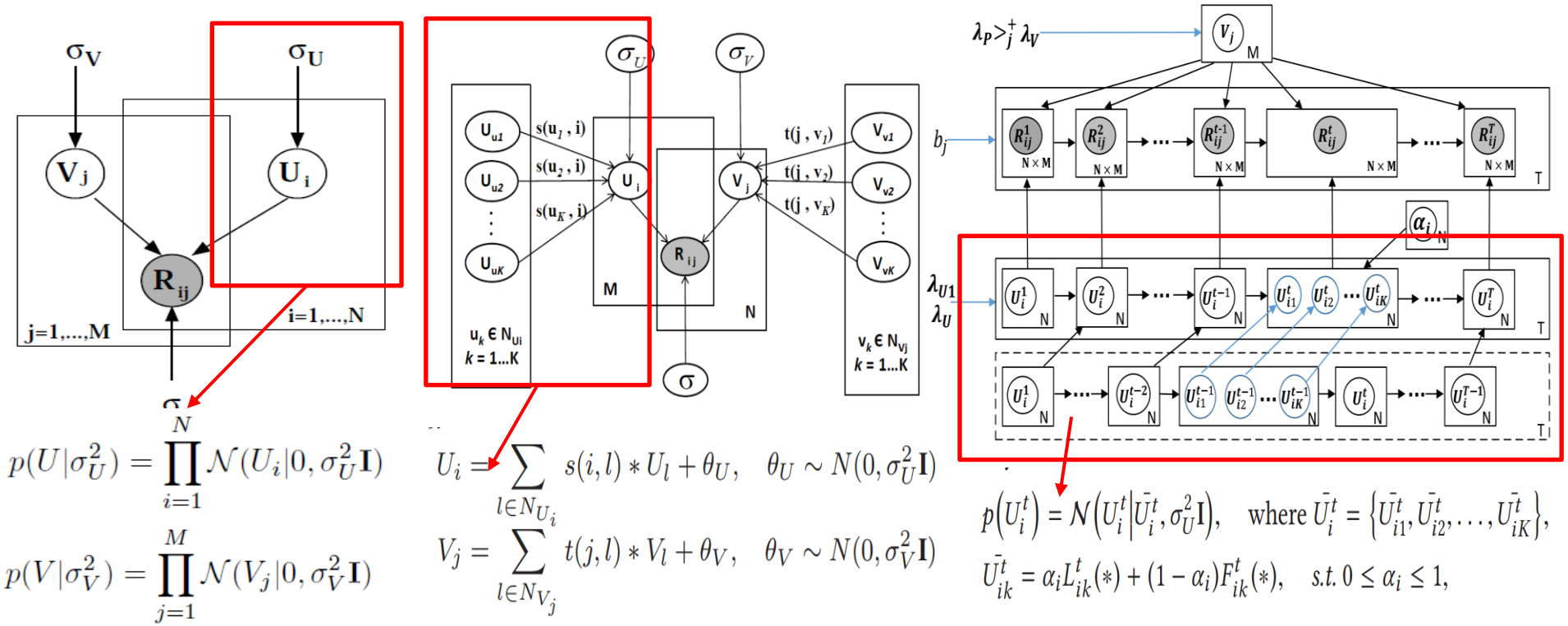


stochastic updates

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



- 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($X|\theta$)

- 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的后验概率分布**
  - MLE和MAP都是只返回了的预估值，最高值
  - **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$N(\mu,\sigma^2)$
正态分布（均值已知）	方差σ^2	倒Γ分布

Beta(α, β)

[PDF] Latent dirichlet allocation
DM Blei, AY Ng, MI Jordan - Journal of mac
… In this section we compare LDA to simple
model, a mixture of unigrams, and the pLSI
☆ 保存 引用 被引用次数：49791 相

- **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49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：

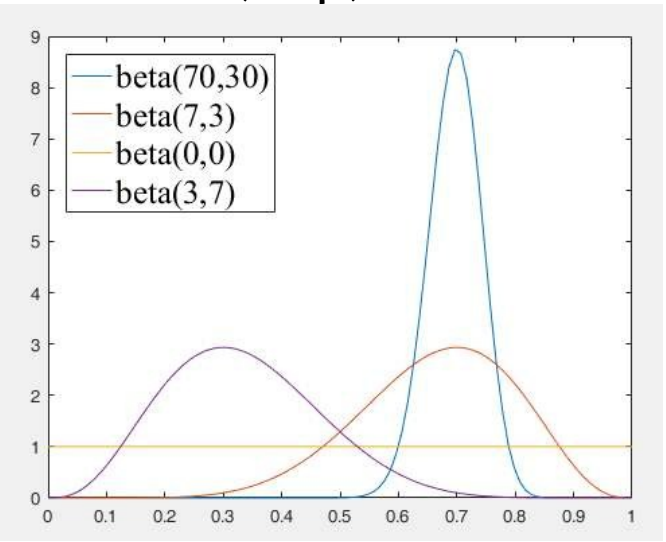
$$P(x_1 = 1) = \int_0^1 P(x_1 = 1|\theta)P(\theta)d\theta \qquad P(x_1 = 1) = \frac{\alpha}{\alpha+\beta}$$

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)

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

  - 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)

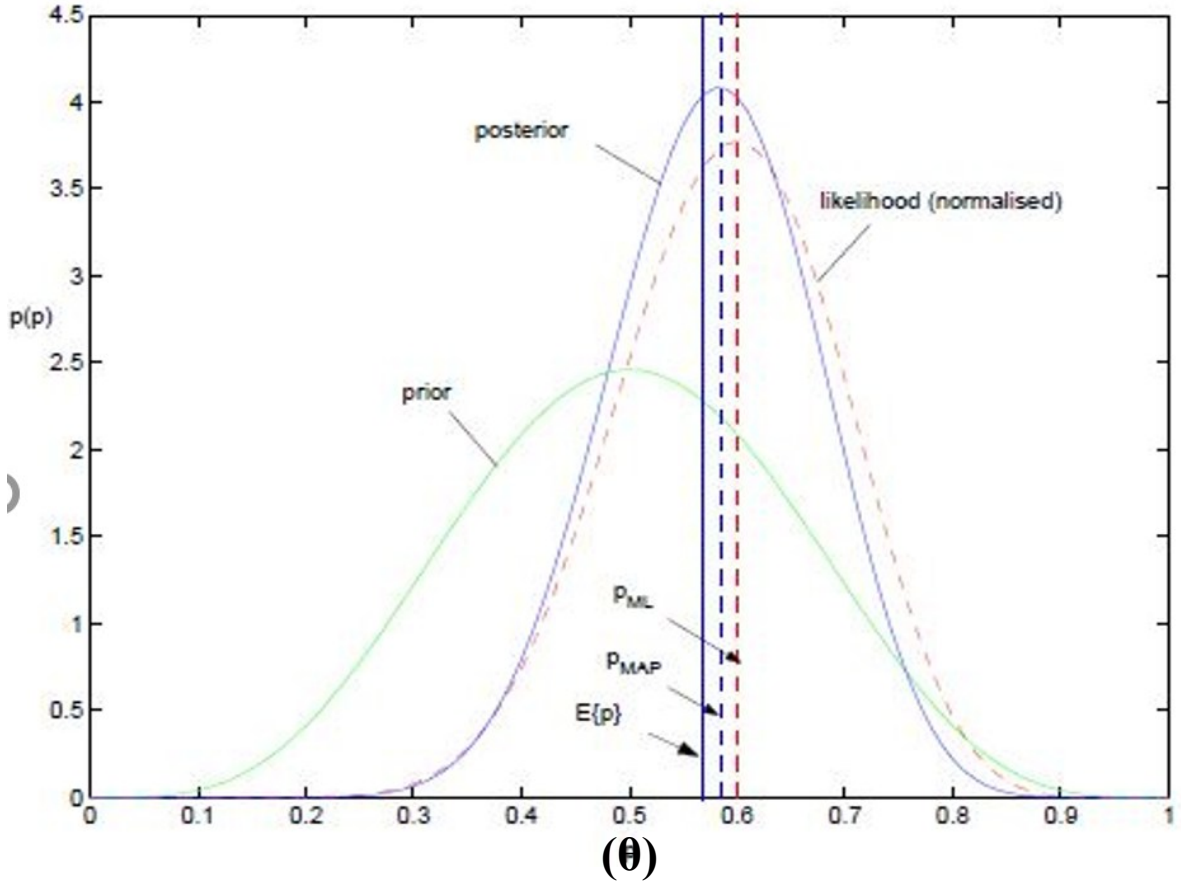
$$\frac{\alpha + M_1}{\alpha + M_1 + \beta + M_2}$$

  - 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

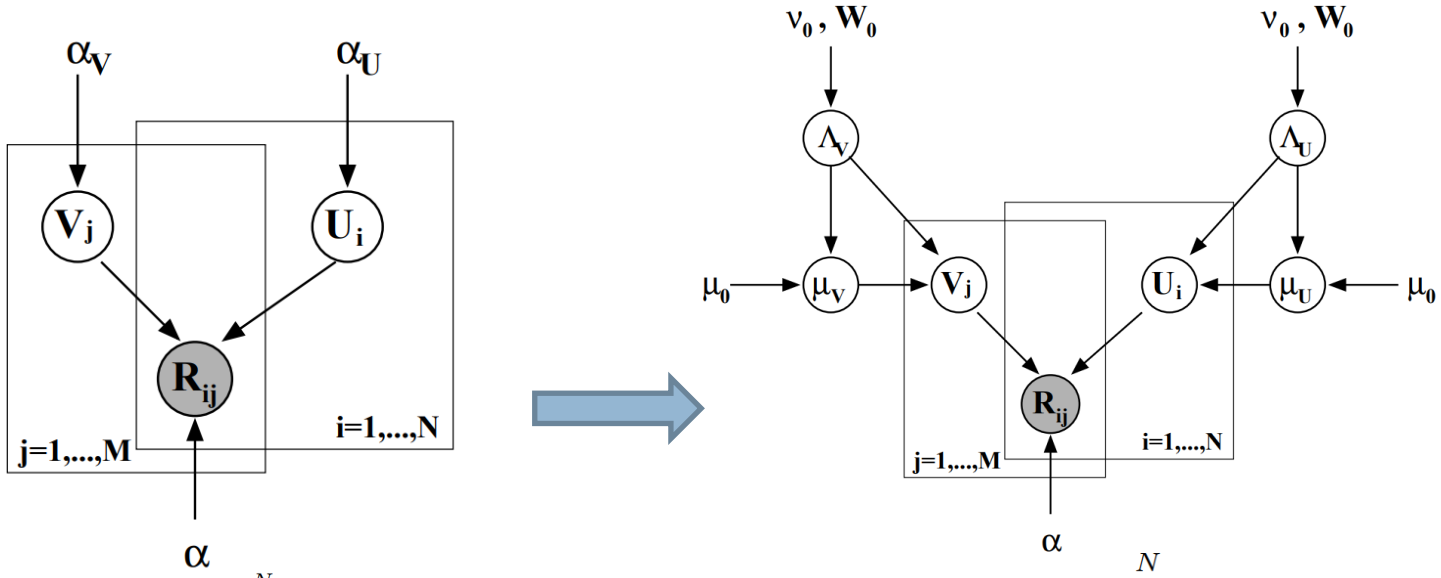
- 贝叶斯估计
  - 抛硬币例子：从理论上来说，贝叶斯估计优于MLE及MAP
  - 存在一些问题：
    - 贝叶斯估计通常需要做积分运算，复杂度较大
    - 有时对积分我们没有一个解析解
    - 有时无法为似然函数likelihood找到合适的共轭先验

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1：案子：猜嫌疑人身高。

    证据：目击者说看到的人很高，目测大概 185cm（观测数据）
  - 估计结果
    - **MLE：完全相信数据：185cm。**

      风险：目击者可能看错，容易激进
    - **MAP：** 数据：185cm，先验：这个片区人员大多 172cm 左右（先验分布）。结合该知识，嫌疑人身高约180cm（后验的概率最高值）

      风险：只报一个最可能值，不告诉你还有别的可能性。
    - **贝叶斯估计：** 分析后验的整条概率曲线。身高 178、179、180（最高点）、181、182 都有不小可能性。嫌疑人身高 95% 可信区间是 [174 cm, 184 cm]；后验均值是 179 cm。

      优势：自带不确定性，适合做风险决策

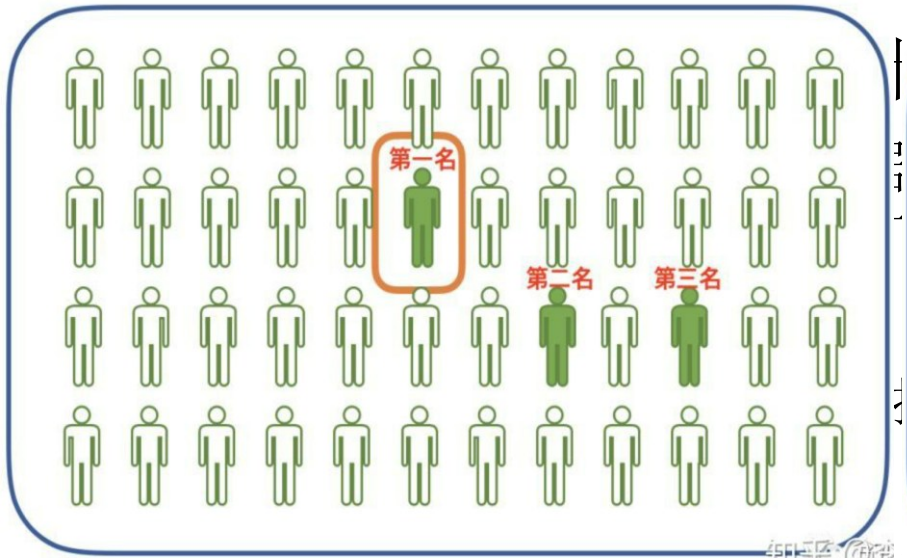
参数估计

- 应用角度理解：MLE，MAP，贝叶斯估计
  - 举例2：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。

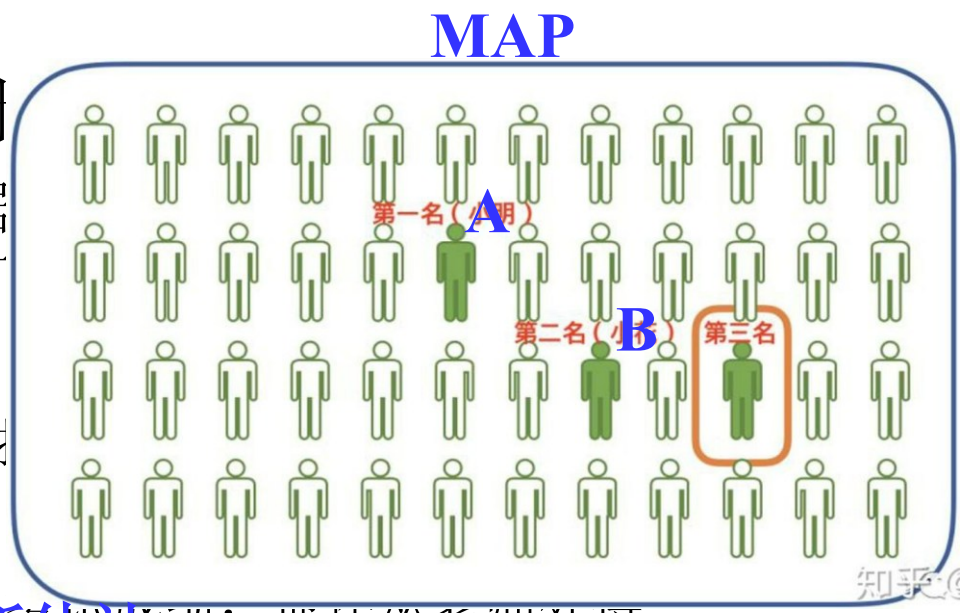
参数估计

- 应用
  - 举例
    朋友
    那么
    - M
      成
    - 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 ，小江选择学生C来解决自己
    - 贝叶斯估计：此时小 每个学生都给出一个答案，并 生一个权重（参数的分布），。

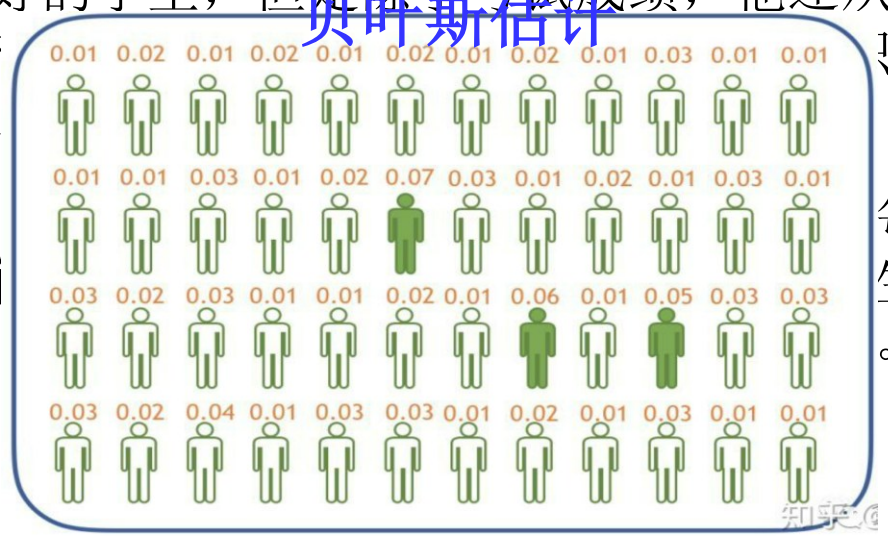
MLE



MAP



贝叶斯估计



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 $P(X|\theta)$ 大
    - 预测：$P(x^*|\hat{\theta})$
  - MAP
    - 估计：寻找 $\hat{\theta}$ 使得 $P(\theta|X)$ 大
    - 预测：$P(x^*|\hat{\theta})$

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

  - 贝叶斯估计
    - 估计：由数据估计出参数的后验 $P(\theta|X)$
    - 预测：$\int_{\Theta} P(x^*|\theta)P(\theta|X)d\theta$



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



参数估计—课后学习

- 线性回归例子：MLE，MAP，贝叶斯方法估计
  - $P(w|y) = \frac{P(y|w)P(w)}{P(y)}$
  - MLE最大化似然函数
  - MAP最大化后验函数
  - 贝叶斯估计需要直接求出整个后验分布