



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

1

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

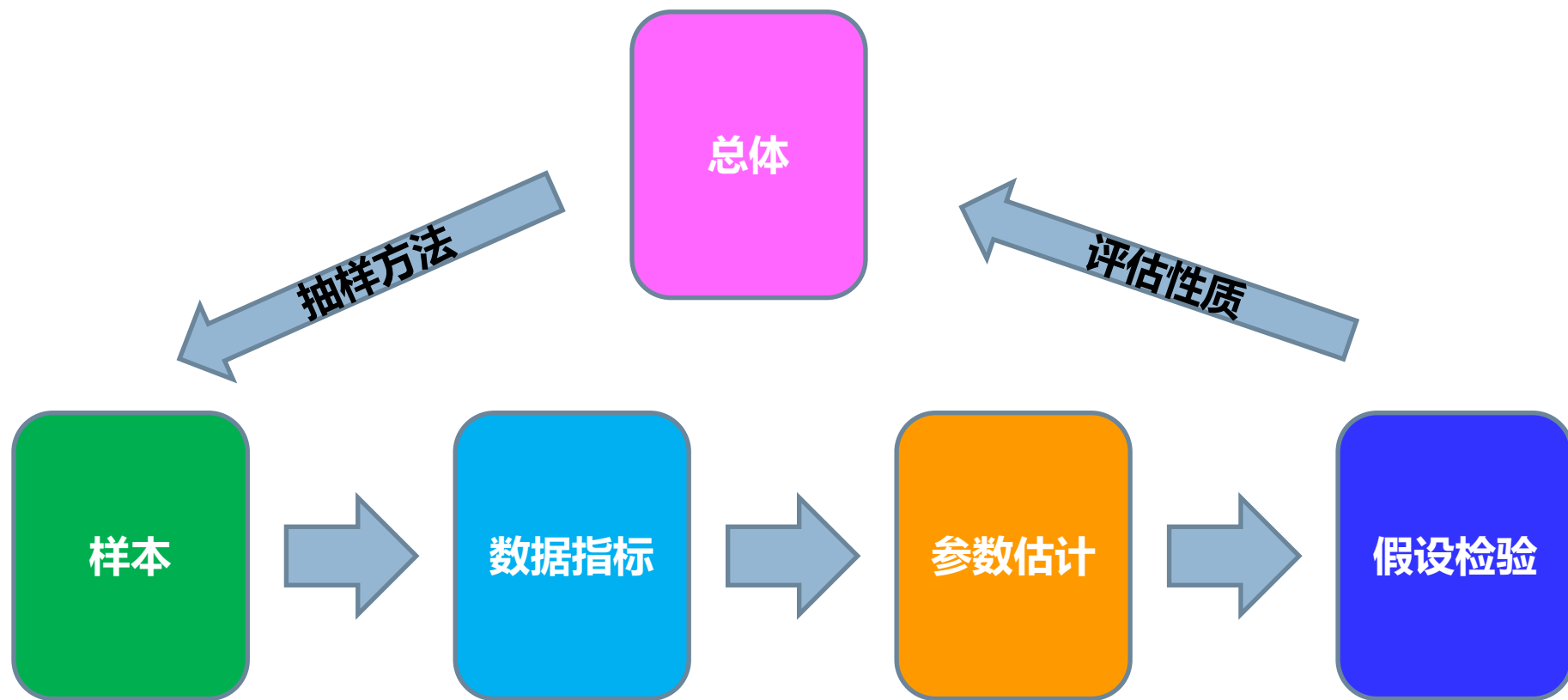
课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html>



回顾：数据统计

2





参数估计—最大后验估计

26

□ 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 θ ,
- n 次实验, 共 k 次正面向上, 目标为估计参数 θ



□ 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

总体参数



k 次
5次

$N-k$ 次
5次

MLE



	正	反
θ	0.5	0.5

□ 频率学派

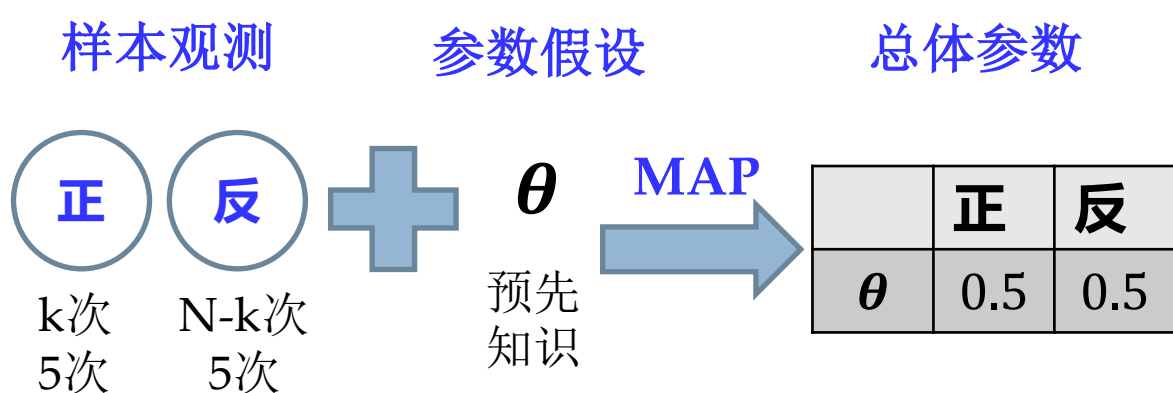
- 完全相信数据
- 世界是确定的
- 事件在多次重复实验中趋于稳定



参数估计—最大后验估计

27

- 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。



□ 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

28

□ 贝叶斯公式:

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

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

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

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

□ 因果概率

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



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

29

□ 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：

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

注：患有 meningitis 的后验概率 仍然非常小



参数估计—最大后验估计

30



□ 最大后验概率估计(MAP)

□ 已知： $x_1, x_2, x_3, \dots, x_N$ 为样本，问：估计总体的参数 θ

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

- $p(\theta|X)$ 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
- $p(\theta)$ 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 $p(X|\theta) p(\theta)$ 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

31

□ 最大后验概率估计(MAP)

□ 已知： $x_1, x_2, x_3, \dots, x_N$ 为样本，问：估计总体的参数 θ

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

□ 整理MAP的优化目标为

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

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

32

□ 最大后验概率估计(MAP)—理解先验 $p(\theta)$

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

- 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

33

□ 最大后验概率估计(MAP)—理解先验 $p(\theta)$

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

□ **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

34

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 $p(\theta|X) \propto p(X|\theta)p(\theta) = \theta^7(1-\theta)^3 \times \frac{1}{10\sqrt{2\pi}} e^{-50(\theta-0.5)^2}$

□ 取对数： $\ln p(\theta|X) \propto 7\ln(\theta) + 3\ln(1-\theta) + \ln\left(\frac{1}{10\sqrt{2\pi}}\right) - 50(\theta-0.5)^2$

□ 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



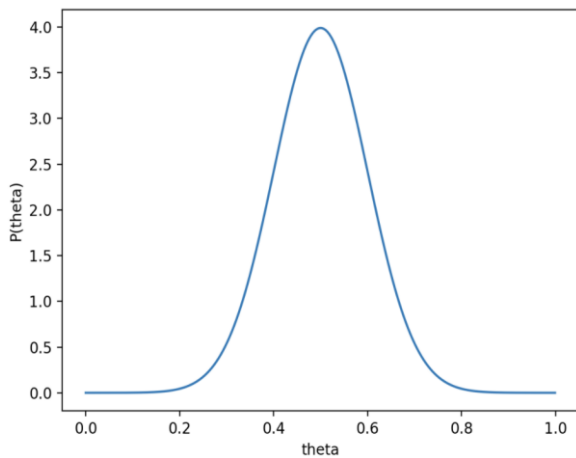
参数估计—最大后验估计

35

最大后验概率估计(MAP) — 理解先验 $p(\theta)$

扔硬币的例子：

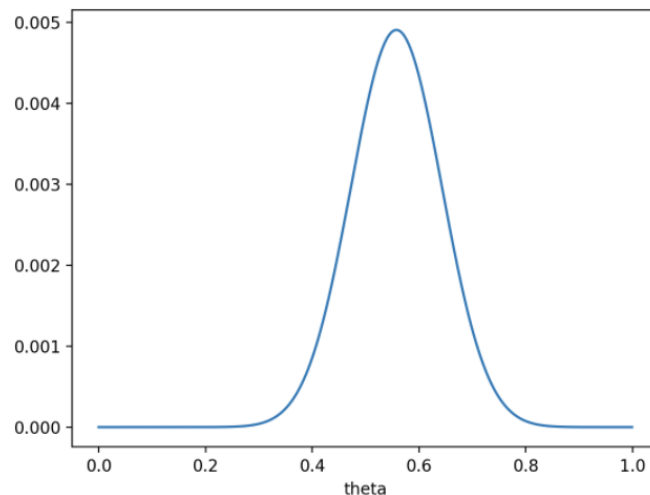
先验 $p(\theta|\mu, \sigma)$
均值0.5，方差0.1



10次实验，其中正面朝上
(θ)7次，反面朝上3次

MLE解得： $\theta=0.7$

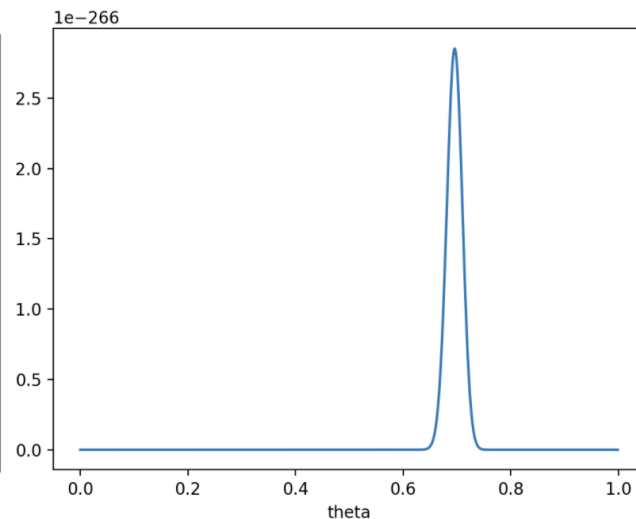
MAP解得： $\theta=0.558$



1000次实验，其中正面朝上
(θ)700次，反面朝上300次

MLE解得： $\theta=0.7$

MAP解得： $\theta=0.696$



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

36

□ 最大后验概率估计(MAP)—理解先验 $p(\theta)$

- 扔硬币的例子：我们期望先验概率（待估计的参数 θ ）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：

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

- 其中，Beta函数是 $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$
- 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

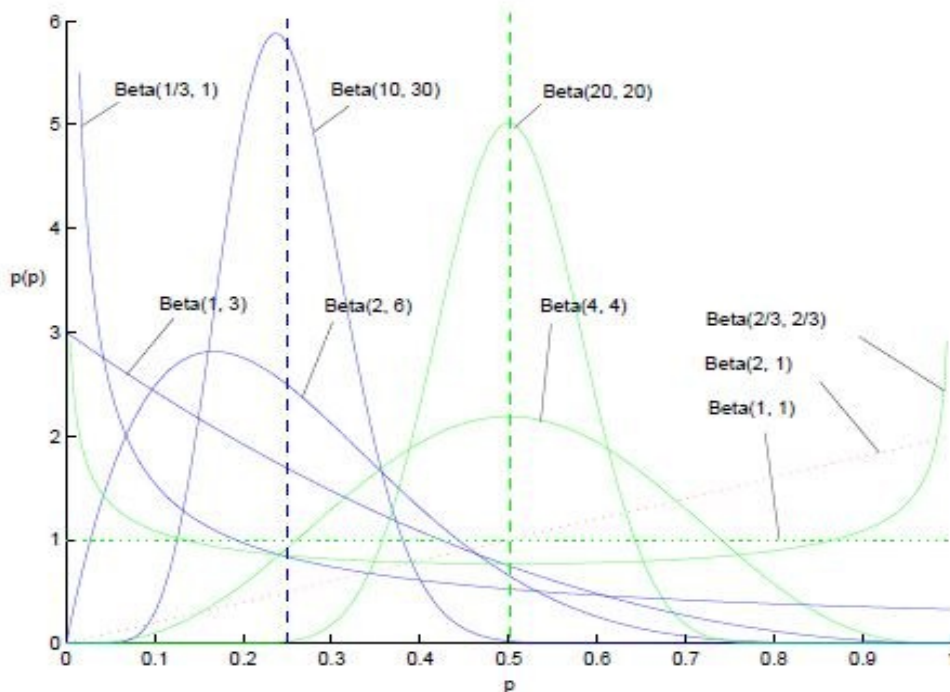


参数估计——最大后验估计

37

最大后验概率估计(MAP)—理解先验 $p(\theta)$

在0.5



以下的

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



参数估计—最大后验估计

正

反

M1次 M2次

38

□ 最大后验概率估计(MAP)—理解先验 $p(\theta)$

- 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：

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

- 取 $\alpha = \beta = 5$ ，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 $\alpha-1$ 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

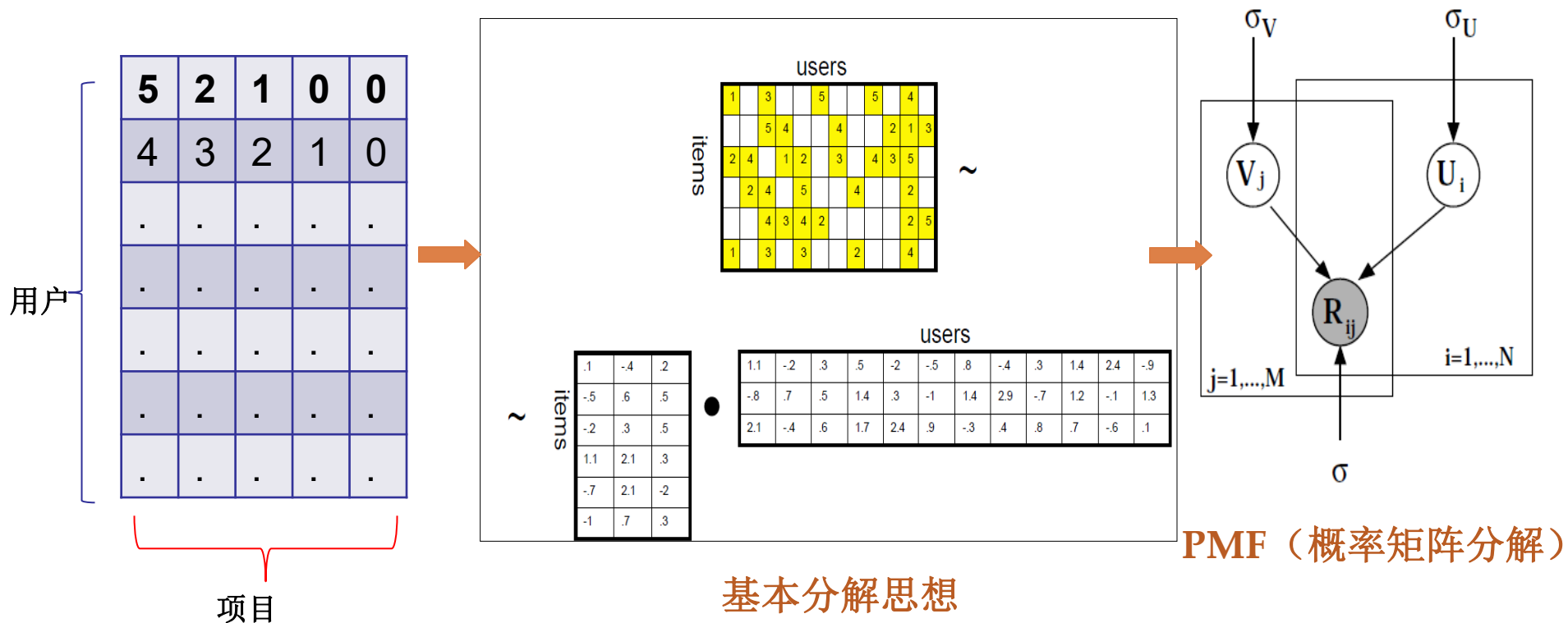
思考：如果先验 $P(\theta=0.5)=1$ ？



参数估计—最大后验估计

39

- 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

40

基于模型的协同过滤—概率矩阵分解

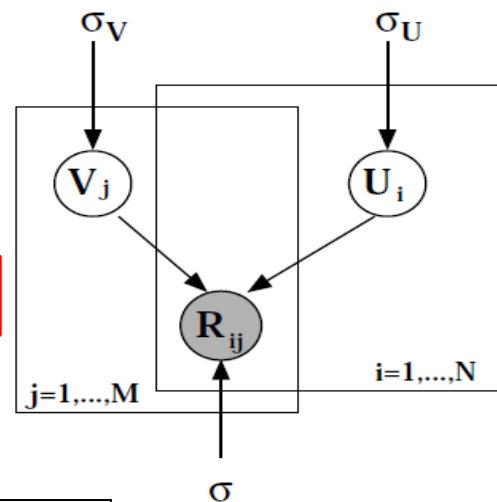
最大后验概率方法估计参数U和V (θ)

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

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

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

41

□ 基于模型的协同过滤—概率矩阵分解

□ MAP learning

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

□ 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

42

□ 基于模型的协同过滤—概率矩阵分解

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

1) Initialize U, V with small, random values

2) repeat

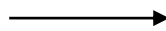
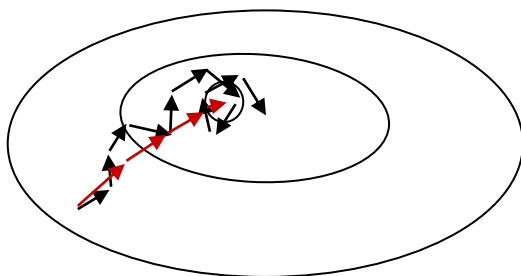
for each record in the training data

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

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

优化方法：随机梯度下降(SGD)

until convergence



stochastic updates



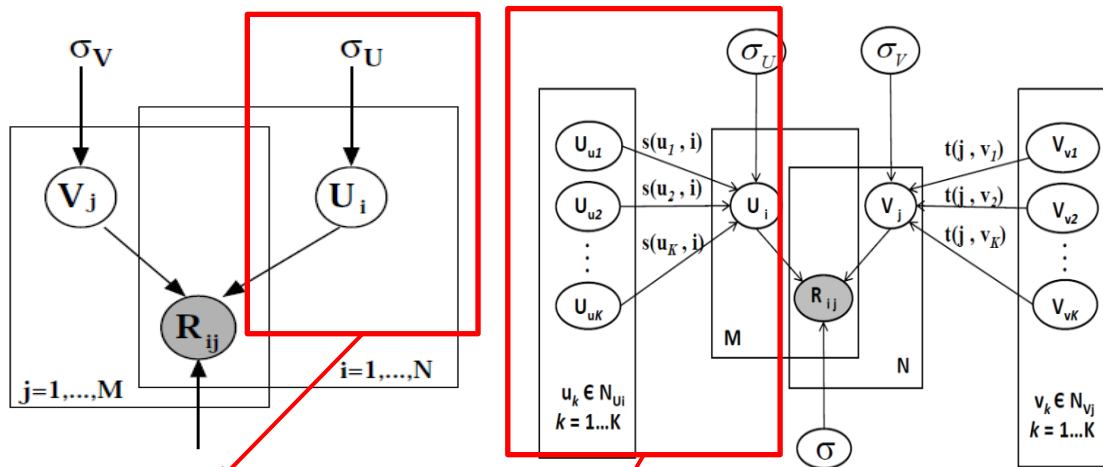
full updates (averaged over all data-items)



参数估计—最大后验估计

43

□ 课后学习：概率矩阵分解

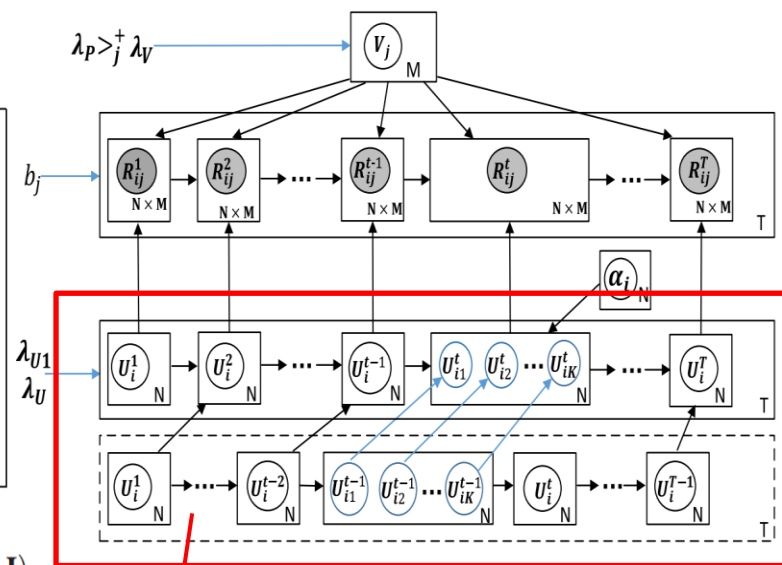


$$p(U|\sigma_U^2) = \prod_{i=1}^N \mathcal{N}(U_i|0, \sigma_U^2 \mathbf{I})$$

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

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

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



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

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

- 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

44

- 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”



参数估计—贝叶斯估计



45

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

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

- 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

46

□ 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） $P(X, \theta)$ ，易得：

$$P(X, \theta) = P(X|\theta)P(\theta)$$

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：

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

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

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

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

一般来说，计算积分是不可能的



参数估计—贝叶斯估计

47



□ 贝叶斯估计

$$P(\theta) \sim \text{Beta}(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

- 下面仍然以抛硬币为例，此时选择Beta分布作为先验，类似MAP：

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

- 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 \mathbf{X}_1 ，那么贝叶斯方法预测该硬币朝上的概率为：

$$\begin{aligned} P(x_1 = 1) &= \int_0^1 P(x_1 = 1 | \theta) P(\theta) d\theta \\ &= \int_0^1 \theta P(\theta) d\theta = \int_0^1 \theta \gamma \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta \end{aligned}$$

- 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

48

□ 贝叶斯估计

$$P(\theta) \sim \text{Beta}(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

- 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：

$$\begin{aligned} P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta} = \frac{\theta^{M_1}(1-\theta)^{M_2} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}}{\int \theta^{M_1}(1-\theta)^{M_2} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} d\theta} \\ &= \frac{\theta^{M_1+\alpha-1}(1-\theta)^{M_2+\beta-1}}{\int \theta^{M_1+\alpha-1}(1-\theta)^{M_2+\beta-1} d\theta} = \text{Beta}(\theta|\alpha + M_1, \beta + M_2) \end{aligned}$$

- 观察：在抛硬币的实验中(似然 $p(X|\theta)$ 为二项分布), 当先验 $P(\theta)$ 为Beta分布时，后验 $P(\theta|X)$ 也为Beta分布，即更新后的参数服从一个新的Beta($\alpha+M_1, \beta+M_2$)分布
这种情况我们称之为Beta分布是二项分布似然 $P(X|\theta)$ 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 $P(X|\theta)$ 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 $\beta(\alpha, \beta)$
泊松分布	均值 λ	Γ 分布 $\Gamma(\alpha, \beta)$
指数分布	均值的倒数 λ	Γ 分布 $\Gamma(\alpha, \beta)$
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

50

□ 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 $\text{Beta}(\alpha+M_1, \beta+M_2)$ ，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

$$P(x^*|X) = \int P(x^*|\theta)P(\theta|X)d\theta \quad \leftarrow p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

□ 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46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：

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



参数估计—贝叶斯估计

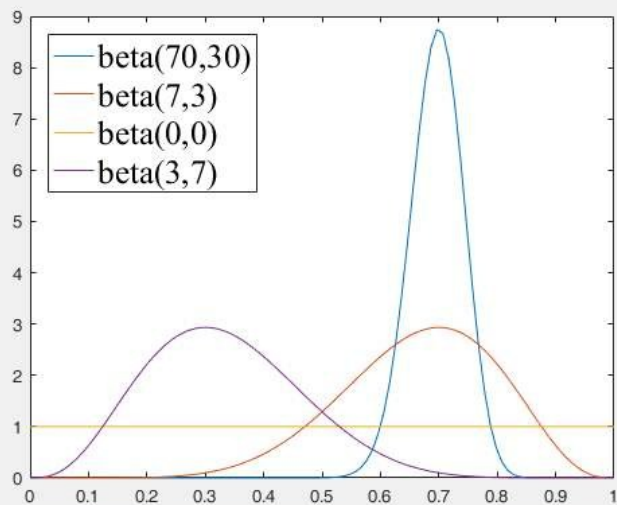
51

□ 贝叶斯估计

- 此时，参数 θ 的取值（期望）就是这个新的Beta分布 $\text{Beta}(\alpha+M_1, \beta+M_2)$ 的均值 (M_1 次正面， M_2 次反面)

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

- $\text{Beta}(\alpha, \beta)$ 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

52

□ 贝叶斯估计

- 此时，参数 θ 的取值（期望）就是这个新的Beta分布 $\text{Beta}(\alpha+M_1, \beta+M_2)$ 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 $\text{Beta}(14+5, 6+5)$ 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）

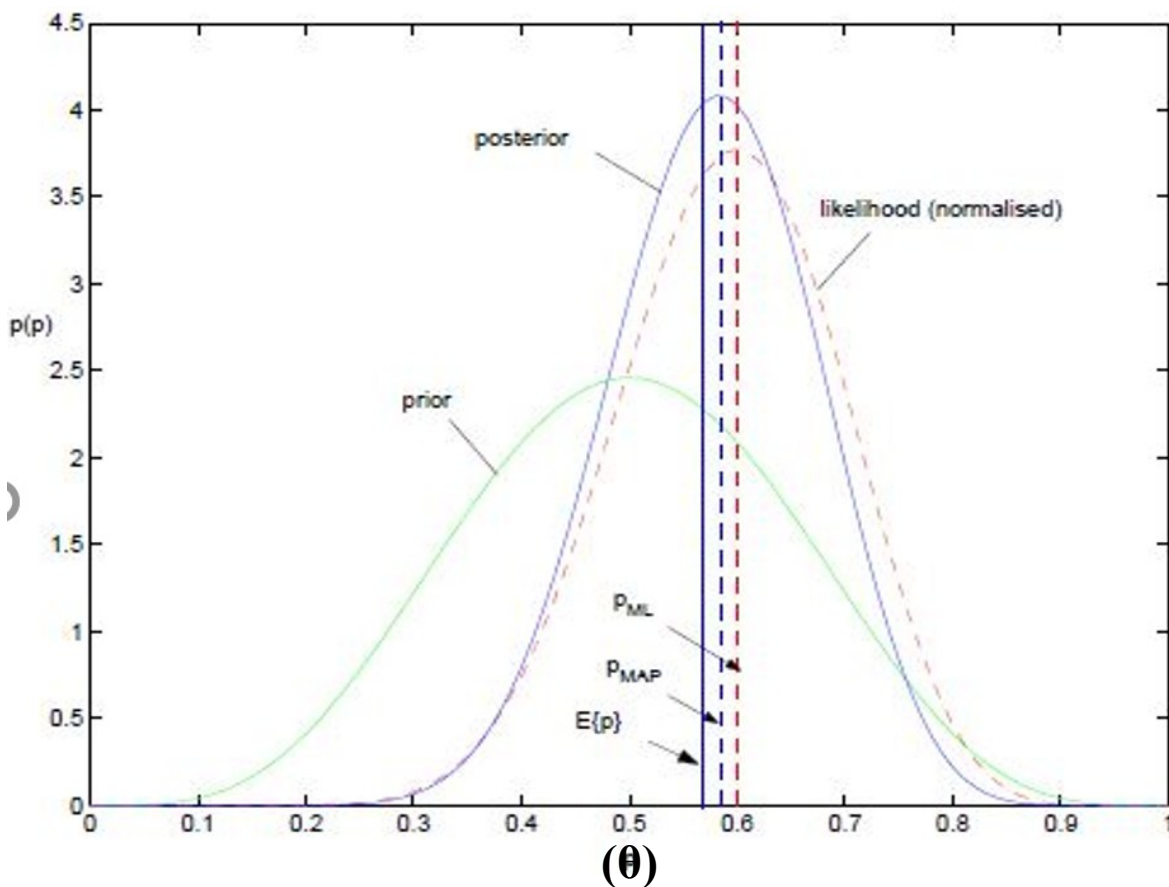


参数估计

53

□ 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

54

□ 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

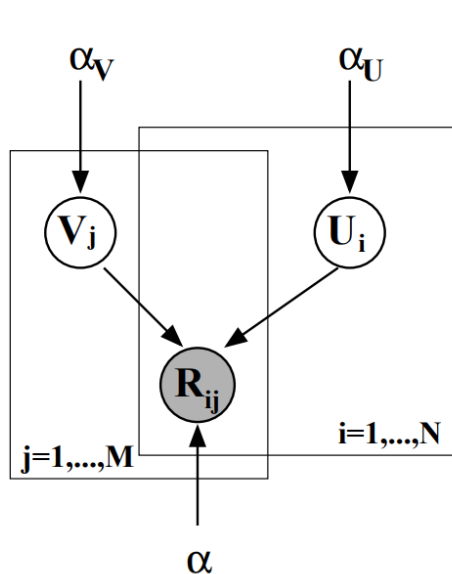
<http://www.mcmchandbook.net/HandbookChapter1.pdf>



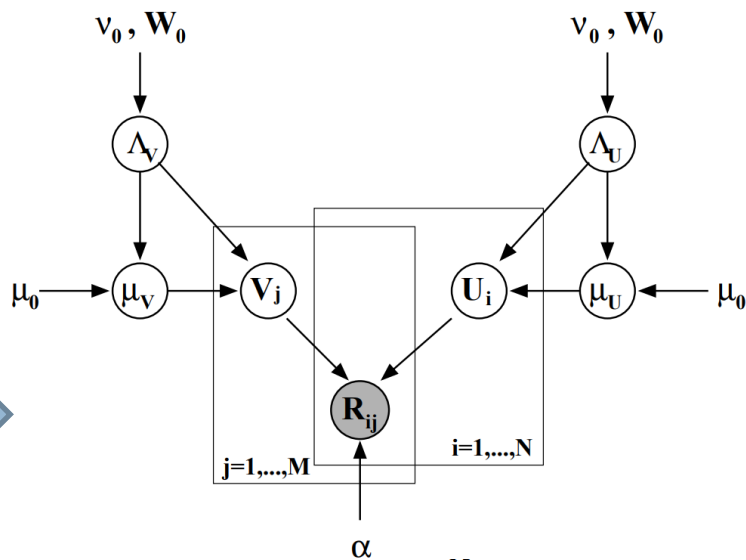
参数估计—贝叶斯估计

55

贝叶斯估计—贝叶斯概率矩阵分解BPMF



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



$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^N \mathcal{N}(U_i|\mu_U, \Lambda_U^{-1}),$$
$$p(V|\mu_V, \Lambda_V) = \prod_{j=1}^M \mathcal{N}(V_j|\mu_V, \Lambda_V^{-1}).$$

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

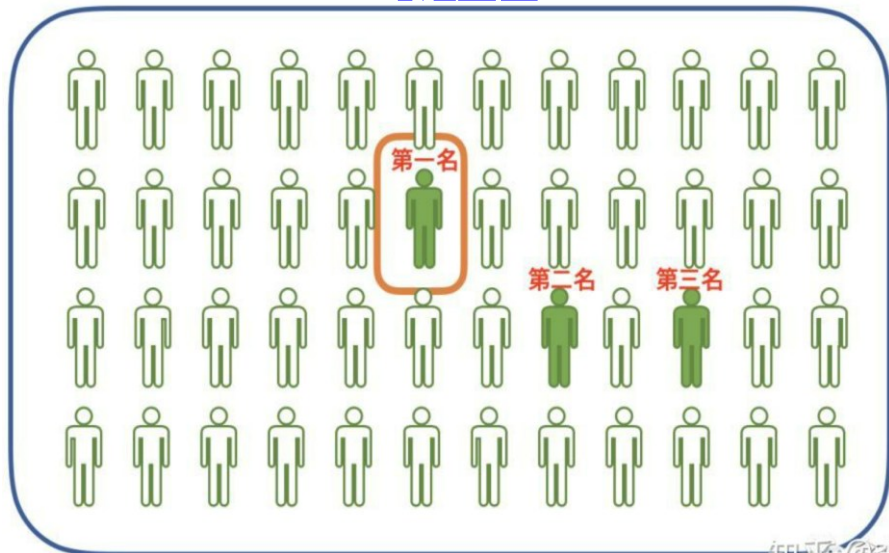
56

□ 应用角度理解：MLE，MAP，贝叶斯估计

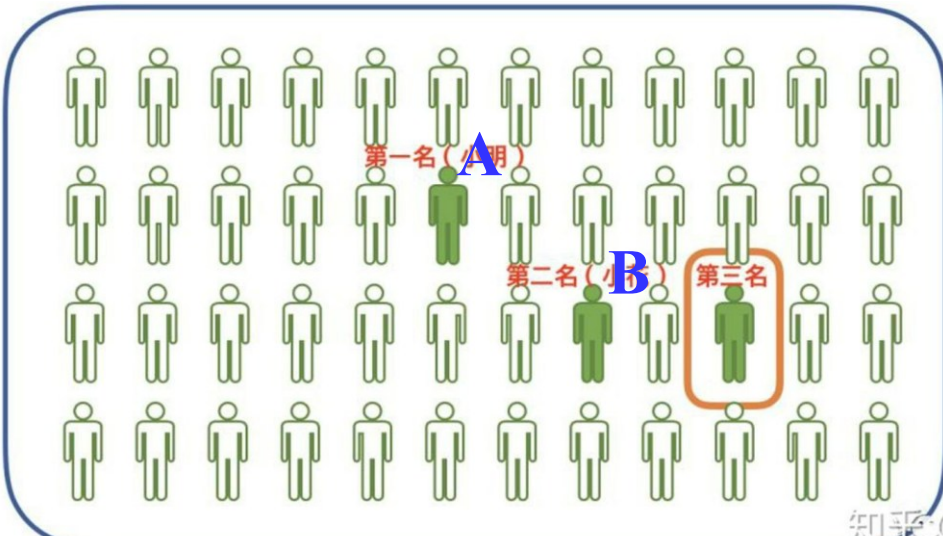
- 举例：假设小江遇到一个计算机难题（数据的预测），碰巧小江有个朋友在大学计算机系当老师，于是他打算找该老师的学生帮忙，那么他该如何寻求帮助呢？
 - **MLE:** 由以往的考试成绩（对应已有数据）排序（A,B,C....）,选出成绩最好的学生A（对应模型中的参数）来解决自己的问题
 - **MAP:** 仍然选择最好的学生，但是除了考试成绩，他还从老师处得知A，B两人考试中有作弊嫌疑（对应先验），结合该知识，小江选择学生C来解决自己的问题
 - **贝叶斯估计:** 此时小江不再寻求单个人的帮助，他会要求每个学生都给出一个答案，并结合考试成绩和老师的提醒给每个学生一个权重（参数的分布），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



MLE



MAP



■ **MAP:** 仍然选择取好的子集，但实际与测试成绩，他选择老师更付
知A, B两 0.01 0.02 0.01 0.02 0.01 0.02 0.01 0.02 0.01 0.03 0.01 0.01 结合该知识，小江选

- 贝叶斯估计都给出多重（参数



他会要求每个学生
给每个学生一个权
最后的解答。



参数估计

58

□ 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: $P(x^*|\hat{\theta})$

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

□ 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

59

- 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

60

□ 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

61

- 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

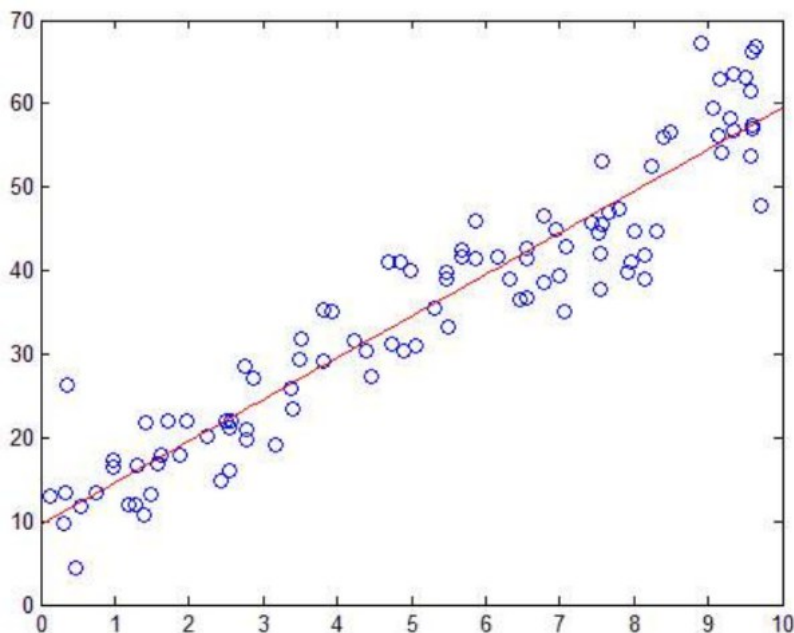


参数估计

62

□ 线性回归：MLE，MAP，贝叶斯估计

- 现有一个数据集，包含 N 个一维样本点，将其记作 $X = (x_1, x_2, \dots, x_N)^T$ ，同时有 N 个一维的标签 $Y = (y_1, y_2, \dots, y_N)^T$ ， $x_i, y_i \in R$ 。
- 求一个参数为 w 的线性模型 $y = wx$ 暂不考虑偏置



最小二乘法损失函数：

$L(w) = \sum_{i=1}^N (y_i - wx_i)^2$ ，
使 $L(w)$ 的值最小的 w^* 便为
所求参数



参数估计

63

□ 求解线性回归: 极大似然估计(MLE)

- 从概率的角度思考最小二乘法, 概率的本质是不确定性
- 引入这种不确定性, 在线性模型 $y = wx$ 中加入一个服从高斯分布的噪声 $\varepsilon \sim N(0, \sigma^2)$ 。即, 我们要求的模型是 $y = wx + \varepsilon$

□ 有 $y|w, x \sim N(wx, \sigma^2)$ \longrightarrow $P(y|w, x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-wx)^2}{2\sigma^2}}$

- 极大似然估计(MLE): 极大似然估计的基本思路
 - 求解的参数是固定的, 在给定数据集 X 的情况下使得标签 Y 出现的概率最大, 即似然函数最大
 - 求解过程的推导如下



参数估计

64

□ 求解线性回归: 极大似然估计(MLE)

$$\begin{aligned}w^* &= \operatorname{argmax}_w P(Y|w, X) \\&= \operatorname{argmax}_w \prod_{i=1}^N P(y_i|w, x_i) \\&= \operatorname{argmax}_w \log(\prod_{i=1}^N P(y_i|w, x_i)) \\&= \operatorname{argmax}_w \log(\sum_{i=1}^N P(y_i|w, x_i)) \\&= \operatorname{argmax}_w \sum_{i=1}^N -\log(\sqrt{2\pi}\sigma) - \frac{(y_i - wx_i)^2}{2\sigma^2} \\&= \operatorname{argmin}_w \sum_{i=1}^N \frac{(y_i - wx_i)^2}{2\sigma^2} \quad \longleftrightarrow \quad \text{最小二乘法的损失函数!}\end{aligned}$$

所求的参数应该使得似然函数最大, 之后的推导和化简可以看出, 最大化似然函数(MLE)的结果等价于最小二乘法的损失函数!



参数估计

65

- 求解线性回归: 极大似然估计(MLE)
 - 使用MLE的方法估计参数和使用LSE的损失函数是等价的
 - 有条件的: 我们假设加入模型的噪声是服从高斯分布的。即模型的概率密度函数为高斯分布, 则LSE与MLE等价

课后思考: 假如加入模型的噪声 ε 不再服从高斯分布, 那么极大似然法最后得到的结果是否和最小二乘法的损失函数一致?





参数估计

66

□ 求解线性回归: 最大后验估计(MAP)

□ 认为参数 w 不再是固定的, 而是服从一个分布的, 是具有不确定性的

□ 则目标函数: $P(w|y) = \frac{P(y|w)P(w)}{P(y)}$ \longrightarrow $P(w|y) \propto P(y|w)P(w)$
后验 似然 先验

□ 加入噪声引入不确定性, 则

$$y|w, x \sim N(wx, \sigma^2) \longleftrightarrow P(y|w, x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-wx)^2}{2\sigma^2}}$$

□ 上式似然函数服从高斯分布

■ 由于高斯分布是自共轭的, 可假设参数 w 服从高斯分布先验: $w \sim N(0, \sigma_0^2)$

注: 与贝叶斯方法不同, 在MAP中, 一般不严格要求 w 服从共轭分布。



参数估计

67

□ 求解线性回归: MAP的推导

$$w^* = \operatorname{argmax}_w P(w|Y)$$

$$\propto \operatorname{argmax}_w P(w)P(Y|w)$$

$$= \operatorname{argmax}_w P(w) \prod_{i=1}^N P(y_i|w, x_i)$$

$$= \operatorname{argmax}_w \log(P(w) \prod_{i=1}^N P(y_i|w, x_i))$$

$$= \operatorname{argmax}_w \log P(w) + \log(\prod_{i=1}^N P(y_i|w, x_i))$$

$$= \operatorname{argmin}_w \frac{w^2}{\sigma_0^2} + \frac{\sum_{i=1}^N (y_i - wx_i)^2}{\sigma^2}$$

$$= \operatorname{argmin}_w \boxed{\sum_{i=1}^N (y_i - wx_i)^2} + \boxed{\frac{\sigma^2}{\sigma_0^2} w^2} \quad (\sigma \neq 0)$$



最小二乘法的损失函数 正则化项

- 最大化后验概率等价于最小二乘法加上一个额外的项
- **机器学习中:** 等价于最小二乘法的损失函数加上L2的正则化项
- 可以认为 $\frac{\sigma^2}{\sigma_0^2}$ 的值很小时, 等价于最小二乘法。



参数估计

68

□ 求解线性回归: 贝叶斯估计

□ MAP: 最大化后验概率

□ 贝叶斯方法: 求解参数 w 的后验分布 $P(w|y) = \frac{P(y|w)P(w)}{P(y)}$

□ 假设: 先验及似然函数都是高斯分布, 则, 后验分布也是一个高斯分布 (共轭性质)

■ 假设 $y|w, x \sim N(wx, \sigma^2)$, 似然函数 $P(y|w, x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-wx)^2}{2\sigma^2}}$

■ 假设先验分布 $w \sim N(0, \sigma_0^2)$, $P(w) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{w^2}{2\sigma_0^2}}$

■ 根据高斯分布的自共轭的性质, 后验分布 $P(w|Y)$ 也是一个高斯分布 $N(\mu_w, \sigma_w^2)$ 。

$$P(w|Y) \propto \exp\left\{-\frac{1}{2\sigma_w^2} w^2 + \frac{\mu_w}{\sigma_w^2} w + \mu_w^2\right\}$$



参数估计

69

□ 求解线性回归: 贝叶斯估计

□ 假设先验及似然函数都是高斯分布, 则后验分布也是一个高斯分布

■ 则假设似然函数 $P(y|w, x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-wx)^2}{2\sigma^2}}$

■ 先验分布 $P(w) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{w^2}{2\sigma_0^2}}$

■ 根据高斯分布的自共轭的性质, 后验分布 $P(w|Y)$ 也是一个高斯分布 $N(\mu_w, \sigma_w^2)$ 。

$$P(w|Y) \propto \exp\left\{-\frac{1}{2\sigma_w^2} w^2 + \frac{\mu_w}{\sigma_w^2} w + \mu_w^2\right\}$$

后验 $P(w|Y) \sim N(\mu_w, \sigma_w^2)$ \longrightarrow $P(w|Y) \propto \exp\left\{-\frac{(w-\mu_w)^2}{2\sigma_w^2}\right\}$

先验 $P(w) \sim N(0, \sigma_0^2)$ \longrightarrow $P(w) \propto \exp\left\{-\frac{w^2}{2\sigma_0^2}\right\}$

似然 $Y|w \sim \prod_{i=1}^N y_i|w$ \longrightarrow $P(Y|w) \propto \prod_{i=1}^N \exp\left\{-\frac{(y_i-wx_i)^2}{2\sigma^2}\right\}$



参数估计

70

□ 求解线性回归: 贝叶斯估计

- 由于已知后验分布 $P(w|Y)$ 为高斯分布, 计算可以使用比较法, 只需要保证指数项一致即可:

后验
$$P(w|Y) \propto \exp\left\{-\frac{1}{2\sigma_w^2} w^2 + \frac{\mu_w}{\sigma_w^2} w + \mu_w^2\right\}$$

似然*先验
$$P(w)P(Y|w) \propto \exp\left\{-\frac{w^2}{2\sigma_0^2}\right\} \prod_{i=1}^N \exp\left\{-\frac{(y_i - wx_i)^2}{2\sigma^2}\right\}$$
$$\propto \exp\left\{\left(-\frac{X^T X}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right) w^2 + \frac{1}{\sigma^2} Y^T X w - \frac{1}{2\sigma^2} Y^T Y\right\}$$

- 根据剩下公式中 w^2 以及 w 的对应系数相等

- 计算出 $\mu_w = \frac{X^T Y \sigma_0^2}{\sigma^2 + X^T X \sigma_0^2}$, $\sigma_w^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + X^T X \sigma_0^2}$, 即为后验分布的参数



小插曲

71

□ Logistic 回归

- Logistic回归模型依据sigmoid函数 $\sigma(z) = \frac{1}{1+e^{-z}}$ 对样本点属于什么类别做出估计。
- 现有一个数据集，包含 N 个一维样本点，将其记作 $X = (x_1, x_2, \dots, x_N)^T$ ，同时有 N 个一维的标签 $Y = (y_1, y_2, \dots, y_N)^T$ ， $x_i, y_i \in R$
- $P(y = 1|x) = \frac{1}{1+e^{-wx}} = p_1, P(y = 0|x) = \frac{e^{-wx}}{1+e^{-wx}} = p_0$ ，由上两式可得 $P(y|x) = p_1^y \cdot p_0^{1-y}$
- 再根据极大似然法对参数 w 进行估计，其推导如下：



小插曲

72

□ 求解Logistic回归: MLE的推导

$$\begin{aligned}w^* &= \operatorname{argmax}_w P(Y|w, X) \\&= \operatorname{argmax}_w \prod_{i=1}^N P(y_i|w, x_i) \\&= \operatorname{argmax}_w \log(\prod_{i=1}^N P(y_i|w, x_i)) \\&= \operatorname{argmax}_w \log(\sum_{i=1}^N P(y_i|w, x_i)) \\&= \operatorname{argmax}_w \sum_{i=1}^N \log(p_1^{y_i} \cdot p_0^{1-y_i}) \\&= \operatorname{argmin}_w \boxed{\sum_{i=1}^N y_i \log(p_1) + (1 - y_i) \log(p_0)}\end{aligned}$$



交叉熵损失函数

神经网络中常见目标函数

线性回归和logistic回归的区别:

- logistic回归模型本质上是分类模型
- 线性回归无法处理非线性分类问题
- 线性回归有闭式解, 但logistic回归没有