第3章
概率密度估计—参数法(第1讲)
Estimation on PDF: Parameter Estimation

张 燕 明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

多模态人工智能系统实验室　中科院自动化所

助教：　杨　奇　(yangqi2021@ia.ac.cn)

　　　　张　涛　(zhangtao2021@ia.ac.cn)

University of Chinese Academy of Sciences

内容提要

- 上讲内容回顾
- 基本概念
- 最大似然估计
- 贝叶斯估计
  - 正态分布下的贝叶斯估计
  - 贝叶斯学习
  - 贝叶斯估计：一个例子
- 特征维数问题

上次课主要内容回顾

- 贝叶斯决策
  - 已知类条件概率密度$p(\mathbf{x}|\omega_i)$ 和类先验分布$P(\omega_i)$，计算样本$\mathbf{x}$的类后验分布

$$p(\omega_i \mid \mathbf{x}) = \frac{p(\omega_i, \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \omega_i) P(\omega_i)}{p(\mathbf{x})}$$

  - 得到类后验概率，如何对$\mathbf{x}$分类？

上次课主要内容回顾

- 贝叶斯决策
  - 决策损失： $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ 　　把 j 类样本决策为 i 类的损失
  - 条件风险： $R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$ 把样本$\mathbf{x}$决策为 i 类的期望损失
  - 最小风险决策：

$$\arg\min_i R(\alpha_i \mid \mathbf{x})$$

  - 最小分类错误率决策：

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0, & \alpha_i = \omega_j \\ 1, & \alpha_i \neq \omega_j \end{cases}$$

$$\arg\min_i R(\alpha_i \mid \mathbf{x}) = \arg\min_i 1 - P(\omega_i \mid \mathbf{x}) = \arg\max_i P(\omega_i \mid \mathbf{x})$$

最大后验概率决策

最小分类错误率 $R^* = \int (1 - \arg\max_i P(\omega_i \mid \mathbf{x})) p(\mathbf{x}) d\mathbf{x}$

对给定的特征空间，最小分类错误率是确定的；要减小分类错误率，只能改进特征空间。

上次课主要内容回顾

- 贝叶斯决策
  - 决策损失： $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ 把 j 类样本决策为 i 类的损失
  - 条件风险： $R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$ 把样本**x**决策为 i 类的期望损失
  - 最小风险决策：

$$\arg\min_i R(\alpha_i \mid \mathbf{x})$$

  - 最小分类错误率决策：

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0, & \alpha_i = \omega_j \\ 1, & \alpha_i \neq \omega_j \end{cases}$$

$$\arg\max_i R(\alpha_i \mid \mathbf{x}) = \arg\max_i P(\omega_i \mid \mathbf{x})$$ 最大后验概率决策

  - Reject recognition

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0, & \alpha_i = \omega_j \\ \lambda_s, & \alpha_i \neq \omega_j \\ \lambda_r, & \text{reject} \end{cases} \quad (\lambda_r < \lambda_s)$$

$$\arg\min_i R_i(\mathbf{x}) = \begin{cases} \arg\max_i P(\omega_i \mid \mathbf{x}), & \max_i P(\omega_i \mid \mathbf{x}) > 1 - \lambda_r / \lambda_s \\ \text{reject}, & \text{otherwise} \end{cases}$$

上次课主要内容回顾

- 判别函数(Discriminant Function)
  - 表征样本 x 属于某一类的广义似然度：

$$g_i(\mathbf{x}) \quad i = 1, \ldots, c$$

  - 多种形式：

$$g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$$

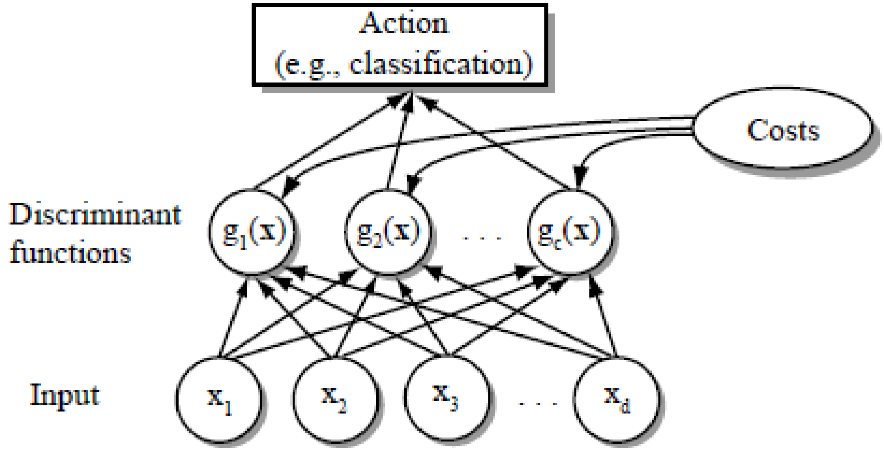
$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$$

$$g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid \omega_i) + \log P(\omega_i)$$

  - 已知类判别函数后的分类准则：

$$\arg\max_i g_i(\mathbf{x})$$



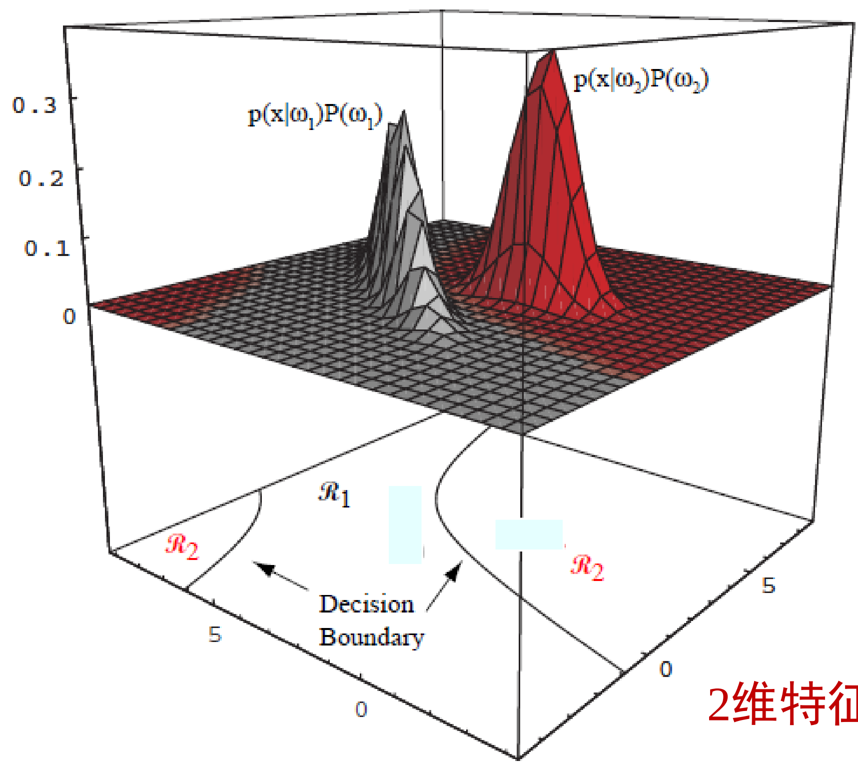
上次课主要内容回顾

- 决策面(Decision Boundary)
  - 特征空间中二类判别函数相等的点的集合

令$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$，决策面由所有满足 $g(\mathbf{x})=0$ 的 $\mathbf{x}$ 组成

正态分布下的一个例子

$$g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$$



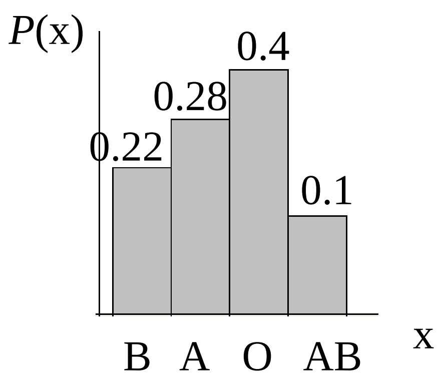
2维特征空间

上次课主要内容回顾

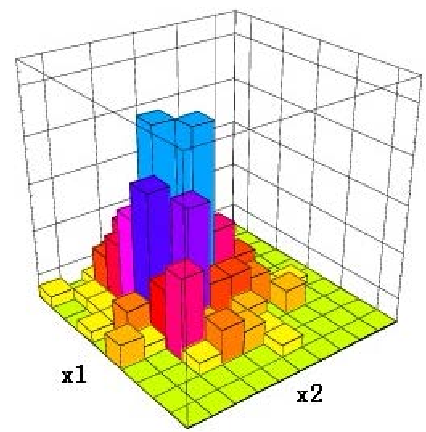
- 高斯分布（正态分布）
  - 1D、多维（记住了？）
  - 协方差矩阵的性质
    - 等密度点轨迹、马氏距离
  - 线性变换的高斯密度
- 高斯分布下的判别函数
  - Linear discriminant function (LDF): $\Sigma_i = \Sigma_j$
  - Quadratic discriminant function (QDF): $\Sigma_i \neq \Sigma_j$

2.7 离散变量贝叶斯决策

- 离散特征变量
  - 例如：问卷调查，每个问题2个或多个选项
    医疗诊断：是否有某个症状
  - 概率分布函数 $P(\mathbf{x}|\omega_i)=P(x_1,x_2,\ldots,x_d\,|\,\omega_i)$
    （非参数、直方图表示）



一维分布（血型）



二维离散分布

- 独立二值特征 (Binary features)
  - 特征独立假设（Naïve Bayes）：

$$P(\mathbf{x}\mid\omega_i)=P(x_1,x_2,\cdots,x_d\mid\omega_i)=\prod_{j=1}^{d}P(x_j\mid\omega_i)$$

  - 每维特征服从伯努利分布（0/1分布）：

$$p_i=P(x_i=1|\omega_1)\quad i=1,\dots,d$$

$$P(\mathbf{x}|\omega_1)=\prod_{i=1}^{d}p_i^{x_i}(1-p_i)^{1-x_i}$$

$$q_i=P(x_i=1|\omega_2)\quad i=1,\dots,d$$

$$P(\mathbf{x}|\omega_2)=\prod_{i=1}^{d}q_i^{x_i}(1-q_i)^{1-x_i}$$

  - 似然比(Likelihood ratio)

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)}=\prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

- 独立二值特征 (Binary features)
  - Discriminant/decision function

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}) = \ln P(\mathbf{x}|\omega_1)P(\omega_1) - \ln P(\mathbf{x}|\omega_2)P(\omega_2)$$

$$= \sum_{i=1}^{d}[x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}] + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

$$= \sum_{i=1}^{d}\ln\frac{p_i}{q_i}\frac{1-q_i}{1-p_i}x_i + \sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i} + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

- g(**x**)为线性判别函数

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0$$

w_i 为每个特征的权重

$$w_i = \ln\frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\dots,d \qquad w_0 = \sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i} + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

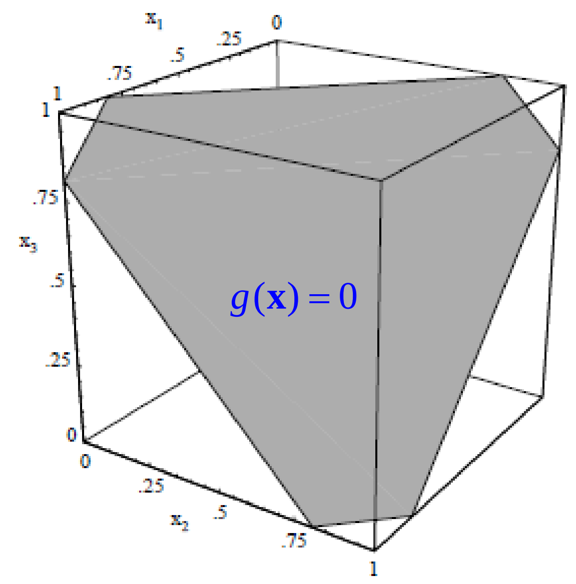
- 一个例子: 3D binary data
  - $P(\omega_1)=0.5$, $P(\omega_2)=0.5$
  - $p_i=0.8$, $q_i=0.5$, $i=1,2,3$

$$P(\mathbf{x}|\omega_1)=\prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i} \qquad P(\mathbf{x}|\omega_2)=\prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

$$g(\mathbf{x})=\sum_{i=1}^{d} w_i x_i + w_0$$

$$w_i = \ln\frac{.8(1-.5)}{.5(1-.8)} = 1.3863$$

$$w_0 = \sum_{i=1}^{3}\ln\frac{1-.8}{1-.5}+\ln\frac{.5}{.5} = -2.7489$$



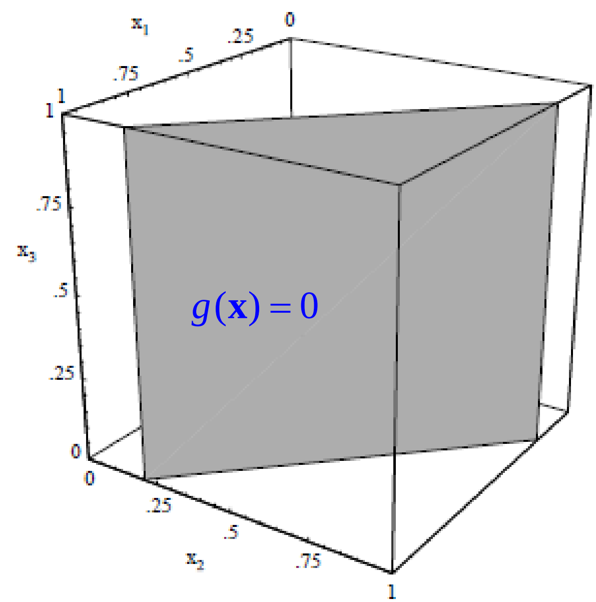
- 另一个例子: 3D binary data
  - $P(\omega_1)$=0.5, $P(\omega_2)$=0.5
  - p_1=p_2=0.8, p_3=0.5; q_i=0.5, i=1,2,3

$$w_i = \ln \frac{.8(1-.5)}{.5(1-.8)} = 1.3863$$

$$(w_1 = w_2)$$

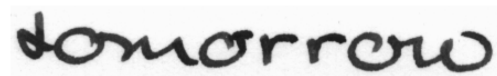
$$w_3 = 0$$

$$\mathrm{w}_0 = 2\ln\frac{1-0.8}{1-0.5} = -1.8326$$



2.8 复合模式分类

(Compound Bayesian Decision Theory and Context)

- 多个样本同时分类 $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n]$ $\boldsymbol{\omega}=\omega(1)\omega(2)\cdots\omega(n)$
  - 比如：字符串识别 

  - Bayesian decision

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

    - 注意：ω类别数巨大(c^n)，$p(\mathbf{X}|\,\boldsymbol{\omega})$存储和估计困难
  - Conditionally independent

$$p(\mathbf{X}\,|\,\omega)=\prod_{i=1}^{n}p(\mathbf{x}_i\,|\,\omega(i))$$

已知类别的条件下，样本之间相互独立

  - Prior assumption
    - Markov chain

$$P(\boldsymbol{\omega})=P[\omega(1),\omega(2),\cdots,\omega(n)]=P[\omega(1)]\prod_{j=2}^{n}P[\omega(j)\,|\,\omega(j-1)]$$

  - Hidden Markov model (Chapter 3)

与复合模式识别类似的问题：多分类器融合

- 有同一个分类问题的K个分类器，对于样本$\mathbf{x}$，怎样使用K个分类结果得到最终分类结果？
  - 一个分类器的输出：离散变量 $e_k \in \{\omega_1,\ldots,\omega_c\}$
  - 多个分类器的决策当作样本$\mathbf{x}$的多维特征，用Bayes方法重新分类

$$P(\omega_i \mid e_1,\ldots,e_K) = \frac{\boxed{P(e_1,\ldots,e_K \mid \omega_i)}P(\omega_i)}{P(e_1,\ldots,e_K)}, \quad i=1,\ldots,c$$

  - 需要估计离散空间的类条件概率

$$P(e_1,\ldots,e_K \mid \omega_i)$$ 指数级复杂度，需要大量样本

  - 特征独立假设（Naïve Bayes）

$$P(e_1,\ldots,e_K \mid \omega_i) = \prod_{k=1}^{K} P(e_k \mid \omega_i)$$

第2章小结

本章，我们探讨了在已知类条件概率密度$p(\mathbf{x}|\omega_i)$ 和类先验分布$P(\omega_i)$的情况下，如何基于贝叶斯决策理论对样本**x**分类的问题

（1）单模式分类：连续特征、离散特征

（2）复合模式分类

（3）多分类器融合

第3章
最大似然和贝叶斯参数估计

3.1 基本概念

- 贝叶斯分类器
  - 已知类先验概率$P(\omega_i)$和类条件概率密度$p(\mathbf{x}|\omega_i)$，按某决策规则确定判别函数和决策面。
  - 但类先验概率和类条件概率密度在实际中往往是未知的。
  - 因此，我们要换一种处理问题的方式："从样本出发来设计分类器"。根据设计方法，可以将分类器分为两类：
    - 估计类先验概率和类条件概率密度函数（产生式方法）
    - 直接估计类后验概率或判别函数（判别式方法）

3.1 基本概念

- **方法分类**
  - 参数估计：
    - 样本所属的类条件件概率密度函数的形式已知，而概率密度函数的参数是未知的
    - 目标是由已知类别的样本集估计概率密度函数的参数
    - 例如，知道样本所属总体为正态分布，而正态分布的参数未知

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

  - 非参数估计：
    - 样本所属的类条件件概率密度函数的形式和参数都是未知的，目标是由已知类别的样本集估计类条件件概率密度函数本身。

3.1 基本概念

- 基本概念
  - 统计量：样本中包含总体的信息，我们希望通过样本集将有关信息估计出来。根据不同要求构造出有关样本的某种函数，在统计学中称为统计量$d(x_1,x_2,\ldots,x_n)$ 。比如，
    - 均值 $\mu=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$
  - 参数空间：将未知待估计参数记为θ，参数θ 的全部允许取值集合构成参数空间，记为Θ。

3.1 基本概念

- 基本概念
  - 点估计：点估计问题就是构造一个统计量$d(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$作为参数$\theta$的估计 $\hat{\theta}$。比如，常用的均值估计：

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

  - 区间估计：与点估计不同，区间估计要求采用(d_1, d_2)作为参数θ 可能取值范围的一种估计。这个区间称为置信区间。这类估计问题称为区间估计。

3.2 最大似然估计

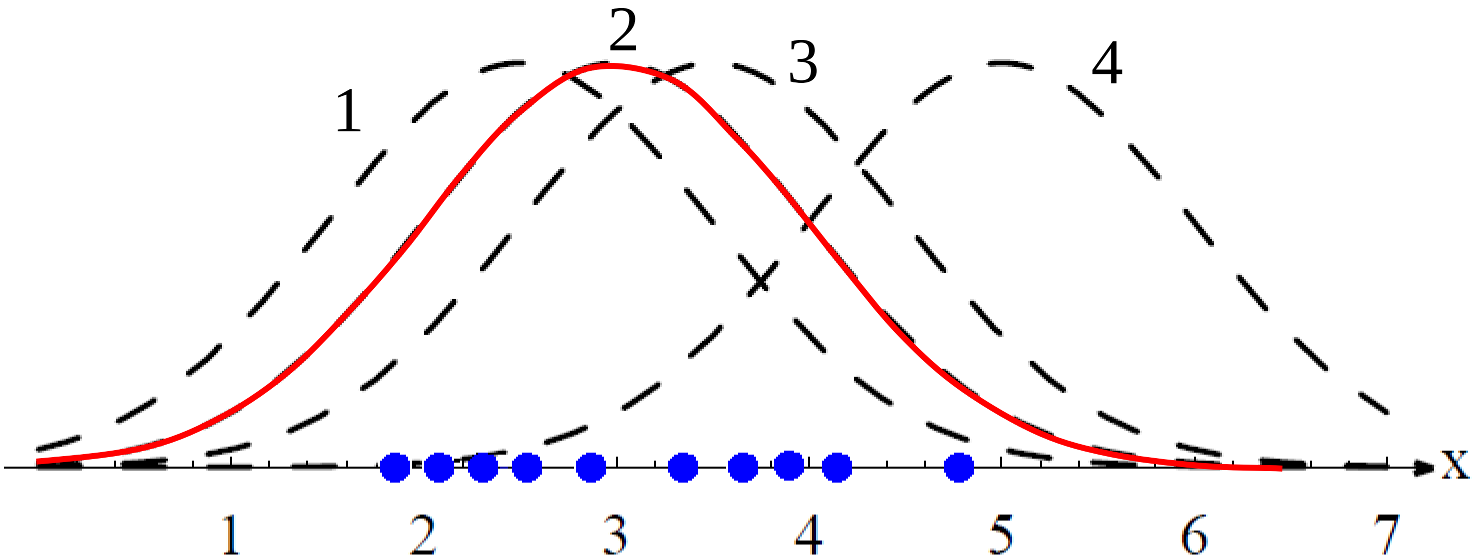
- 基本假设
  - 独立同分布假设：每类样本均是从类条件概率密度 $p(\mathbf{x}|\omega_i)$中独立抽取出来的。
  - $p(\mathbf{x}|\omega_i)$具有确定的函数形式，只是其中的参数θ未知：
    - 比如，当 $\mathbf{x}$ 服从一维正态分布$N(\mu,\sigma^2)$，未知的参数为$\boldsymbol{\theta}=[\mu,\sigma]^T$，为一个二维向量。
  - 各类样本只包含本类的分布信息：即不同类别的参数是独立的。可以分别处理c个独立问题。

3.2 最大似然估计

- 基本原理
  - 已知随机抽取的n个样本(观测值)，最合理的参数估计应该是使得从该模型中能抽取这n个样本的概率最大。
- 直观想法：
  - 一个随机试验如有若干个可能的结果：A，B，C，…。若仅作一次试验，结果A出现，则认为试验条件（模型参数）对A出现有利，也即A出现的概率很大。
    - 一般地，事件A发生的概率与参数θ相关，A发生的概率记为$P(A|\theta)$，则θ的估计应该使上述概率达到最大，这样的θ顾名思义称为极大似然估计。

3.2 最大似然估计

- 例如：我们观察到10个一维空间中的样本。现假定其来自于4个高斯分布中的一个。哪一个最有可能？



- 基本原理
  - 设样本集包含n个样本$D=\{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$，这些样本是从概率密度函数$p(\mathbf{x}|\boldsymbol{\theta})$中独立抽取的，则获得 n 个样本的联合概率为：

$$l(\boldsymbol{\theta}) = P(D\,|\,\boldsymbol{\theta}) = P(\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\,|\,\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i\,|\,\boldsymbol{\theta})$$

  - $l(\theta)$是θ的函数，描述了在不同参数取值下取得当前样本集的可能性。
  - $l(\theta)$被称为参数θ相对于样本集D的似然函数。
    - 似然函数给出了从总体中抽出$\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n$这$n$个样本的概率。

3.2 最大似然估计

- 方法描述
  - 令$l(\boldsymbol{\theta})$为样本集D的似然函数，$D=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$。如果$\hat{\boldsymbol{\theta}}$是参数空间$\Theta$中能使$l(\theta)$极大化的$\theta$值，那么$\hat{\boldsymbol{\theta}}$就是$\boldsymbol{\theta}$的最大似然估计量，即

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta})$$

  - 为计算方便，通常采用对数似然函数：

$$H(\boldsymbol{\theta}) = \ln(l(\boldsymbol{\theta})) = \ln\prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln(p(\mathbf{x}_i \mid \boldsymbol{\theta}))$$

$$\arg\max l(\boldsymbol{\theta}) = \arg\max H(\boldsymbol{\theta})$$

- **问题求解**

$$H(\boldsymbol{\theta}) = \ln(l(\boldsymbol{\theta})) = \ln\left(\prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \ln(p(\mathbf{x}_i \mid \boldsymbol{\theta}))$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta})$$

当$l(\theta)$可微时：

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0, \quad \text{or} \quad \frac{\partial H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

对于多维情形$\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_m]^T$，梯度向量为零：

$$\nabla_{\boldsymbol{\theta}}(l(\boldsymbol{\theta})) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_2}, \ldots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_m}\right]^T = \mathbf{0}$$

用梯度上升法求解 $\boldsymbol{\theta}' = \boldsymbol{\theta} + \eta \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

• 问题求解

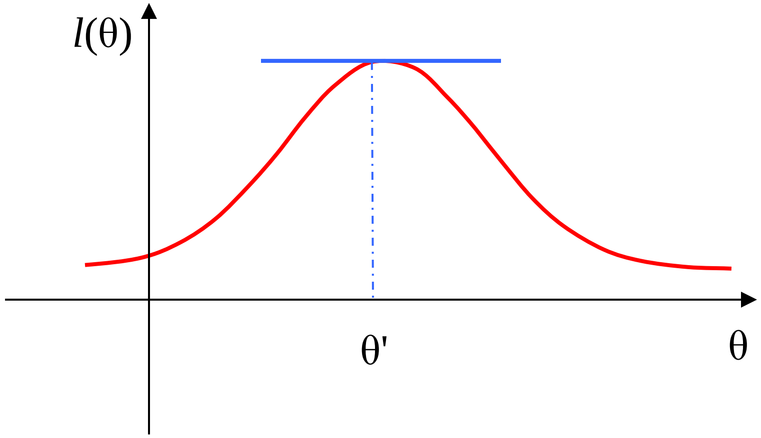
$$H(\boldsymbol{\theta}) = \ln(l(\boldsymbol{\theta})) = \ln\left(\prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \ln(p(\mathbf{x}_i \mid \boldsymbol{\theta}))$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta})$$

当$l(\theta)$是可微凹函数时：

$$\arg\max l(\boldsymbol{\theta}) \Leftrightarrow \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

梯度等于0是最优解的充要条件

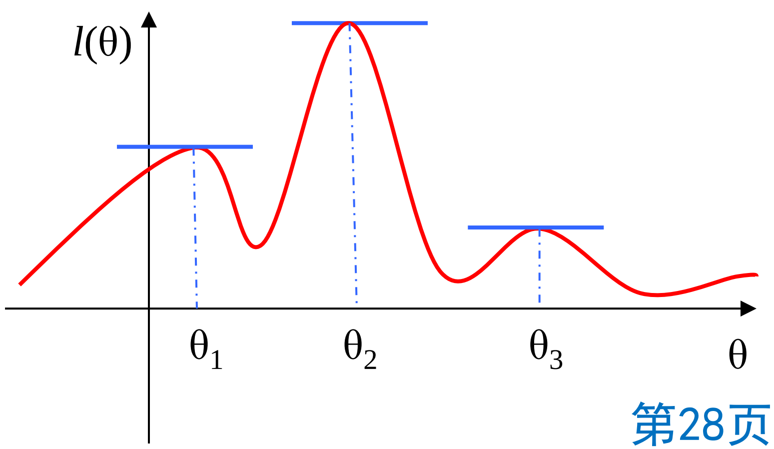


当$l(\theta)$是一般可微函数时：

$$\arg\max l(\boldsymbol{\theta}) \Rightarrow \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

梯度等于0是最优解的必要条件

在高维空间中，寻找全局最优解是极困难的，通常满足于局部最优解。



3.2 最大似然估计

– 一个袋子里装有白球与黑球，但是不知道它们之间的比例。现有放回地抽取10次，结果获得8次黑球2次白球，估计袋子中的黑球的比例。
  - 最大似然估计法：设抽到黑球的概率为p，取到8次黑球2次白球的概率为：$l(p)=\binom{10}{8}p^8(1-p)^2$
  - 计算$l(p)$和$H(p)$的最优解：

$$\frac{\partial l(p)}{\partial p}=\binom{10}{8}\frac{\partial p^8(1-p)^2}{\partial p}=\binom{10}{8}(10p^9-18p^8+8p^7)=0 \quad \Rightarrow \quad \hat{p}=0.8$$

$$\frac{\partial \ln l(p)}{\partial p}=\frac{\partial \ln\binom{10}{8}+8\ln p+2\ln(1-p)}{\partial p}=\frac{8}{p}-\frac{2}{1-p}=0 \quad \Rightarrow \quad \hat{p}=0.8$$

- 例子2：高斯分布下的最大似然估计
  - 将在下面的推导中用到如下几个预备公式（见清华大学张贤达老师著的《矩阵分析与应用》第五章：“梯度分析应用”）：

$$tr(\mathbf{A}) = \sum_{i=1}^{d} A_{ii}, \quad \text{where } \mathbf{A} = (A_{ij}) \in R^{d\times d}$$

$$s = tr(s), \quad \text{if } s \text{ is } a \text{ scalar} \Rightarrow \mathbf{x}^T\mathbf{A}\mathbf{x} = tr\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right), \text{where } \mathbf{x} \in R^d, \mathbf{A} \in R^{d\times d}$$

$$\frac{\partial\left|\mathbf{\Sigma}\right|}{\partial\mathbf{\Sigma}} = \left|\mathbf{\Sigma}\right|\left(\mathbf{\Sigma}^{-1}\right), \quad \text{if } \mathbf{\Sigma} \text{ is symmetrical}$$

$$\frac{\partial tr\left(\mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{B}\right)}{\partial\mathbf{\Sigma}} = \left(-\mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{A}\mathbf{\Sigma}^{-1}\right)^T \Rightarrow \frac{\partial\left(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right)}{\partial\mathbf{\Sigma}} = -\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^T\mathbf{\Sigma}^{-1}, \quad \text{if } \mathbf{\Sigma} \text{ is symmetrical}$$

- **例子2：高斯分布下的最大似然估计**

$$\mathbf{x} = [x_1, x_2, \ldots, x_d]^T \in R^d,$$
$$\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_d]^T \in R^d,$$
$$\boldsymbol{\Sigma} \in R^{d \times d}$$

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

$$\ln p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \ln[(2\pi)^d |\boldsymbol{\Sigma}|] - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, n$$

$$\nabla_{\boldsymbol{\theta}} (H(\boldsymbol{\theta})) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \mathbf{0}$$

(where $\boldsymbol{\theta} = \boldsymbol{\mu}$, and (or) $\boldsymbol{\Sigma}$)

中国科学院大学 University of Chinese Academy of Sciences

- 估计$\boldsymbol{\mu}$

$$\ln p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\ln[(2\pi)^d |\boldsymbol{\Sigma}|] - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, n$$

$$\frac{\partial H(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \frac{\partial \ln p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$\sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) = \mathbf{0} \quad \Rightarrow \quad \sum_{i=1}^{n} \mathbf{x}_i - n\hat{\boldsymbol{\mu}} = 0 \quad \Rightarrow \quad \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

$$\frac{\partial\left(\mathbf{x}^T \mathbf{A}\mathbf{x}\right)}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}, \text{ where } \mathbf{x} \in R^d, \mathbf{A} \in R^{d\times d} \text{ is a real symmetrical matrx}$$

- **估计Σ**

$$\ln p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\ln[(2\pi)^d |\boldsymbol{\Sigma}|] - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, n$$

$$\Downarrow$$

$$\frac{\partial H(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = \sum_{i=1}^{n} \frac{\partial \ln p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \frac{\partial \ln[(2\pi)^d |\boldsymbol{\Sigma}|]}{\partial \boldsymbol{\Sigma}} - \frac{1}{2} \frac{\partial \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)}{\partial \boldsymbol{\Sigma}} \right)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \frac{\partial \ln[|\boldsymbol{\Sigma}|]}{\partial \boldsymbol{\Sigma}} - \frac{1}{2} \frac{\partial \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)}{\partial \boldsymbol{\Sigma}} \right)$$

(复合函数求偏导) $$= \sum_{i=1}^{n} \left(-\frac{1}{2} \frac{1}{|\boldsymbol{\Sigma}|} \frac{\partial |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}} - \frac{1}{2} \frac{\partial \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)}{\partial \boldsymbol{\Sigma}} \right)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \frac{1}{|\boldsymbol{\Sigma}|} |\boldsymbol{\Sigma}| \boldsymbol{\Sigma}^{-1} - \frac{1}{2} \frac{\partial \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)}{\partial \boldsymbol{\Sigma}} \right)$$

$$= \sum_{i=1}^{n} \left(-\frac{1}{2} \boldsymbol{\Sigma}^{-1} - \frac{1}{2} \left(-\boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \right) \right)$$

$$= -\frac{1}{2} \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \left(\mathbf{I} - (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \right)$$

- **估计Σ（接上页）**

$$\frac{\partial H(\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}}=\mathbf{0} \quad \Rightarrow \quad -\frac{1}{2}\hat{\boldsymbol{\Sigma}}^{-1}\sum_{i=1}^{n}\left(\mathbf{I}-(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T\hat{\boldsymbol{\Sigma}}^{-1}\right)=\mathbf{0}$$

$$\Rightarrow \quad \sum_{i=1}^{n}\left(\mathbf{I}-(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T\hat{\boldsymbol{\Sigma}}^{-1}\right)=\mathbf{0}$$

$$\Rightarrow \quad \sum_{i=1}^{n}\left(\hat{\boldsymbol{\Sigma}}-(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T\right)=\mathbf{0}$$

$$\Rightarrow \quad n\hat{\boldsymbol{\Sigma}}-\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T=\mathbf{0}$$

$$\Rightarrow \quad \hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$$

- 例子2：高斯分布下的最大似然估计
  - μ、Σ均未知

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

一维：$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i, \quad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

(直接考虑以下问题也得得到同样的结论)

$$\max \sum_{i=1}^{n}\ln p(x_i \mid \mu, \sigma), \quad \text{where } p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

3.3 贝叶斯估计

- 贝叶斯估计与最大似然估计
  - 贝叶斯估计是概率密度估计中另一类主要的参数估计方法。其结果在很多情况下与最大似然法十分相似，但是，两种方法对问题的处理视角是不一样的。
    - 最大似然估计是将待估计的参数当作未知但固定的变量，其任务是根据观测数据估计其在参数空间中的取值。
    - 贝叶斯估计将待估计的参数视为一个随机变量，其中的一个核心任务是根据观测数据对参数的分布进行估计。

3.3 贝叶斯估计

- 基本方法
  - 参数先验分布 $p(\boldsymbol{\theta})$：是指在没有任何数据时，有关参数θ 的分布情况 （根据领域知识或经验）
  - 给定样本集 $D=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$，数据独立采样，且服从数据分布：

$$p(D\,|\,\boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n\,|\,\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i\,|\,\boldsymbol{\theta})$$

  - 利用贝叶斯公式计算参数的后验分布 $p(\boldsymbol{\theta}|D)$：

$$p(\boldsymbol{\theta}\,|\,D) = \frac{p(D\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(D)}$$

$p(\boldsymbol{\theta}|D)$中融合了先验知识和数据信息

$$p(x\mid\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\qquad p(D)=\sum_i p(D\mid\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)$$

- 基本方法
  - $p(D)$是与参数无关的归一化因子，根据全概率公式(连续）：

$$p(D)=\int_{\boldsymbol{\theta}} p(D\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

对于一般情况，计算$p(D)$十分困难

  - 可得贝叶斯参数估计中的后验概率密度函数：

$$p(\boldsymbol{\theta}\mid D)=\frac{p(D\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}=\frac{\prod_{i=1}^n p(\mathbf{x}_i\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}}\prod_{i=1}^n p(\mathbf{x}_i\mid\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}=\alpha\prod_{i=1}^n p(\mathbf{x}_i\mid\boldsymbol{\theta})p(\boldsymbol{\theta})$$

3.3 贝叶斯估计

- 如何使用$p(\theta \mid D)$获得关于数据的分布？
  - 得到 $p(\boldsymbol{\theta} \mid D)$ 只是获得了关于参数$\boldsymbol{\theta}$的后验分布，并没有像最大似然估计那样获得参数$\boldsymbol{\theta}$ 的具体取值。
  - 方法一：可对$p(\boldsymbol{\theta}|D)$采样，计算平均值

$$\hat{\boldsymbol{\theta}} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\theta}_i \qquad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta} \mid D) \qquad i = 1, \ldots, M$$

  - 方法二：最大后验估计(Maximum A Posteriori estimation, MAP)

$$\hat{\boldsymbol{\theta}} = \arg\max\, p(\boldsymbol{\theta} \mid D)$$

$$\Leftrightarrow \hat{\boldsymbol{\theta}} = \arg\max\, p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

$$\Leftrightarrow \hat{\boldsymbol{\theta}} = \arg\max\, \ln p(D \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$$

PR/ML方法中普遍使用的L2正则，等价于假设参数服从$N(0, \mathbf{I})$

3.3 贝叶斯估计

- 方法三：后验数据分布（完整的贝叶斯方法）
  - 我们的最终目的是根据D中的样本来估计概率密度函数$p(\mathbf{x}|D)$。
    - 比如，假定观测样本服从正态分布$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，给定$D$，可以估计得到具体的 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的取值，代入如下公式可得关于样本的密度分布函数：

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

- 后验数据分布
  - 但现在获得了有关θ的后验估计$p(\boldsymbol{\theta}\,|D)$，如何估计$p(\mathbf{x}|D)$？考虑全概率公式和边际分布：

$$
\begin{aligned}
p(\mathbf{x}\,|\,D) &= \int_{\boldsymbol{\theta}} p(\mathbf{x},\boldsymbol{\theta}\,|\,D)d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \frac{p(\mathbf{x},\boldsymbol{\theta},D)}{p(D)}d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \frac{p(\mathbf{x}\,|\,\boldsymbol{\theta},D)\quad p(\boldsymbol{\theta},D)}{p(D)}d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta},D)\quad p(\boldsymbol{\theta}\,|\,D)\,d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})\quad p(\boldsymbol{\theta}\,|\,D)\,d\boldsymbol{\theta}
\end{aligned}
$$

在给定参数θ时，样本分布与训练集D无关

不同参数的密度函数的加权平均

  - 积分通常很难计算，使用近似方法：

$$
\hat{p}(\mathbf{x}\,|\,D) = \frac{1}{M}\sum_{i=1}^{M} p(\mathbf{x}|\boldsymbol{\theta}_i) \qquad \boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}\,|\,D) \qquad i=1,\ldots M
$$

M个不同参数的密度函数的平均

- **正态分布下的贝叶斯参数估计**
  - 先考虑一维情形，假定$X \sim N(\mu, \sigma^2)$且仅μ未知。
  - 假定参数μ的先验概率也服从正态分布：

$$\mu \sim N(\mu_0, \sigma_0^2)$$

  - 第一个任务：给定样本集D，在上述条件下，估计关于参数的后验分布$p(\mu \mid D)$。
  - 回顾我们前面得到的公式：

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \alpha \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

- **正态分布下的贝叶斯参数估计：一维**
  - $p(x|\mu) = N(\mu, \sigma^2)$, $p(\mu) = N(\mu_0, \sigma_0^2)$

$$p(\boldsymbol{\theta} \mid D) = \alpha \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

（应用后验估计）

$$p(\mu \mid D) = \alpha \prod_{i=1}^{n} p(\mathbf{x}_i \mid \mu) p(\mu)$$

$$= \alpha \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{1}{2} \frac{(\mu - \mu_0)^2}{\sigma_0^2} \right)$$

$$= \alpha' \exp\left\{ -\frac{1}{2} \left(\sum_{i=1}^{n} \left(\frac{(x_i - \mu)^2}{\sigma^2} \right) + \frac{(\mu - \mu_0)^2}{\sigma_0^2} \right) \right\}$$

$$= \alpha'' \exp\left(-\frac{1}{2} \left(\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2} \right) \mu \right) \right)$$

- **正态分布下的贝叶斯参数估计:一维**

$$p(\mu|D)=\alpha''\exp\left(-\frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right)$$

  – $p(\mu|D)$是关于μ的二次函数的exp函数，因此，也是一个正态分布密度函数。

  – $p(\mu|D)$ is said to be **reproducing density**, because this is true for any number of training samples, $p(\mu|D)$ remains normal as the number of $\boldsymbol{n}$ of samples is increased

- **正态分布下的贝叶斯参数估计:一维**
  - As $p(\boldsymbol{\mu} | D)$ is a normal density function, we can rewritten it as follows:

$$p(\mu | D) \sim N(\mu_n, \sigma_n^2) = \frac{1}{\sqrt{2\pi\sigma_n}} \exp\left(-\frac{1}{2} \frac{(\mu - \mu_n)^2}{\sigma_n^2} \right)$$

  - But, at the same time, we also get its formulation as

$$p(\mu | D) = \alpha'' \exp\left(-\frac{1}{2}\left(\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2} \right)\mu \right) \right)$$

$$\Downarrow$$

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}, \quad \frac{\mu_n}{\sigma_n^2} = \underline{\frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}}, \quad \text{where } \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **正态分布下的贝叶斯参数估计:一维**
  - 进一步可解得:

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0, \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

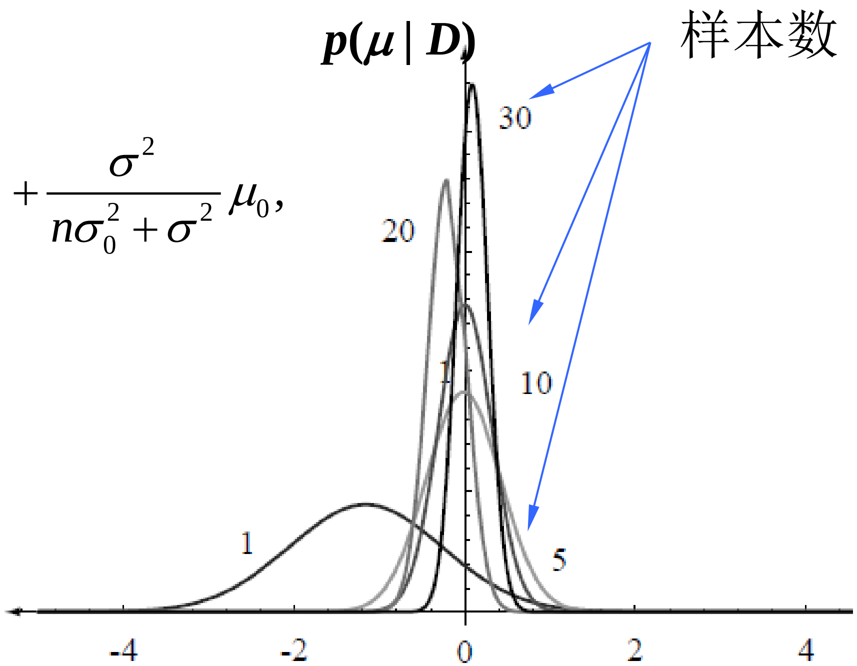
  - These equations show how **the prior information** is combined with **the empirical information** in the samples to obtain the a posterior density $p(\mu|D)$.
  - μ_n : represents our best guess for μ after obtaining n samples.
  - σ_n^2: **measures the uncertainty about the guess of μ**.
  - Because σ_n^2 decreases **monotonically with n**, each additional observation will help decrease our uncertainty about the true value of μ.

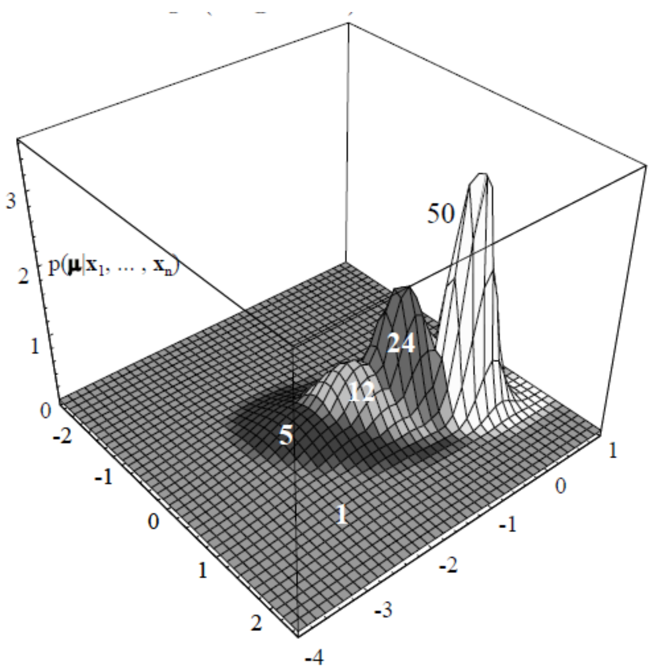
(这种先验起到了平滑的效果，导致了更加鲁棒的估计)

- # 正态分布下的贝叶斯参数估计:一维
  - Because $(\sigma_n)^2$ decreases monotonically with n, each additional observation will help decrease our uncertainty about the true value of μ. As n increase, $p(\mu\ |D)$ becomes more and more sharply peaked, approaching a Dirac delta function as n approaches infinity.

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0,$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$





- **正态分布下的贝叶斯参数估计:一维**
  - 现在，我们希望获得后验数据分布

$$p(\mathbf{x} \mid D) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

$$p(x \mid D) = \int_{\mu} p(x|\mu)\, p(\mu \mid D)\, d\mu$$

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

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left(-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right) f(\sigma, \sigma_n)$$

where $f(\sigma, \sigma_n) = \int_{\mu} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2+\sigma_n^2}\right)^2\right) d\mu$

3.4 正态分布下的贝叶斯估计

- 贝叶斯估计与最大似然估计

贝叶斯估计： $p(x|D) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$

最大似然估计： $p(x|D) \sim N(\hat{\mu}_n, \sigma^2)$

$$\mu_n = \frac{n\sigma^2}{n\sigma_0^2 + \sigma^2}\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0, \qquad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2},$$

3.4 正态分布下的贝叶斯估计

- 正态分布下的贝叶斯参数估计
  - 多元情形（高维情形）：

$$p(\mathbf{x}\,|\,\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

$$p(\boldsymbol{\theta}\,|\,D) = \alpha \prod_{i=1}^{n} p(\mathbf{x}_i\,|\,\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

$$= \alpha' \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}^T \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \right) \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \left(\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} \mathbf{x}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \right) \right)$$

$$= \alpha'' \exp\left(-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right)$$

- 进一步，我们有：

$$p(\boldsymbol{\theta} \mid D) = \alpha'' \exp\left(-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right) \Rightarrow p(\boldsymbol{\theta} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$$

$$\Rightarrow \boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}, \quad \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n = n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \qquad \hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \hat{\boldsymbol{\mu}}_n + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n}\boldsymbol{\Sigma}$$

$$\because \ (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}$$

- **Data posterior distribution:**

$$p(\mathbf{x} \mid D) = \int_{\boldsymbol{\mu}} p(\mathbf{x} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid D) d\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

3.5 贝叶斯学习

- **一般情形下的贝叶斯估计（总结）**
  - The basic assumption are summarized as follows:
    - The form of the density $p(\mathbf{x}|\boldsymbol{\theta})$ is assumed to be known, but the value of the parameter vector is not known exactly.
    - Our initial knowledge about $\boldsymbol{\theta}$ is assumed to be contained in a known prior density $p(\boldsymbol{\theta})$
    - The rest knowledge about $\boldsymbol{\theta}$ is contained in a set D of n samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ drawn independently according to the unknown probability density $p(\mathbf{x})$
  - **The basic problem** is to compute the posterior density $p(\boldsymbol{\theta}|D)$ about parameter θ and the posterior density $p(\mathbf{x}|D)$ about data.

3.5 贝叶斯学习

$p(\boldsymbol{\theta}), p(\mathbf{x}|\boldsymbol{\theta}) \rightarrow p(\boldsymbol{\theta}|D) \rightarrow p(\mathbf{x}|D)$

- **一般情形下的贝叶斯估计（总结）**
  - The basic problem is to compute the posterior density $p(\boldsymbol{\theta}|D)$ about the parameter θ . By Bayes formula we have:

$$p(\boldsymbol{\theta}\,|\,D) = \frac{p(D\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

where $P(D\,|\,\boldsymbol{\theta}) = P(\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_n\,|\,\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i\,|\,\boldsymbol{\theta})$

  - After that, we can obtain $p(\mathbf{x}|D)$ as follows:

$$p(\mathbf{x}\,|\,D) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,D)d\boldsymbol{\theta}$$

3.5 贝叶斯学习

- 遇到的困难
  - 除了一些特殊的分布(共轭分布)之外，对于一般情形，积分很难计算：

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$p(\mathbf{x} \mid D) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}$$

  - 参数先验 $p(\theta)$ 怎么选取？对结果有何影响？
  - 给定D，我们真的能通过$p(\mathbf{x}|D)$将$p(\mathbf{x})$估计得很好吗？或者说，随着D中样本的增多，$p(\mathbf{x}|D)$收敛于$p(\mathbf{x})$吗？

- **贝叶斯学习的迭代计算公式**
  - 记$D^n=\{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$，由于样本是独立选样，则：

$$p(D^n|\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})\,p(D^{n-1}|\boldsymbol{\theta}) = p(\mathbf{x}_n|\boldsymbol{\theta})\,p(\mathbf{x}_{n-1}|\boldsymbol{\theta})\,p(D^{n-2}|\boldsymbol{\theta}) = \cdots$$

  - 于是有如下迭代公式：

$$p(\boldsymbol{\theta}|D^n) = \frac{p(D^n|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D^n|\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{p(\mathbf{x}_n|\boldsymbol{\theta})\,p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(\mathbf{x}_n|\boldsymbol{\theta})\,p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

$$= \frac{p(\mathbf{x}_n|\boldsymbol{\theta})}{\int p(\mathbf{x}_n|\boldsymbol{\theta})\dfrac{p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\boxed{\int p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}}}d\boldsymbol{\theta}} \cdot \frac{p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\boxed{\int p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}}}$$

$$= \frac{p(\mathbf{x}_n|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|D^{n-1})}{\int p(\mathbf{x}_n|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|D^{n-1})d\boldsymbol{\theta}}$$

一个固定的数

$$\left(\because\ p(\boldsymbol{\theta}|D^{n-1}) = \frac{p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int p(D^{n-1}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})d\boldsymbol{\theta}}\right)$$

3.5 贝叶斯学习

- **参数迭代学习方法**
  - 为统一表示，记参数先验分布$p(\theta)$为$p(\theta|D^0)$，表示没有样本情形下的参数概率密度估计。
  - 记$D^n=\{x_1, x_2, \ldots, x_n\}$，随着样本的增加，可以得到一系列对参数概率密度函数的估计：

$$p(\boldsymbol{\theta}),\ p(\boldsymbol{\theta}|\mathbf{x}_1),\ p(\boldsymbol{\theta}|\mathbf{x}_1,\mathbf{x}_2), \ldots,\ p(\boldsymbol{\theta}|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n), \cdots$$

  - 一般来说，随着样本的数目的增加，上述序列函数逐渐尖锐，逐步趋向于以θ的真实值为中心的一个尖峰。当样本无穷多时，此时将收敛于一个脉冲函数（参数真值）。

3.6 贝叶斯估计：一个例子

- **一个例子**
  - 假设一维随机变量**X**服从[0, θ]上的均匀分布:

$$p(x|\theta) = U(0,\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

基于先验知识，我们知道0 < θ < 10，并希望利用迭代的贝叶斯方法从样本**{ 4, 7, 2, 8 }**中，估计参数θ.

3.6 贝叶斯估计：一个例子

- **一个例子**
  - **Before any data arrive, we have** $p(\theta|D^0) = p(\theta) = U(0,10)$ **.**
  - **When our first data point $x_1 = 4$ arrives, then**

$$p(\theta|D^1) = \frac{p(x_1|\theta)p(\theta|D^0)}{\int p(x_1|\theta)p(\theta|D^0)d\theta} = \alpha p(x_1|\theta)p(\theta|D^0) = \alpha\frac{1}{\theta}\cdot\frac{1}{10} \propto \frac{1}{\theta}$$

$$p(\theta|D^1) \propto \begin{cases} 1/\theta, & 4 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

where throughout we will ignore the normalization.

因为θ一定要大于等于观测值x

中国科学院大学 University of Chinese Academy of Sciences

3.6 贝叶斯估计：一个例子

- **一个例子**
  - **When the next data point x_2= 7 arrives, we have**

$$p(\theta|D^2) \propto p(x_2|\theta)p(\theta|D^1) = \frac{1}{\theta^2} \quad \Rightarrow p(\theta|D^2) \propto \begin{cases} 1/\theta^2, & 7 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

  - **When the next data point x_3= 2 arrives, we have**

$$p(\theta|D^3) \propto p(x_3|\theta)p(\theta|D^2) = \frac{1}{\theta^3} \quad \Rightarrow p(\theta|D^3) \propto \begin{cases} 1/\theta^3, & 7 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

  - **When the next data point x_4= 8 arrives, we have**

$$p(\theta|D^4) \propto p(x_4|\theta)p(\theta|D^3) = \frac{1}{\theta^4} \quad \Rightarrow p(\theta|D^4) \propto \begin{cases} 1/\theta^4, & 8 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

……

  - **When data point x_n arrives, we have**

$$p(\theta|D^n) \propto p(x_n|\theta)p(\theta|D^{n-1}) = \frac{1}{\theta^n} \quad \Rightarrow p(\theta|D^n) \propto \begin{cases} 1/\theta^n, & \max\{D^n\} \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

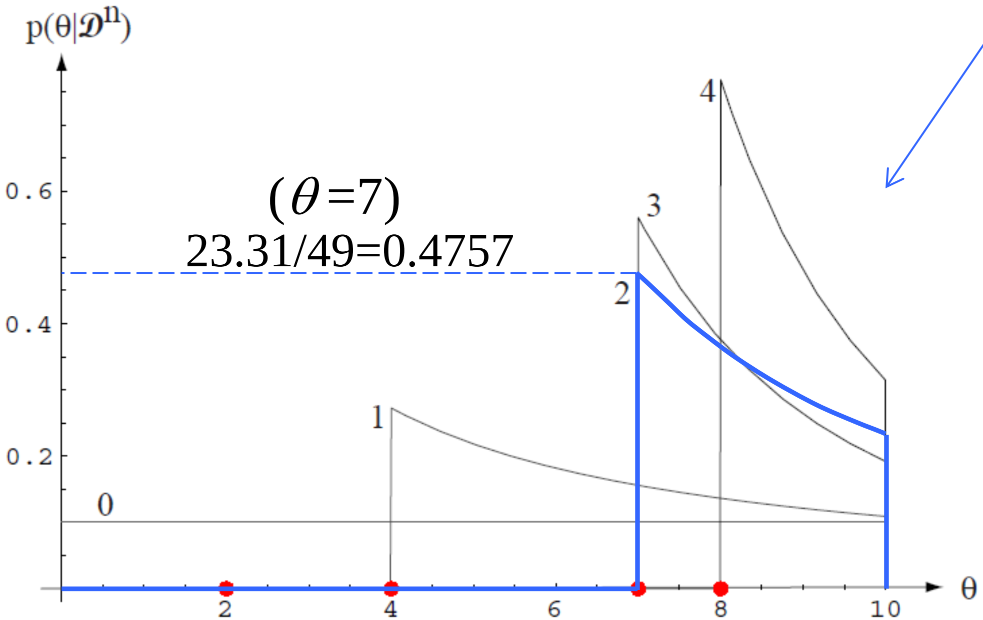
比如：

$$p(\theta \mid D^2) = \alpha \times \begin{cases} 1/\theta^2, & 7 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

$$\alpha = \frac{1}{\int_7^{10} \frac{1}{\theta^2} d\theta} = \frac{1}{1/7 - 1/10} = 23.3100$$

- **一个例子**
  - 关于参数θ的分布 的调整过程：

D = { 4, 7, 2, 8 }

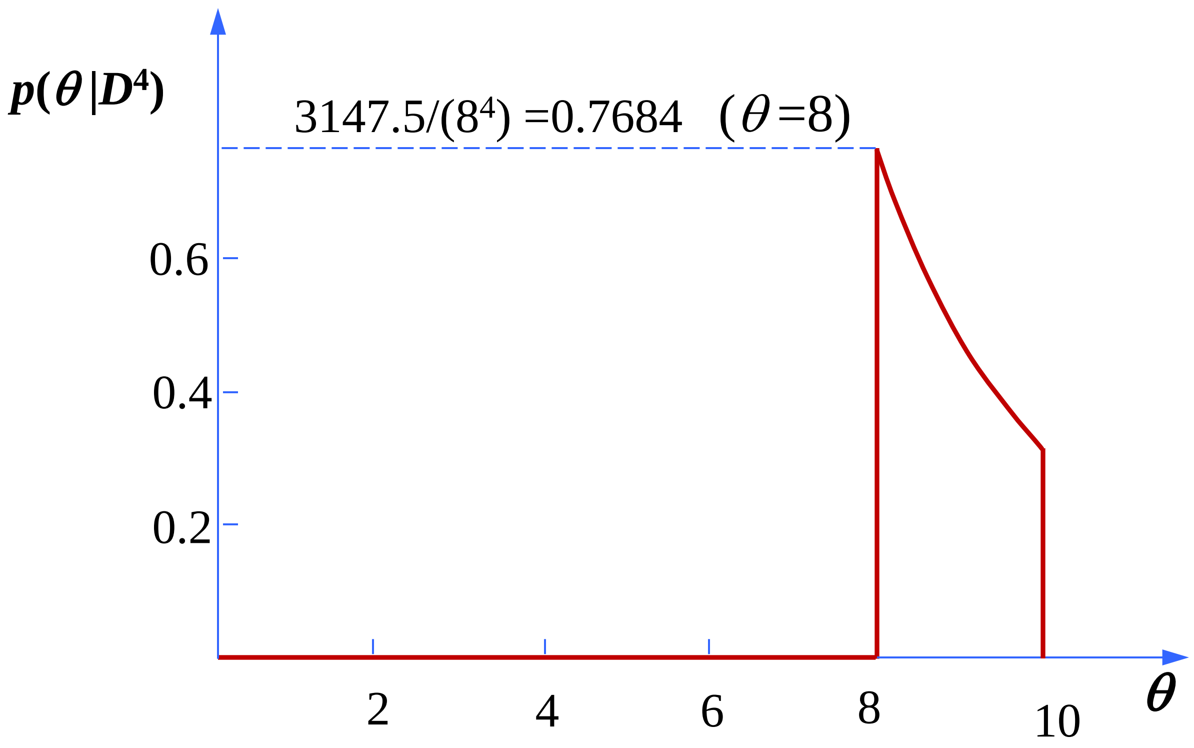


$$p(\theta \mid D^4) \propto \begin{cases} 1/\theta^4, & 8 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

$$p(\theta \mid D^3) \propto \begin{cases} 1/\theta^3, & 7 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

$$p(\theta \mid D^0) \propto \begin{cases} 1/10, & 0 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases} \Rightarrow p(\theta \mid D^1) \propto \begin{cases} 1/\theta, & 4 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases} \Rightarrow p(\theta \mid D^2) \propto \begin{cases} 1/\theta^2, & {\color{red}7} \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

- **一个例子**
  - 参数θ的最后估计结果：



$$p(\theta \mid D^4) = \alpha \times \begin{cases} 1/\theta^4, & 8 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases} \qquad \alpha = \frac{1}{\int_8^{10} \frac{1}{\theta^4} d\theta} = \frac{1}{\frac{1}{3}\left(\frac{1}{8^3} - \frac{1}{10^3}\right)} = 3147.5$$

- **一个例子**
  - 样本的后验分布

$$p(\theta \mid D^4) = \begin{cases} 3147.5/\theta^4, & 8 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

$$p(\mathbf{x} \mid D) = \int_\theta p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

$$\boxed{p(x \mid D) = \begin{cases} 0.1134, & 0 \le x \le 8 \\ 786.875\left(\dfrac{1}{x^4} - \dfrac{1}{10^4}\right), & 8 < x \le 10 \end{cases}}$$

中国科学院大学 University of Chinese Academy of Sciences

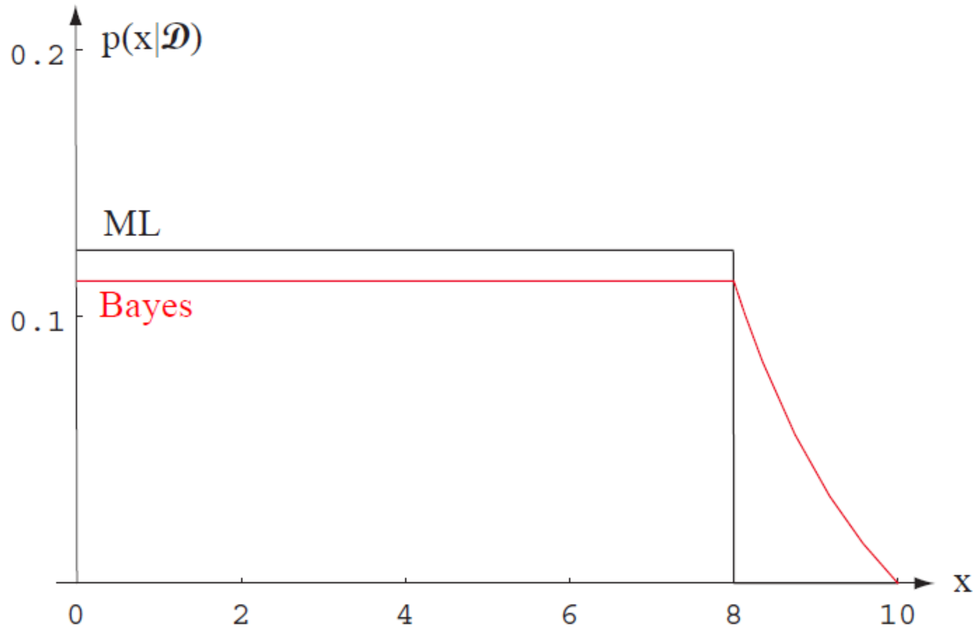
3.6 贝叶斯估计：一个例子

- **一个例子**
  - 我们来看看其最大似然估计，对于数据，其似然函数为：

$$l(\theta) = p(x_1, x_2, x_3, x_4 \mid \theta) = \frac{1}{\theta^4}$$

显然， $l(\theta)$ 单调递减，θ 越小，$l(\theta)$ 越大。但同时，θ 一定要大于等于最大观测数据。在现有样本**{4,7,2,8}**中，使似然函数$l(\theta)$取值最大的θ只能等于8。所以由于是均匀分布，所以θ的最大似然估计值为8。

- **一个例子**
  - **样本的后验分布**



Whereas the maximum-likelihood approach estimates a point in θ space, the Bayesian approach instead estimates a distribution. This figure illustrates the difference of these estimations finally on the data density.

3.7 特征维数问题

- 模式分类与特征的关系
  - 贝叶斯决策(0-1损失)：$\omega^* = \arg\max_j p(\omega_j|x)$
  - 特征空间给定时，贝叶斯分类错误率就确定了，即分类性能的理论上限就确定了（与分类器、学习算法无关）
- 增加特征有什么好处
  - 判别性：类别间有差异的特征有助于分类
- 带来什么问题
  - 泛化性能，Overfitting
  - 计算、存储

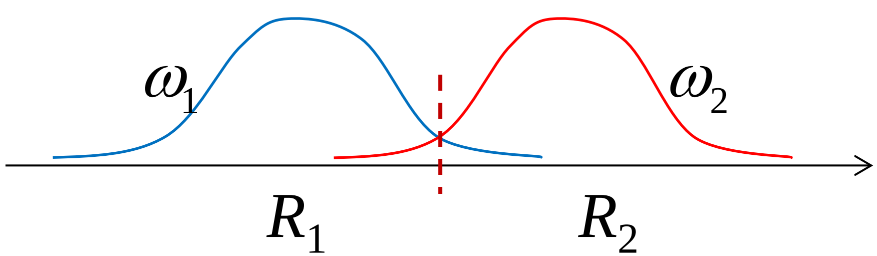
3.7 特征维数问题：分类错误率与特征的关系

- 高斯分布（两类问题）：
  - $p(\mathbf{x}|\omega_j)\sim N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma})$, j=1,2, 等协方差矩阵
  - Bayes error rate

$$\begin{aligned}P(error) &= P(\mathbf{x}\in R_2,\omega_1)+P(\mathbf{x}\in R_1,\omega_2)\\ &= \int_{R_2} p(\mathbf{x}|\omega_1)P(\omega_1)d\mathbf{x}+\int_{R_1} p(\mathbf{x}|\omega_2)P(\omega_2)d\mathbf{x}\\ &= P(\mathbf{x}\in R_2|\omega_1)P(\omega_1)+P(\mathbf{x}\in R_1|\omega_2)P(\omega_2)\end{aligned}$$

$$\Downarrow$$

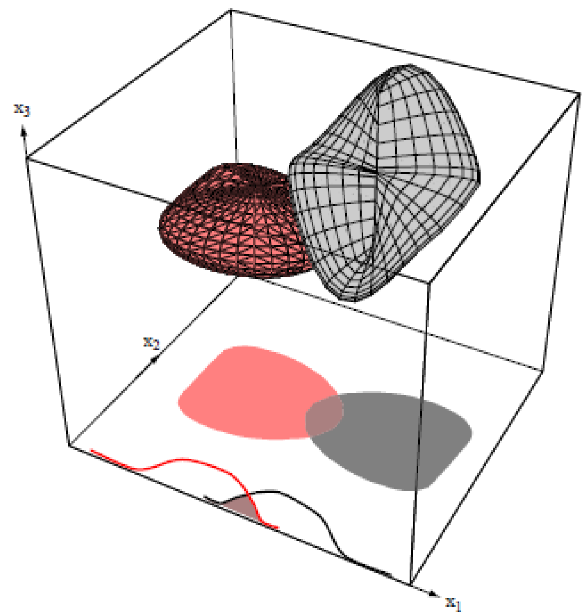
$$P(error)=\frac{1}{\sqrt{2\pi}}\int_{r/2}^{+\infty} e^{-\mu^2}d\mu,\qquad r^2=(\mathbf{u}_1-\mathbf{u}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{u}_1-\mathbf{u}_2)$$



- 高斯分布（两类问题）：
  - Conditionally independent case $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_1^2, \ldots, \sigma_d^2)$
    - 每一维的二类均值之间距离反映区分度，从而决定错误率
    - 特征增加有助于减小错误率 (因为 r^2 增大)

$$r^2 = \sum_{i=1}^{d} \frac{(\mu_{i1} - \mu_{i2})^2}{\sigma_i^2}, \qquad r^2 = (\mathbf{u}_1 - \mathbf{u}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{u}_1 - \mathbf{u}_2)$$

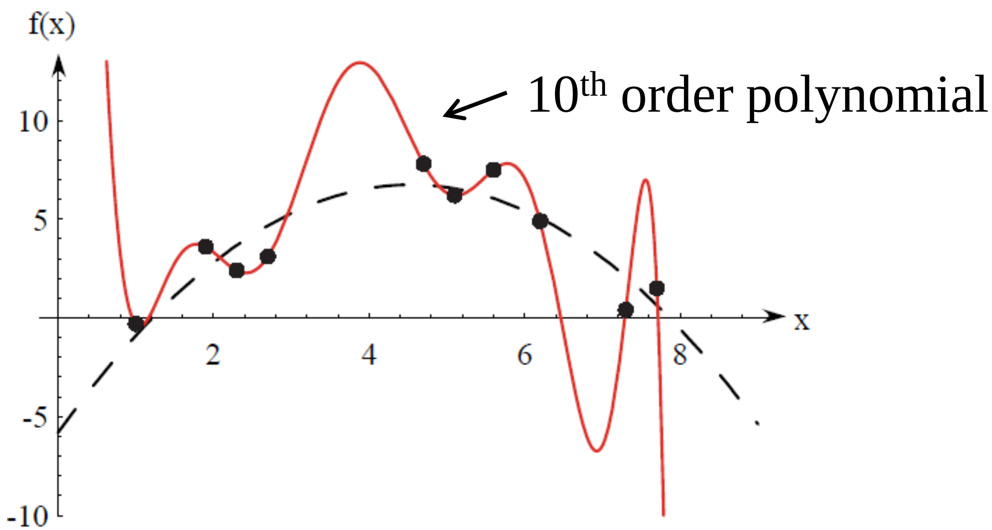
- 特征维数决定可分性的例子
  - 3D空间完全可分
  - 2D和1D投影空间有重叠



然而，增加特征也可能导致分类性能更差，因为有模型估计误差 (wrong model)

3.7 特征维数问题：过拟合(Overfitting)

- 过拟合
  - 特征维数高、训练样本少导致模型参数估计不准确
    - 比如协方差矩阵需要样本数在d以上
- 过拟合的例子



$$f(x) = ax^2 + bx + c + \varepsilon, \quad \text{where } p(\varepsilon) = \mathcal{N}(0, \sigma^2)$$

完美拟合训练数据，但测试误差很大

3.7 特征维数问题：过拟合(Overfitting)

- 克服办法
  - 特征降维：特征提取(变换)、特征选择
  - 参数共享/平滑
    - 方法一：共享协方差矩阵$\boldsymbol{\Sigma}_0$
    - 方法二：Shrinkage (a.k.a. Regularized Discriminant Analysis)：

第i类的协方差矩阵：

$$\boldsymbol{\Sigma}_i(\alpha)=\frac{(1-\alpha)n_i\boldsymbol{\Sigma}_i+\alpha n\boldsymbol{\Sigma}}{(1-\alpha)n_i+\alpha n},$$

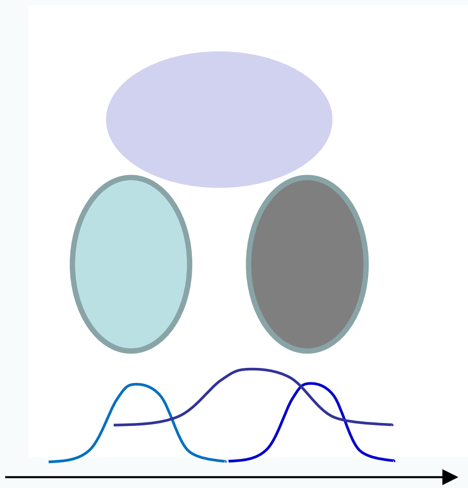
(启发式方法)

$$\boldsymbol{\Sigma}(\beta)=(1-\beta)\boldsymbol{\Sigma}+\beta\mathbf{I}$$

用第i类数据

用所有数据

扩展：开放集分类的特征维数问题



- 开放集分类问题
  - 已知类别： $\omega_i, i=1,\ldots,c$
  - 后验概率 $\sum_{i=1}^{c+1} P(\omega_i \mid \mathbf{x}) = 1$
  - ω_{c+1}无训练样本，测试样本作为outlier拒识
- 特征维数问题
  - 区分c+1个类别比区分c个类别需要更多的特征
  - 如果分类器训练时瞄准区分c个已知类别
    测试时易造成outlier 与 已知类别样本的混淆
  - 因此，在c类样本上训练分类器时，要使特征表达具有区分更多类别的能力：
    - 比如，训练神经网络时加入数据重构损失(类似auto-encoder)作为正则项
      比如，生成一些假想类样本(通过组合已知类别样本)
    - …

下次课内容

- 第3章
  - 期望最大法
  - 隐马尔可夫模型

致谢

- PPT由向世明老师提供

Thank All of You!

(Questions?)

张燕明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

中科院自动化研究所· 模式识别国家重点实验室