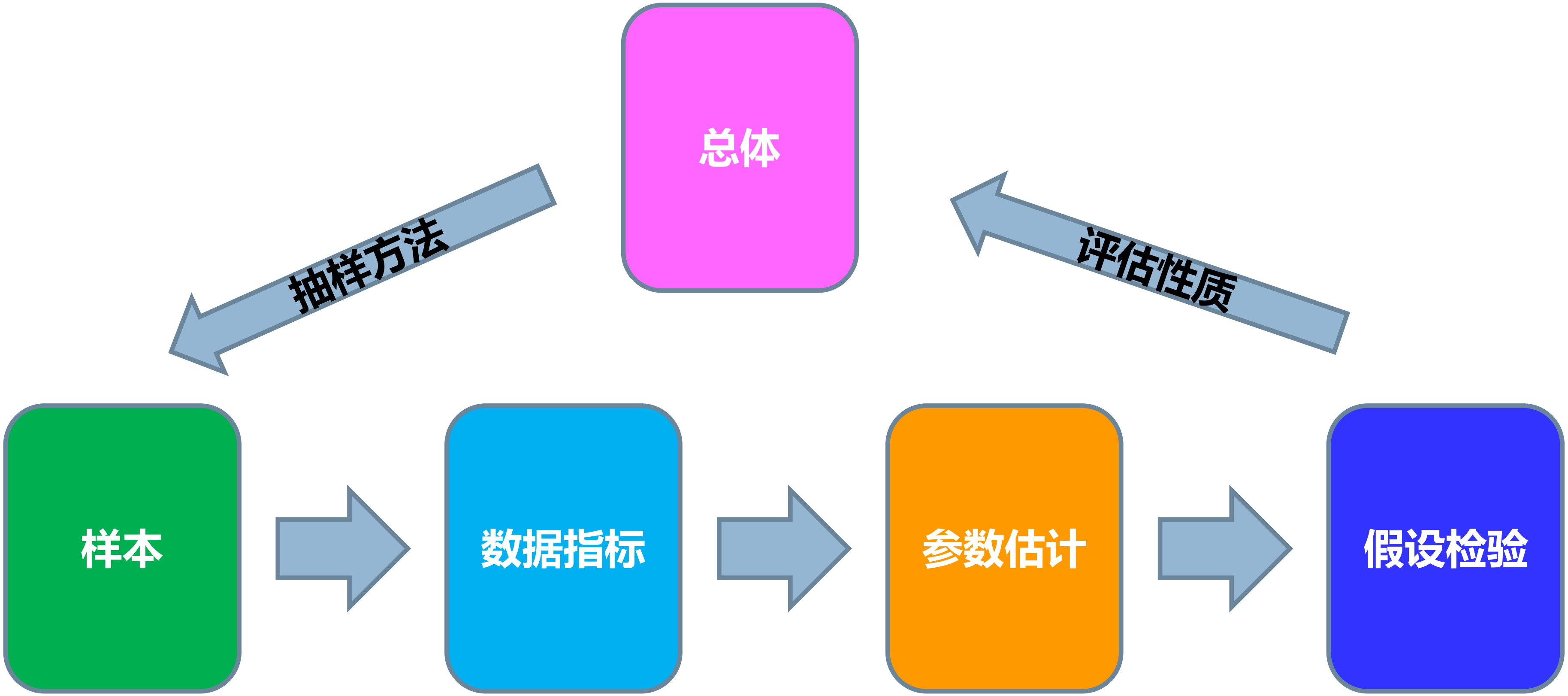
数据科学导论
Introduction to Data Science

第三章 数据统计基础

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

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html

回顾：数据统计



数据统计

- 数据分布
- 参数估计
- 假设检验
- 抽样方法

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

参数估计

- 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
    - 简单来说，直接以样本指标来估计总体指标
    - 总体的某个特征值，如数学期望、方差和相关系数等
  - 区间估计：从总体中抽取的样本，根据一定的正确度与精确度的要求，构造出适当的区间，以作为总体的分布参数(或参数的函数)的真值所在范围的估计
    - 用数轴上的一段经历或一个数据区间，表示总体参数的可能范围。这一段距离或数据区间称为区间估计的置信区间

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—矩估计

- 矩估计
  - 原理：大数定律： n趋近于无穷，样本矩趋近于总体矩
    - 矩估计是基于“替换”思想，即用样本矩估计总体矩
      - 均值，方差
  - 随机变量的矩
    - K阶原点矩：$E\left(X^k\right)=\frac{1}{n}\sum_{i=1}^{n}X_i^k$
    - K阶中心距：$E\left([X-E(X)]^k\right)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$
      - 一阶原点矩表示期望
      - 二阶中心矩表示方差
      - 三阶中心矩表示偏度
      - 四阶中心矩表示峰度

$$\mu=\frac{\sum_{i=1}^{n}x_i}{n} \qquad \sigma^2=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

$$SK=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}} \qquad K=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^2}$$

参数估计—矩估计

- 矩估计
  - 原理：大数定律： n趋近于无穷，样本矩趋近于总体矩
    - 矩估计是基于“替换”思想，即用样本矩估计总体矩
      - 均值，方差
  - 随机变量的矩
    - K阶原点矩：$E\left(X^k\right)=\frac{1}{n}\sum_{i=1}^{n}X_i^k$
    - K阶中心距：$E\left([X-E(X)]^k\right)=$
      - 一阶原点矩表示期望
      - 二阶中心矩表示方差
      - 三阶中心矩表示偏度
      - 四阶中心矩表示峰度

$$\mu=\frac{\sum_{i=1}^{n}x_i}{n} \qquad \sigma^2=\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

$$SK=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}} \qquad K=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}$$

课后练习：思考并推导矩估计与数据统计指标的关系

参数估计—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。

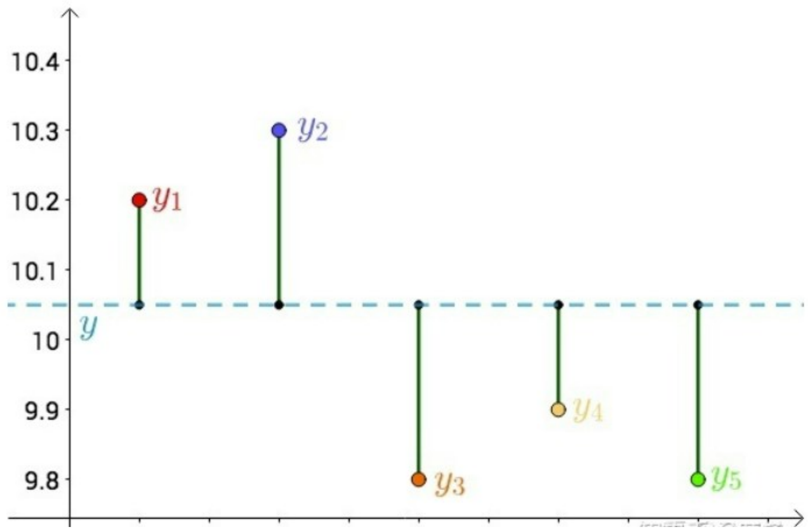
即估计结果为罐中白球所占的比例70% $= \frac{7}{10}$

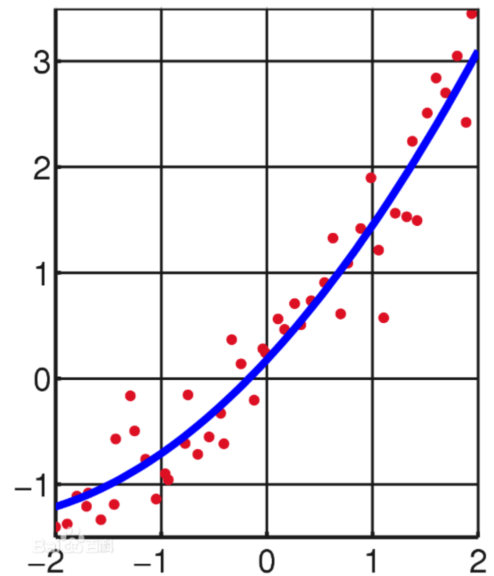
符合直观

(独立同分布，无偏估计)

参数估计—最小二乘估计

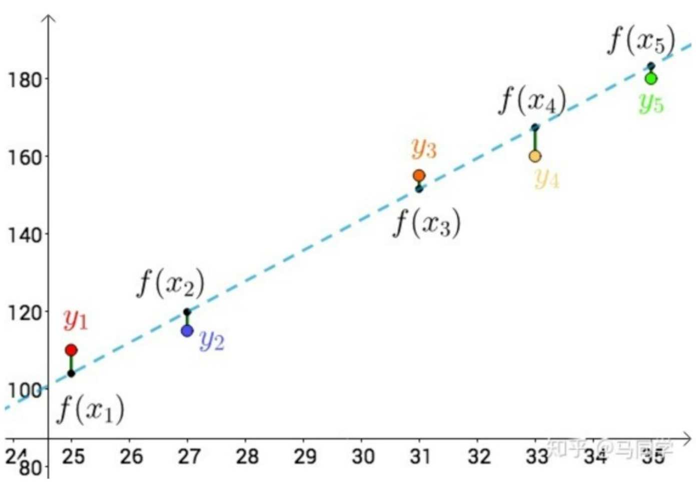
- 最小二乘估计(Least Square Estimate, LSE)
  - 总体的模型：用样本数据拟合总体的参数估计量，即估计值与观测值之差的平方和最小
  - 目标：最小化估计值θ与观测值$\hat{\theta}$之差的平方和
  - $\min L(\theta) = \sum_{i=1}^{N}(\theta - \hat{\theta}_i)^2$





参数估计—最小二乘估计

- 最小二乘估计(LSE)
  - 常用于线性回归分析做参数估计
  - 给定数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，假设模型$f(X|\theta)$
  - 例:在线性回归模型$f(X|\theta)=\theta_0+\theta_1x+\theta_1x^2+\cdots+\theta_nx^n$ $= \boldsymbol{\theta}^{T}\boldsymbol{X}$
  - 目标：$\min L(\theta)=\sum_{i=1}^{N}(f(X|\theta)-Y)^2$
  - 求解：一阶导数为0：$\frac{\partial L(\theta)}{\partial\theta_0}$=0，$\frac{\partial L(\theta)}{\partial\theta_1}$=0，$\cdots$，$\frac{\partial L(\theta)}{\partial\theta_n}$=0



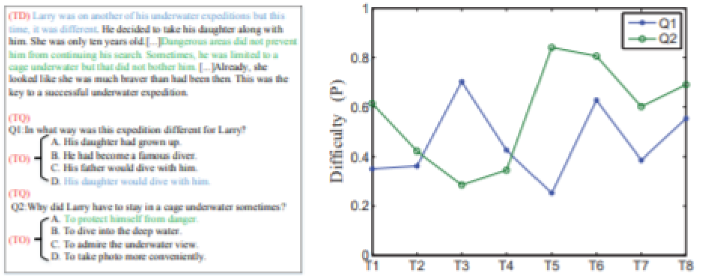
课后学习：最小二乘矩阵求解方法

参数估计—最小二乘估计

- 最小二乘估计—建模案例

Question Difficulty Prediction for READING Problems in Standard Tests

$$\mathcal{J}(\Theta) = \sum_{Q_i} (P_i - \mathcal{M}(Q_i))^2 + \lambda_{\Theta} ||\Theta_{\mathcal{M}}||^2, \qquad (5)$$

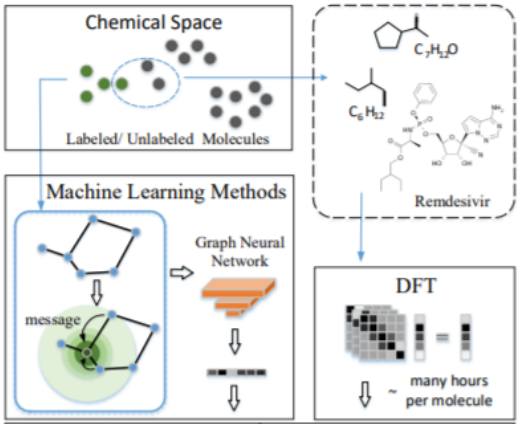


(a) A READING problem (b) Difficulties in tests

ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction

Traditionally, MPGNN is trained in a supervised manner where all the labels are given and we usually use mean square loss (MSE) between predictions and labels $\boldsymbol{y}_i$ (i.e. the labeled properties in $\mathcal{D}_l$) to guide the optimization of the model parameters:

$$\mathcal{L}_p = \sum_{i=1}^{N_l} ||\boldsymbol{y}_i - f_{\theta}(\boldsymbol{z}_{G_i})||^2. \qquad (4)$$



Property	Value
HOMO	-0.38eV
LUMO	0.08eV
U	-196.43eV

✓ Zhenya Huang, Enhong Chen, Question Difficulty Prediction for READING Problems in Standard Tests, AAAI2017

✓ Zhongkai Hao, Zhenya Huang,, Qi Liu, Enhong Chen, ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction, KDD 2020

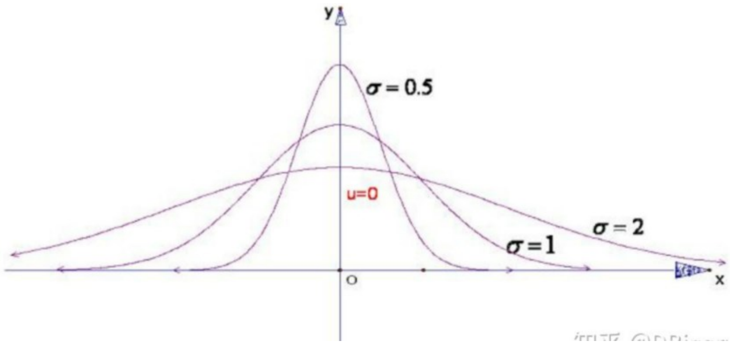
参数估计

- 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θ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： $L(x_1, x_2, \cdots, x_n|\theta) = \boldsymbol{C_n^k \theta^k (1-\theta)^{n-k}}$
  - 对数似然函数： $\ln L(X|\theta) = \ln \boldsymbol{C_n^k} + \mathbf{k} \ln\boldsymbol{\theta} + (\boldsymbol{n}-\boldsymbol{k})\ln(\mathbf{1}-\boldsymbol{\theta})$
  - 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—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
  - $\mathrm{L}(X|\theta)$ 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**

正 反 + $\boldsymbol{\theta}$ MAP

k次 5次　N-k次 5次　预先知识

	正	反
$\boldsymbol{\theta}$	0.5	0.5

- **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：

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

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

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

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

- 因果概率

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

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**

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(θ|X)是后验概率，估计的目标：已知数据X，求参数θ的值
    - p(X|θ)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|θ)最大
    - MAP的目标是：求参数θ，使得似然函数p(X|θ) 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$\theta$

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

  - 整理MAP的优化目标为

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

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$\boldsymbol{p(\theta)}$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 → 硬币可能是不均匀的
    - 期末考试：通常认为学霸分数高 → 学霸当天发挥不好
    - 导师批评：通常认为学生犯了错误 → 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(θ)
  - 扔硬币的例子：10次实验，其中**正面朝上(参数：$\boldsymbol{\theta}$)**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$\boldsymbol{\theta}$**=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($\theta|\mu,\sigma$)

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

  - 求解MAP $\mathrm{p}(\theta|\boldsymbol{X})\propto\mathrm{p}(\boldsymbol{X}|\theta)\mathrm{p}(\theta)=\theta^7(1-\theta)^3\times\frac{1}{10\sqrt{2\pi}}e^{-50(\theta-0.5)^2}$
  - 取对数：$\boldsymbol{ln}\,\mathrm{p}(\theta|\boldsymbol{X})\propto 7\boldsymbol{ln}(\theta)+3\boldsymbol{ln}(1-\theta)+\boldsymbol{ln}\left(\frac{1}{10\sqrt{2\pi}}\right)-50(\theta-0.5)^2$
  - 求导解得：$\boxed{\hat{\theta}\approx 0.558}$
  - 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

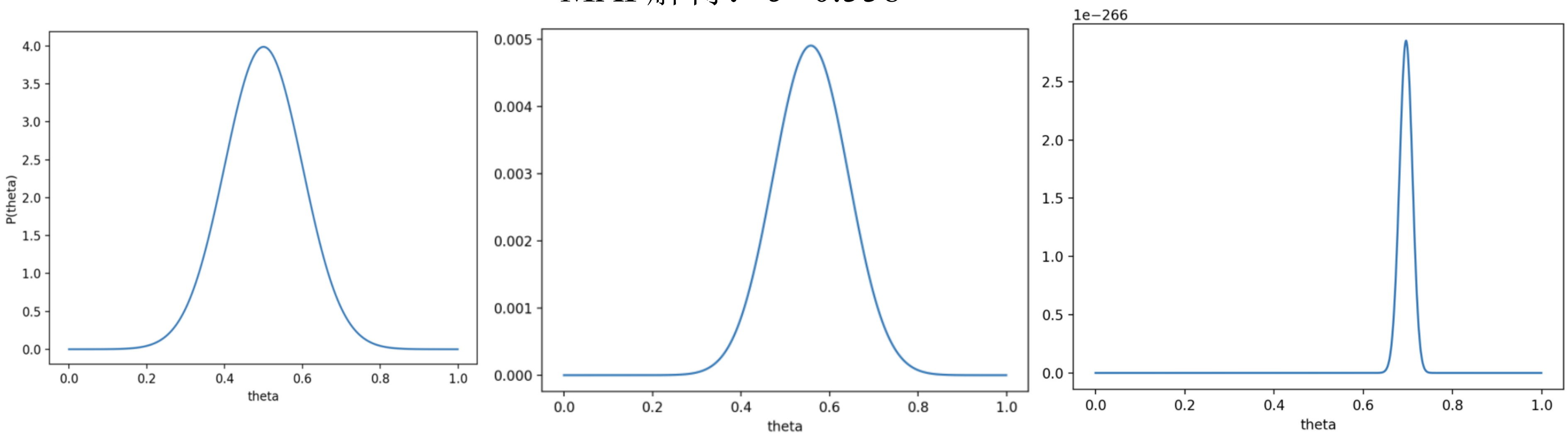
参数估计—最大后验估计

- 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

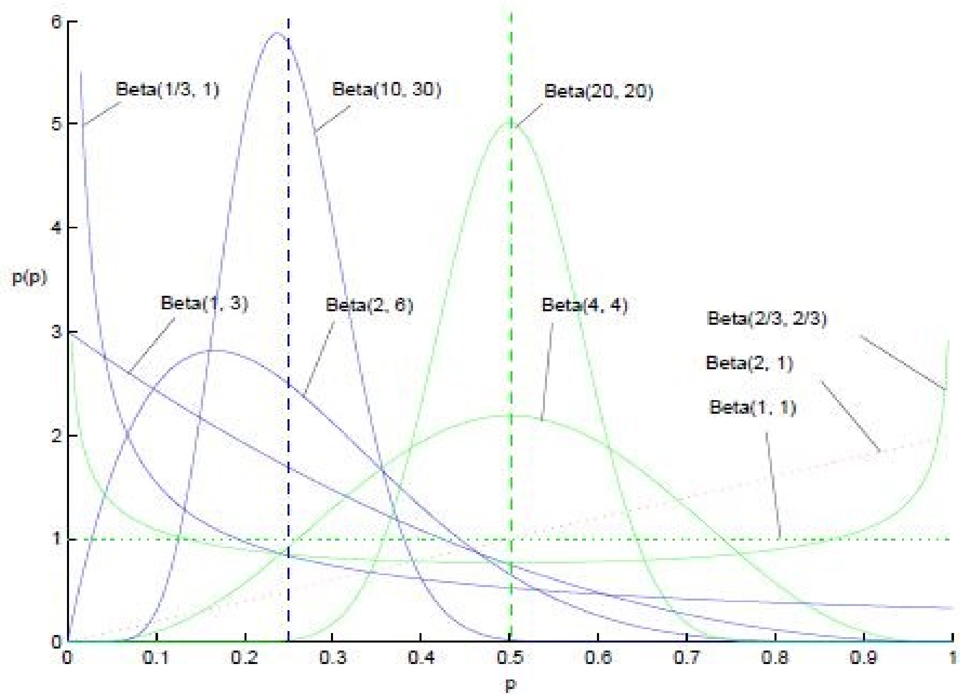


Fig. 1. Density functions of the beta distribution with different symmetric and asymmetric parametrisations.

下的

参数估计—最大后验估计

正　反

M1次 M2次

- 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 $\boldsymbol{\alpha} = \boldsymbol{\beta} = 5$，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

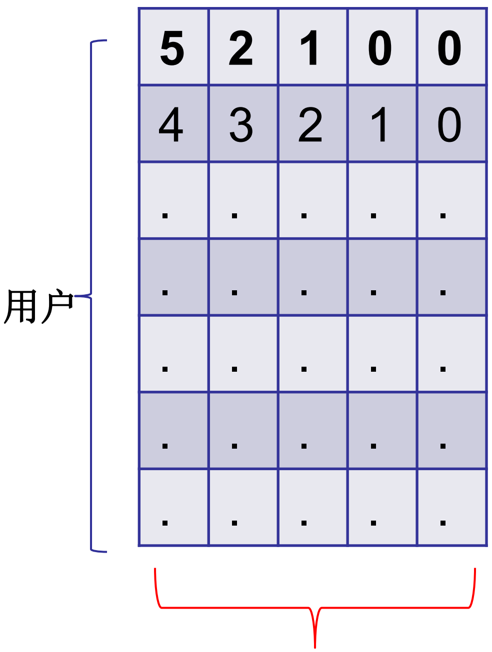
$$\hat{\boldsymbol{\theta}}_{MAP} = \frac{\boldsymbol{M_1 + \alpha - 1}}{\boldsymbol{M_1 + M_2 + \alpha + \beta - 2}} = \frac{\boldsymbol{M_1 + 4}}{\boldsymbol{M_1 + M_2 + 8}}$$

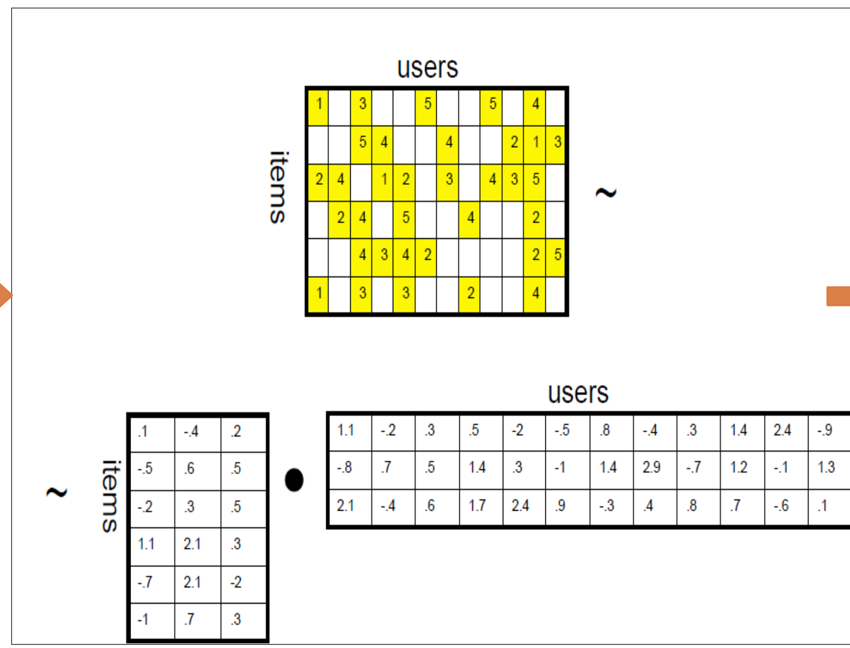
  - 与MLE相比，结果中多了α-1和$\alpha+\beta$-2，即先验作用，且超参数越大，为了改变先验分布传递的belief所需要的观察值就越多
  - 同样表明“**硬币一般是均匀的**”这一先验对参数估计的影响

思考：如果先验P(θ=0.5)=1？

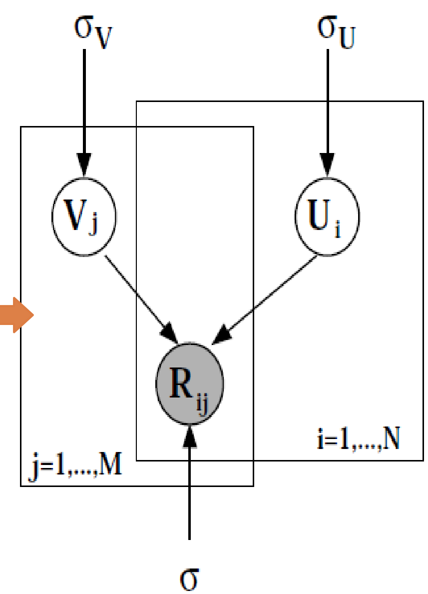
参数估计—最大后验估计

- 基于模型的协同过滤—概率矩阵分解
  - 面向评分预测的模型





基本分解思想



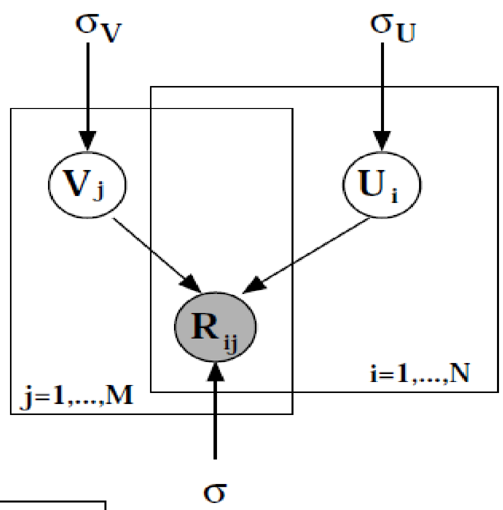
PMF（概率矩阵分解）

➢ 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： $p(U,V\,|\,R,\sigma^2,\sigma_U^2,\sigma_V^2)$

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



    - MLE似然

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

    - 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.

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

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

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

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

1)Intialize U ,V with small, random values

2)repeat

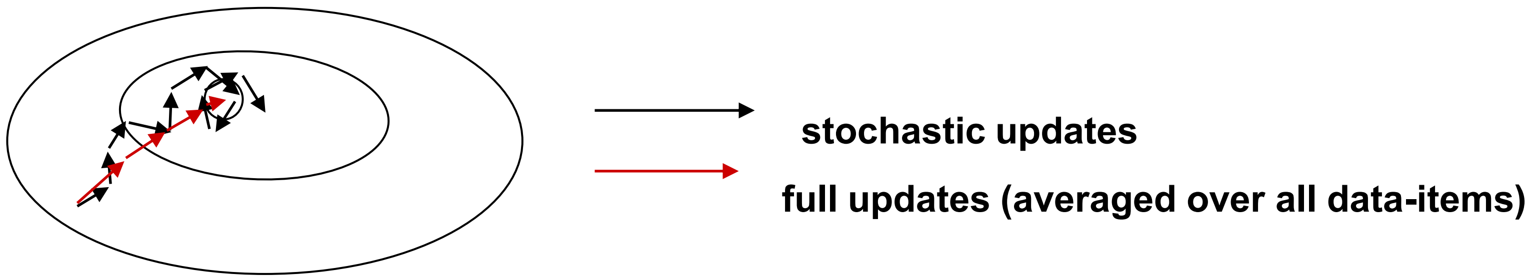
for each record in the training data

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

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

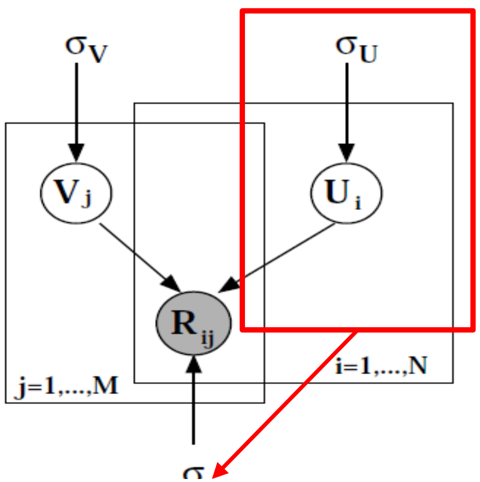
until convergence

优化方法：随机梯度下降(SGD)



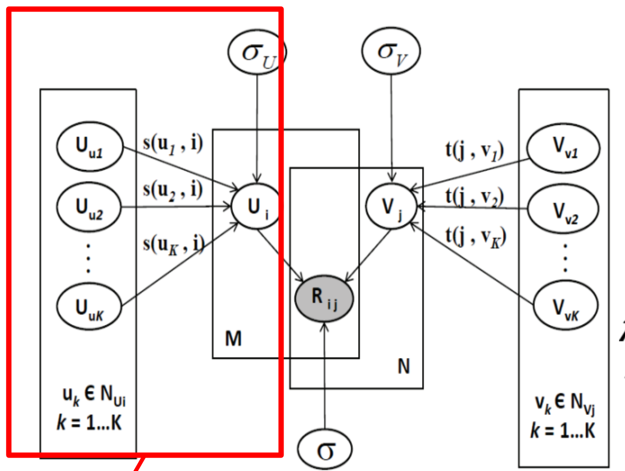
参数估计—最大后验估计

- 课后学习：概率矩阵分解



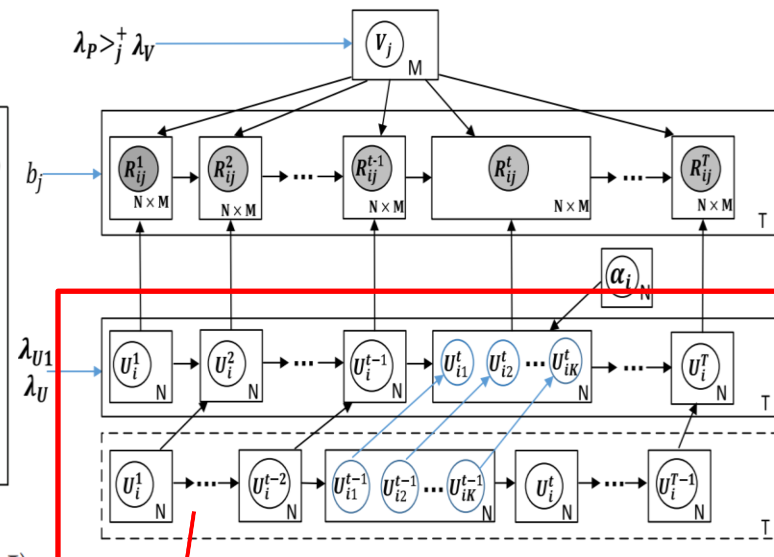
$$p(U|\sigma_U^2) = \prod_{i=1}^{N} \mathcal{N}(U_i|0, \sigma_U^2\mathbf{I})$$

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



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

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



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

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

- **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**