模式识别

张志飞

zhifeizhang@tongji.edu.cn

第三章 最大似然估计和贝叶斯估计

3.1 引言

监督参数估计

3.2 最大似然估计

3.3 贝叶斯估计

3.4 无监督参数估计

3.5 期望最大算法

3.1 引言

- **背景：**

实际分类问题的概率结构完整信息很难获知，通常仅知总体分布的模糊信息及训练样本。

需要用训练样本估计先验概率和类条件概率密度。

如何估计先验概率？ $\hat{P}(\omega_i) = \frac{N_i}{N}$，$N_i$是训练集N个样本中$\omega_i$类样本数。

如何用训练样本估计类条件概率密度？

类条件密度

贝叶斯公式

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

总括

概率密度函数的估计方法分两大类：

- 两大类方法
  - 参数估计（函数形式已知，但参数未知）
    - 最大似然估计
    - 贝叶斯估计和贝叶斯学习
  - 非参数估计（函数形式和参数均未知）
    - Parzen窗估计
    - Kn-近邻估计

基本概念

- **统计量：** 针对不同要求构造出样本的某种函数，这种函数在统计学种称为统计量。
- **参数空间：** 参数估计中，总体分布的概率密度函数形式已知，而参数未知，记为θ，在统计学中，将θ的全部容许值组成的集合称为参数空间，记为Θ。
- **估计量：** 构造一个统计量$d(X_1,\cdots,X_n)$作为参数θ的估计$\hat{\theta}$，在统计学中称$\hat{\theta}$为θ的估计量。d是观测向量$(X_1,\cdots,X_n)$的函数。
- **无偏性**
  - 若$E_\theta[\hat{\theta}_n]=\theta$，则称$\hat{\theta}_n$是无偏的；
  - 若$\lim\limits_{n\to\infty} E_n[\hat{\theta}_n]=\theta$，则称$\hat{\theta}_n$渐进无偏。即，样本数趋于无穷时才具有无偏性。
- **有效性：** 若一种估计的方差比另一种估计的方差小，则称方差小的估计更有效。
- **一致估计：** 如果对于任意给定的正数ε，总有$\lim\limits_{n\to\infty} P(|\hat{\theta}_n-\theta|>\varepsilon)=0$，则称$\hat{\theta}$是$\theta$的一致估计。

同济大学 TONGJI UNIVERSITY

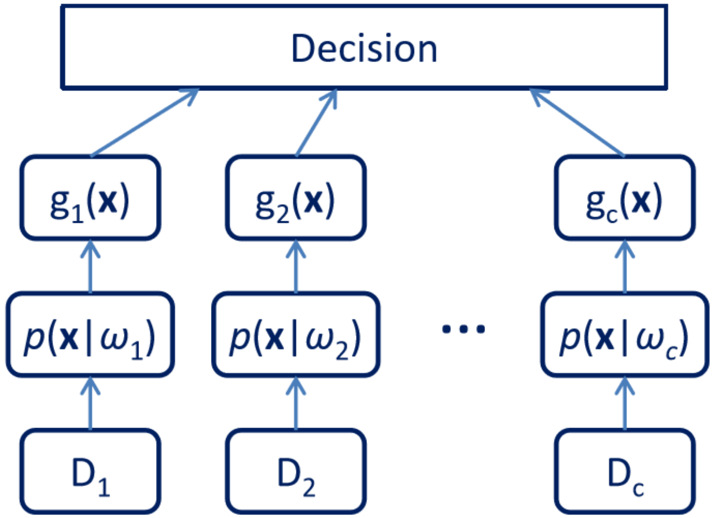
3.2 最大似然估计

1. 基本假设

类条件概率密度$p(X|\omega_j)$函数形式已知，参数**未知但确定**，记作θ_j，将$p(X|\omega_j)$改写为$p(X|\omega_j,\theta_j)$或$p(X|\theta_j)$，j=1,2,...,c。

- 每类样本集D_j中的样本都是从密度为$p(X|\omega_j)$的总体中独立抽出，即D_j中的**样本是独立同分布的**。
- 各类样本只包含本类的分布信息，即不同类别的**参数θ_j是各自独立的**。

分而治之！



在独立性假设下，可将原问题看作c个独立的问题。即，每一类独立地按照概率密度$p(X|\theta)$抽取样本集D，用D估计出参数θ。

2. 基本原理

- $\mathcal{D}=\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$，设各样本按条件概率密度$p(\mathbf{x}|\theta)$从总体中独立抽取，有

$$p(\mathcal{D}|\theta)=p(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n|\theta)=\prod_{k=1}^{n}p(\boldsymbol{X}_k|\theta)$$

将 $p(\mathcal{D}|\theta)$称作**参数$\boldsymbol{\theta}$**相对于样本集$\mathcal{D}$的**似然函数**。

- 样本集确定后，上述函数仅为θ的函数。它反映的是在不同参数取值下取得当前样本集的可能性。
- **似然**的本意：基于已知观测，θ取什么值可使观测值最可能出现。
- **最大似然估计**是使似然函数$p(\mathcal{D}|\theta)$达到最大的参数值$\hat{\theta}$.

$$\boldsymbol{\hat{\theta}}=\arg\max_{\boldsymbol{\theta}}\boldsymbol{p}(\mathcal{D}|\boldsymbol{\theta})$$

同济大学 TONGJI UNIVERSITY

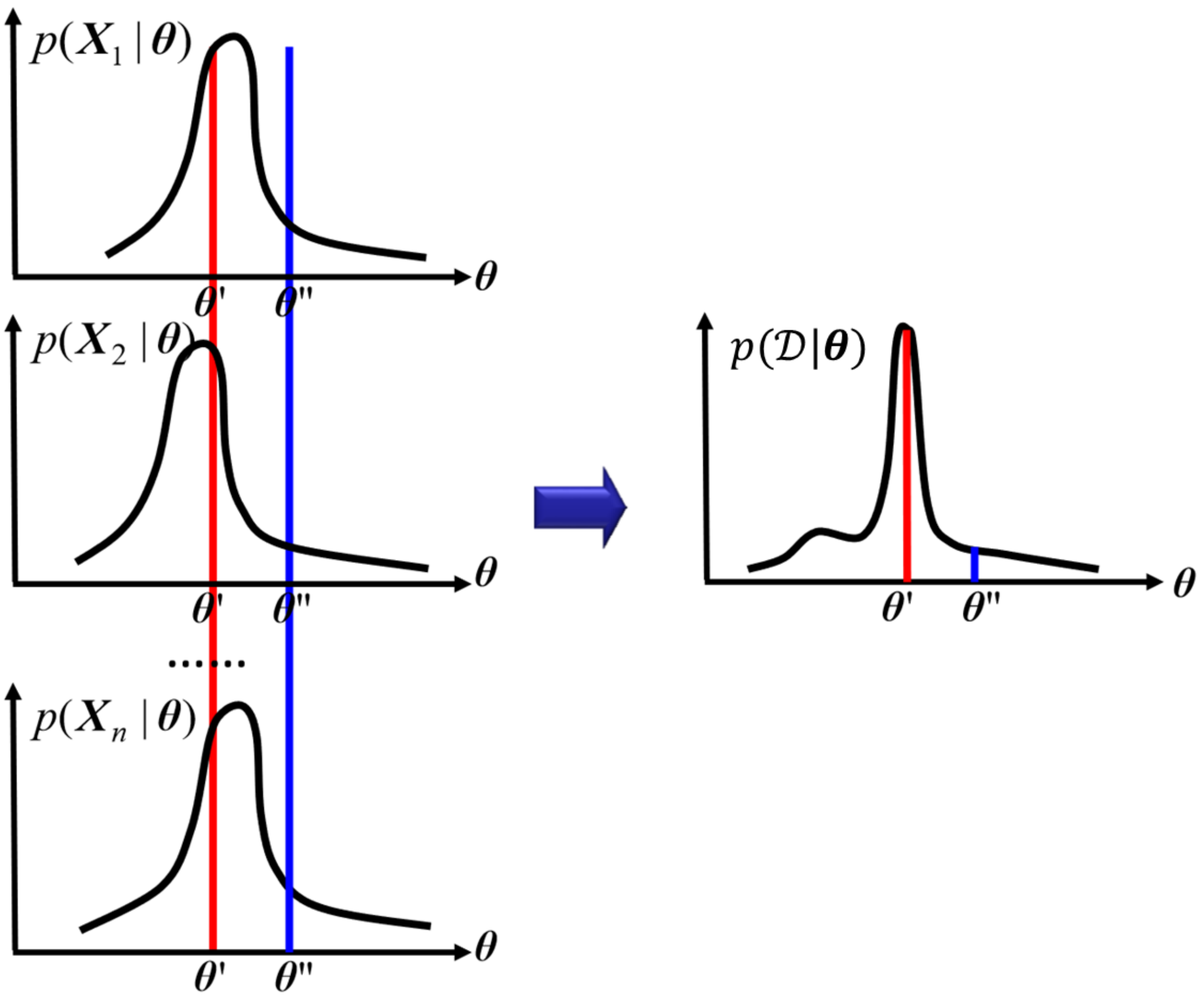


图3.1 最大似然估计示意图

3. 微分求解法

当似然函数$\boldsymbol{p(\mathcal{D}|\theta)}$是$\boldsymbol{\theta}$的可微函数时，可用微分法求解$\hat{\theta}$。

最大似然估计的求解

- 样本集 $\mathcal{D}=\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\}$
- 似然函数 $p(\mathcal{D}|\theta)=p(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n|\theta)=\prod_{k=1}^{n}p(\boldsymbol{X}_k|\theta)$
- 对数似然函数 $l(\theta)=\ln p(\mathcal{D}|\theta)=\sum_{k=1}^{n}\ln p(\boldsymbol{X}_k|\theta)$
- $\nabla_{\boldsymbol{\theta}}l(\theta)=\frac{\partial}{\partial\boldsymbol{\theta}}\ln p(\mathcal{D}|\theta)=\sum_{k=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ln p(\boldsymbol{X}_k|\theta)=0$
- 求解$\nabla_{\theta}l(\theta)=0$ 的方程，得到参数$\boldsymbol{\theta}$的最大似然估计。

例1 单个未知变量

朱丽叶在任何约会中都可能迟到，迟到时间记为随机变量X，服从$[0,\theta]$上的均匀分布，参数θ是未知的。经过N次观测，发现朱丽叶迟到时间的**最大值为X'**。求θ的最大似然估计

若$\theta_1 \le X \le \theta_2$呢？

θ的似然函数为，

$$p(X|\theta)=\begin{cases} 1/\theta, & 若\ 0 \le X \le \theta \\ 0, & 其他 \end{cases}$$

θ 越小，似然函数越大。而N次观测的样本集中 X' 为最大值。显然θ不能小于X'，因此，θ的最小可能值是X'，这时，θ的最大似然估计量为$\hat{\theta}=X'$。

此例说明，并不是所有的概率密度形式都可以用微分求解法。
造成此种困难的原因是似然函数在最大值处没有零斜率。

例2 正态分布的ML参数估计(情况1: μ未知)

设某一随机向量服从正态分布，其协方差矩阵Σ已知，但均值向量 μ 未知，试确定 μ 的最大似然估计。

解： 设有n 个观测$\mathbf{x}_1, \ldots, \mathbf{x}_n$，用来估计均值向量$\mu$。

对于观测$\mathbf{x}_k$ (d维)，$k = 1, \ldots, n$，其对数似然函数为，

$$\ln p(\mathbf{x}_k|\mu) = -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(\mathbf{x}_k - \mu)^T\Sigma^{-1}(\mathbf{x}_k - \mu)$$

求导，$\nabla_\mu \ln p(\mathbf{x}_k|\mu) = \Sigma^{-1}(\mathbf{x}_k - \mu)$

n个观测，$\nabla_\mu \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\mu) = \sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \mu)$

最大似然估计应是方程 $\sum_{K=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \hat{\mu}_{ML}) = 0$ 的解。

得，$\hat{\mu}_{ML} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$

总体均值的最大似然估计是样本均值。几何意义,均值向量是云团质心。

同濟大學 TONGJI UNIVERSITY

例3 正态分布的ML参数估计(情况2: μ和Σ未知)

考虑单变量的情况。待估计参数向量$\boldsymbol{\theta} = (\theta_1, \theta_2)$，$\theta_1 = \mu$，$\theta_2 = \sigma^2$。

解：有 $\boldsymbol{n}$ 个观测$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$，其中$\boldsymbol{x}_k$的对数似然函数为，

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln[2\pi\theta_2] - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

其偏导，

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

对总的对数似然函数求偏导，并令其等于零，有

$$\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}(x_k-\hat{\theta}_1)=0 \quad 和 \quad -\sum_{k=1}^{n}\frac{1}{\hat{\theta}_2}+\sum_{k=1}^{n}\frac{(x_k-\hat{\theta}_1)^2}{\hat{\theta}_2^2}=0$$

用$\hat{\mu}_{\mathrm{ML}}=\hat{\theta}_1$，$\hat{\sigma}_{\mathrm{ML}}^2=\hat{\theta}_2$ 替换后，

$$\boxed{\hat{\mu}_{\mathrm{ML}}=\frac{1}{n}\sum_{k=1}^{n}x_k} \qquad \boxed{\hat{\sigma}_{\mathrm{ML}}^2=\frac{1}{n}\sum_{k=1}^{n}(x_k-\widehat{\mu}_{\mathrm{ML}})^2}$$

- **推广到多变量情况**，有估计向量$\hat{\boldsymbol{\mu}}$ 和 $\widehat{\boldsymbol{\Sigma}}$

$$\boxed{\hat{\boldsymbol{\mu}}_{\mathrm{ML}}=\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k} \qquad \boxed{\widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}=\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\mu}_{\mathrm{ML}})(\mathbf{x}_k-\hat{\mu}_{\mathrm{ML}})^T}$$

最大似然估计量的偏差

- $N(\mu,\Sigma)$均值和方差的最大似然估计 $\hat{\mu}_{ML}$和$\hat{\Sigma}_{ML}$，

$$E[\hat{\boldsymbol{\mu}}_{ML}] = E\left[\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\right] = \boldsymbol{\mu} \qquad E[\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}] = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

- $\hat{\mu}_{ML}$是无偏的，$\hat{\Sigma}_{ML}$是有偏但渐进无偏的。
- $\boldsymbol{\Sigma}$的无偏估计:

$$\tilde{\Sigma} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

3.3 贝叶斯估计

贝叶斯估计基本思想：

贝叶斯估计方法与最大似然估计方法有本质不同，它把参数向量$\boldsymbol{\theta}$本身看成一个随机变量，根据观测数据对参数的分布进行估计，即后验概率密度$p(\theta|\mathcal{D})$。

贝叶斯学习，则是把贝叶斯估计的原理用于直接从数据对概率密度函数进行迭代估计。

概率密度估计与参数分布

- **原问题：** 估计概率密度。假设$p(\mathbf{x}|\theta)$函数形式已知，参数θ未知且不固定；
- **目标：** 根据给定的样本集$\mathcal{D}=\{X_1,\ldots,X_n\}$，找到未知参数$\theta$的一个估计量，使得由此带来的风险最小。

$$p(\mathbf{x}|\mathcal{D})=\int p(\mathbf{x},\theta|\mathcal{D})d\theta=\int p(\mathbf{x}|\theta,\mathcal{D})p(\theta|\mathcal{D})d\theta=\int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})d\theta$$

贝叶斯估计最核心公式

$$p(\mathbf{x}|\mathcal{D})=\int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})d\theta$$

这一关键等式将概率密度$p(\mathbf{x}|\mathcal{D})$估计和参数$\theta$后验分布 $p(\theta|\mathcal{D})$联系起来。 在已知p(x|θ) 的条件下，训练样本可通过后验密度$p(\theta|\mathcal{D})$ 对$p(\mathrm{x}|\mathcal{D})$的估计施加影响。

同济大学 TONGJI UNIVERSITY

贝叶斯估计

适用范围： 概率密度函数形式已知，但参数未知且不固定。

方法： 用类似于**最小风险判决**的方法来估计未知随机参数。

假设： $\boldsymbol{\theta}$取值的参数空间$\boldsymbol{\Theta}$是一个连续空间；

用$\lambda(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta})$标记真实参数为$\boldsymbol{\theta}$，得到的估计量为$\widehat{\boldsymbol{\theta}}$时承担的损失。

样本X下的条件风险

$$R(\widehat{\boldsymbol{\theta}}|\boldsymbol{X})=\int_{\Theta}\lambda(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}$$

总的平均风险（贝叶斯风险）

$$\begin{aligned}\bar{R} &= \int_{E_d} R(\widehat{\boldsymbol{\theta}}|\boldsymbol{X})p(\boldsymbol{X})d\boldsymbol{X} \\ &= \int_{E_d} p(\boldsymbol{X})\int_{\Theta}\lambda(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}d\boldsymbol{X}\end{aligned}$$

$\boldsymbol{\theta}$的贝叶斯估计是使得贝叶斯风险最小化的估计量$\widehat{\boldsymbol{\theta}}$。

显然，贝叶斯风险与$\lambda(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta})$的选择有关！

最常用的损失函数是平方误差损失函数, 即, $\lambda(\hat{\theta}|\theta)=(\theta-\hat{\theta})^2$

定理3.1　关于贝叶斯估计量的定理

如果选择损失函数$\lambda(\hat{\theta}|\theta)=(\theta-\hat{\theta})^2$, 则

$$\hat{\theta}=E(\theta|X)=\int_{\Theta}\theta p(\theta|X)d\theta$$

[证明]

$$R(\hat{\theta}|X)=\int_{\Theta}\lambda(\hat{\theta}|\theta)p(\theta|X)d\theta=\int_{\Theta}(\theta-\hat{\theta})^2p(\theta|X)d\theta$$

$$=\int_{\Theta}(\theta-E(\theta|X)+E(\theta|X)-\hat{\theta})^2p(\theta|X)d\theta$$

$$=\int_{\Theta}(\theta-E(\theta|X))^2p(\theta|X)\,d\theta+\int_{\Theta}(E(\theta|X)-\hat{\theta})^2p(\theta|X)\,d\theta$$

$$+2\int_{\Theta}(\theta-E(\theta|X))(E(\theta|X)-\hat{\theta})p(\theta|X)d\theta$$

[证明]（续）

$$R(\hat{\theta}|X) = \int_{\Theta} (\theta - E(\theta|X))^2 p(\theta|X)\, d\theta + \int_{\Theta} (E(\theta|X) - \hat{\theta})^2 p(\theta|X)\, d\theta$$

$$+2\int_{\Theta} (\theta - E(\theta|X))(E(\theta|X) - \hat{\theta})p(\theta|X)d\theta$$

$$\int_{\Theta} (\theta - E(\theta|X))(E(\theta|X) - \hat{\theta})p(\theta|X)d\theta$$

$$= (E(\theta|X) - \hat{\theta})\int_{\Theta} (\theta - E(\theta|X))p(\theta|X)d\theta$$

$$= (E(\theta|X) - \hat{\theta})\left[\int_{\Theta} \theta p(\theta|X)d\theta - E(\theta|X)\int_{\Theta} p(\theta|X)d\theta\right]$$

$$= (E(\theta|X) - \hat{\theta})(E(\theta|X) - E(\theta|X)) = 0$$

[证明] (续)

$$R(\hat{\theta}|X)=\int_{\Theta}(\theta-E(\theta|X))^2 p(\theta|X)\,d\theta+\int_{\Theta}(E(\theta|X)-\hat{\theta})^2 p(\theta|X)\,d\theta$$

易见：第一项非负且其取值与$\hat{\theta}$无关；

第二项也非负，但其取值与$\hat{\theta}$有关。

故：欲使贝叶斯风险最小化，需选择估计量使第二项最小化。

即 $\hat{\theta}=E(\theta|X)=\int_{\Theta}\theta p(\theta|X)d\theta$ 证毕

结论：以上定理给出了估计待求参数的方法。

$$\hat{\theta} \Rightarrow p(\theta|X) \Rightarrow p(\theta|\mathcal{D})=\frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}=\frac{p(\mathcal{D}|\theta)p(\theta)}{\int_{\Theta}p(\mathcal{D}|\theta)p(\theta)d\theta}$$

贝叶斯估计步骤

在已知训练样本集$\mathcal{D}=\{X_1,\ldots,X_n\}$的情况下，用$p(\theta|\mathcal{D})$作为$p(\theta|X)$的估计。

$p(\theta|\mathcal{D})$和估计量$\hat{\theta}$通过如下步骤获得：

(1) 确定θ的先验概率密度$p(\theta)$.

(2) 根据$p(\mathcal{D}|\theta)=p(X_1,X_2,\cdots,X_n|\theta)=\prod_{k=1}^{n}p(X_k|\theta)$,由训练样本$\mathcal{D}$求出以$\theta$为参数的联合概率密度$p(\mathcal{D}|\theta)$.

(3) 利用贝叶斯公式，求出θ的后验概率密度

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

(4) 用$p(\theta|\mathcal{D})$替代$p(\theta|X)$，计算$\hat{\theta}=\int_{\Theta}\theta p(\theta|\mathcal{D})d\theta$.

一旦得到θ的贝叶斯估计$\hat{\theta}$，待求类条件概密可随之确定。

贝叶斯学习

不经过参数估计，直接根据样本集推断总体的概率分布。

在独立性假设下，采取分而治之的策略，只需要考虑给定 $\mathcal{D}=\{X_1,X_2,\cdots,X_n\}$，如何推断总体$X$的后验概密 $p(X|\mathcal{D})$?

求解思路：通过 $p(\theta|\mathcal{D})$，求解 $p(X|\mathcal{D})$。

设 $p(X,\theta)$为X和θ的联合概率密度，则

在输入样本集$\mathcal{D}$给定的条件下，$p(X|\theta,\mathcal{D})$仅与θ有关

收敛?

$$p(X)=\int_\Theta p(X,\theta)d\theta \Rightarrow p(X|\mathcal{D})=\int_\Theta P(X,\theta|\mathcal{D})d\theta$$

$$p(X,\theta)=p(X|\theta)p(\theta) \Rightarrow p(X,\theta|\mathcal{D})=p(X|\theta,\mathcal{D})p(\theta|\mathcal{D})$$

$$p(X|\mathcal{D})=\int_\Theta p(X|\theta,\mathcal{D})p(\theta|\mathcal{D})\,d\theta=\int_\Theta p(X|\theta)p(\theta|\mathcal{D})\,d\theta$$

讨论：$p(X|\mathcal{D})$是否收敛于$p(X)$?

引入标记$\mathcal{D}^n = \{X_1, \cdots, X_n\}$,

假设样本集中的各样本是独立抽取的，则n>1时，有

$$p(\mathcal{D}^n|\theta) = p(X_n|\theta)p(X_{n-1}|\theta)\cdots p(X_2|\theta)p(X_1|\theta) = p(X_n|\theta)p(\mathcal{D}^{n-1}|\theta)$$

$$p(\theta|\mathcal{D}^n) = \frac{p(\mathcal{D}^n|\theta)p(\theta)}{\int_\Theta p(\mathcal{D}^n|\theta)p(\theta)d\theta}$$ 贝叶斯公式

$$= \frac{p(X_n|\theta)\boxed{p(\mathcal{D}^{n-1}|\theta)p(\theta)}}{\int_\Theta p(X_n|\theta)\boxed{p(\mathcal{D}^{n-1}|\theta)p(\theta)}d\theta}$$ 独立性

$$p(\mathcal{D}^{n-1}|\theta)p(\theta) = p(\mathcal{D}^{n-1})p(\theta|\mathcal{D}^{n-1})$$

$$= \frac{p(X_n|\theta)p(\mathcal{D}^{n-1})p(\theta|\mathcal{D}^{n-1})}{\int_\Theta p(X_n|\theta)p(\mathcal{D}^{n-1})p(\theta|\mathcal{D}^{n-1})d\theta}$$ 贝叶斯公式

$$= \frac{p(X_n|\theta)p(\theta|\mathcal{D}^{n-1})}{\int_\Theta p(X_n|\theta)p(\theta|\mathcal{D}^{n-1})d\theta}$$ 约去$p(\mathcal{D}^{n-1})$

同济大学 TONGJI UNIVERSITY

实现参数θ在线学习的递推公式

$$p(\theta|\mathcal{D}^n) = \frac{p(X_n|\theta)p(\theta|\mathcal{D}^{n-1})}{\int_\Theta p(X_n|\theta)p(\theta|\mathcal{D}^{n-1})d\theta}$$

n=1时 $p(\theta|\mathcal{D}^0) = p(\theta)$

n=2时 $$p(\theta|\mathcal{D}^1) = \frac{p(X_1|\theta)P(\theta|\mathcal{D}^0)}{\int_\Theta p(X_1|\theta)P(\theta|\mathcal{D}^0)d\theta}$$

$\cdots\cdots$

n=n时 $$p(\theta|\mathcal{D}^n) = \frac{p(X_n|\theta)P(\theta|\mathcal{D}^{n-1})}{\int_\Theta p(X_n|\theta)P(\theta|\mathcal{D}^{n-1})d\theta}$$

$\cdots\cdots$

$$p(\theta), p(\theta|\mathcal{D}^1), p(\theta|\mathcal{D}^2), \cdots, p(\theta|\mathcal{D}^{n-1}), p(\theta|\mathcal{D}^n), \cdots$$

$\boldsymbol{p(\theta), p(\theta|\mathcal{D}^1), p(\theta|\mathcal{D}^2), \cdots, p(\theta|\mathcal{D}^{n-1}), p(\theta|\mathcal{D}^n), \cdots}$

随着n值的增加，θ的相应后验概率密度一般会变得越来越尖锐。

若上述概率密度函数序列在$n \to \infty$时, 收敛于以真值参数θ为中心的狄拉克δ函数，则称相应的学习过程为贝叶斯学习过程。

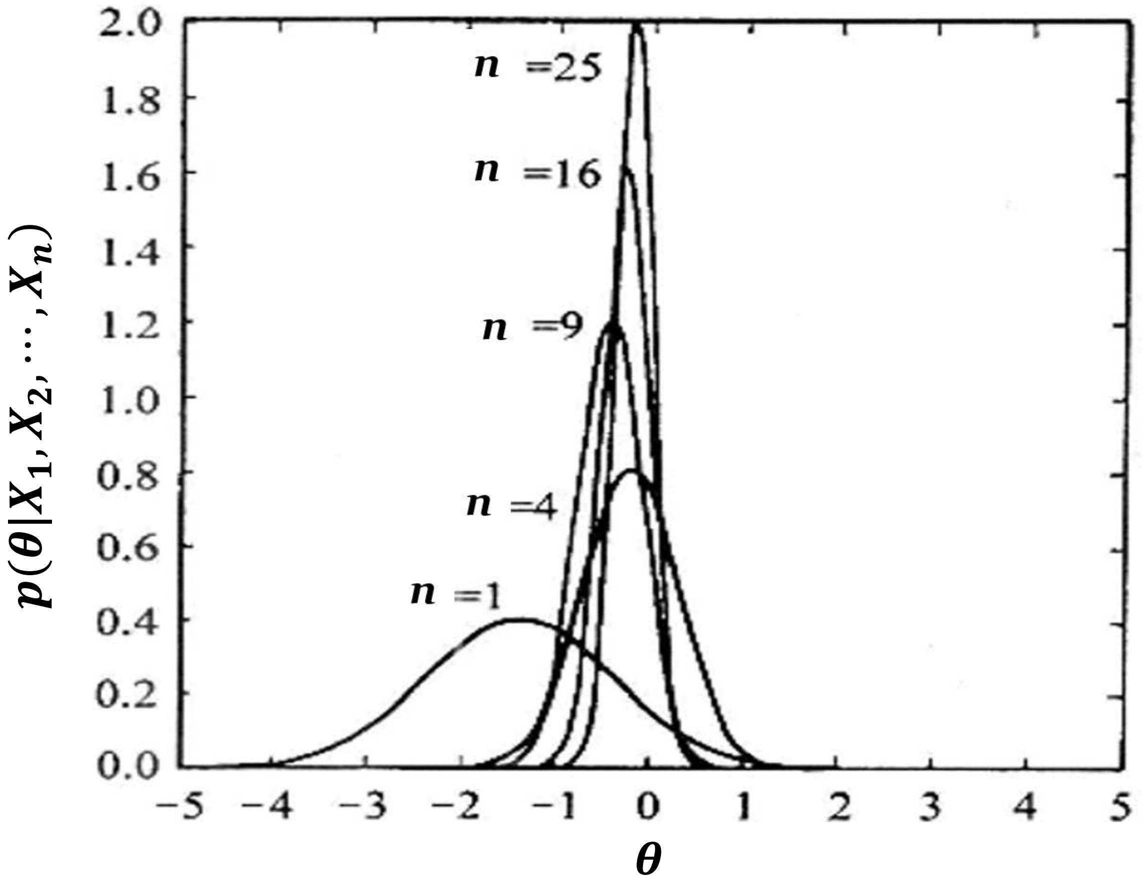


图3.2 贝叶斯学习过程示意图

$$\lim_{n\to\infty} p(\boldsymbol{X}|\mathcal{D}^n) = p(\boldsymbol{X}|\mathcal{D}^{n\to\infty}) = p\left(\boldsymbol{X}\middle|\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}\right) = p(\boldsymbol{X})$$

$\boldsymbol{p(X|\mathcal{D}^n)}$收敛于$\boldsymbol{p(X)}$。

例4 递归的贝叶斯学习

假设一维样本服从均匀分布

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

最初只知道θ有界，如$0\le\theta\le 10$ (无信息或“平”的先验概率，即$p(\theta)\sim U(0,10)$)。已有样本集$\mathcal{D}=\{4,7,2,8\}$，其中每一样本均依概率密度$p(x)$独立抽取。

如何用递归贝叶斯学习方法来估计θ和概率密度函数$p(x)$?

- 一开始，未有样本到达之前，$p(\theta|\mathcal{D}^0)=p(\theta)=U(0,10)$。
- 第一个样本$x_1=4$到达，得改善了的估计，

$$p(\theta|\mathcal{D}^1)\propto p(x_1|\theta)p(\theta|\mathcal{D}^0)=\begin{cases}1/\theta & 4\le\theta\le 10\\ 0 & \text{其他}\end{cases}$$

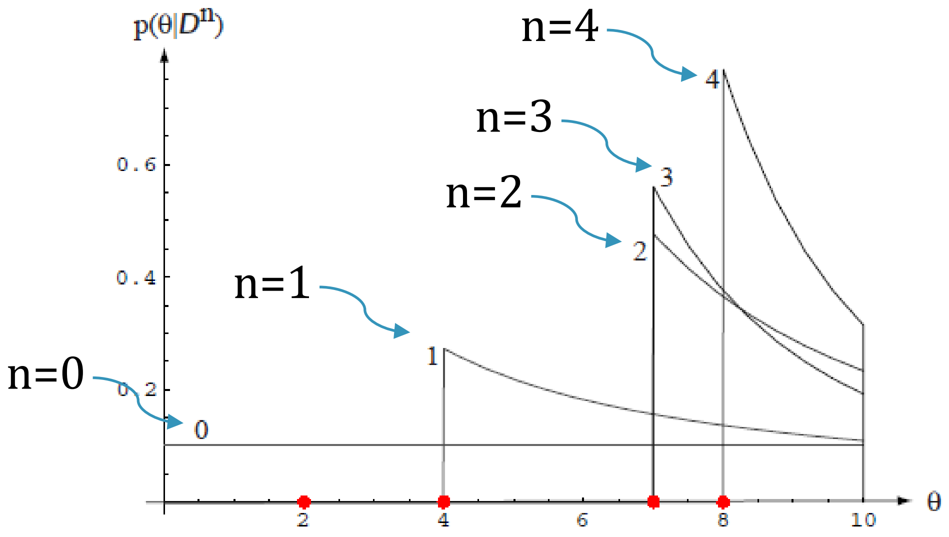
为简便，忽略归一化

- 第二个样本$x_2=7$到达，进一步改善估计，

$$p(\theta|\mathcal{D}^2)\propto p(x_2|\theta)p(\theta|\mathcal{D}^1)=\begin{cases}1/\theta^2 & 7\le\theta\le 10\\ 0 & \text{其他}\end{cases}$$

- 依此类推，

$$p(\theta|\mathcal{D}^n)\propto p(x_n|\theta)p(\theta|\mathcal{D}^{n-1})=\begin{cases}1/\theta^n & max[\mathcal{D}^n]\le\theta\le 10\\ 0 & \text{其他}\end{cases}$$



对于$\boldsymbol{\theta}$的估计，使用了题中全部样本后，

- 最大似然估计结果： $\hat{\theta} = 8$
- 贝叶斯估计结果：
  - **n=0，$p(\boldsymbol{\theta}|\mathcal{D}^n)$ 是位于0到10之间的均匀分布；**
  - 当更多样本加入后，$\boldsymbol{\theta}$后验密度在最大样本点处形成尖峰。

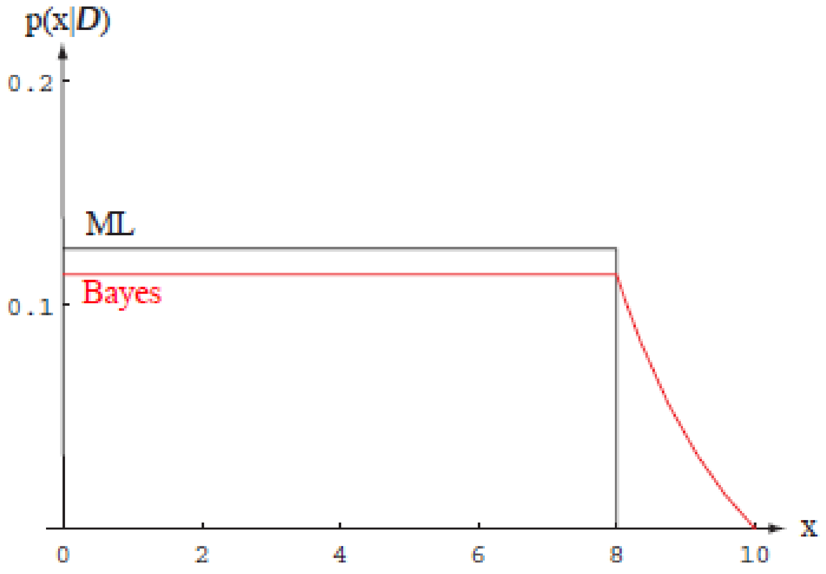


图3.3 最大似然估计与贝叶斯估计

对概率密度 $p(\mathbf{x}|\mathcal{D})$ 的估计，用题中的样本集，

- 最大似然法结果：$p(x|\mathcal{D}) \sim U(0,8)$均匀分布
- 贝叶斯法结果：
  - 在$0 \le x \le 8$ 是均匀分布，即$p(x|\mathcal{D}) \sim U(0,8)$
  - 在$8 \le x \le 10$ 有一个小的拖尾。表明先验信息影响仍在。

贝叶斯参数估计：高斯情况

- 计算后验密度$p(\theta|\mathcal{D})$
- 计算类条件密度$p(\mathbf{x}|\mathcal{D})$

单变量情况： $p(x|\mu) \sim N(\mu, \sigma^2)$，$\mu$是唯一未知参数

1. **计算后验密度分布：$p(\mu|\mathcal{D})$**

设μ的先验密度服从已知的高斯分布 $p(\mu) \sim N(\mu_0, \sigma_0^2)$，$\mu_0$和$\sigma_0^2$均已知。

设从密度为$p(\mu)$的总体中抽取一个μ, 则x的密度$p(x|\mu)$就确定了。从该总体中**独立**抽取n个样本$x_1, \cdots, x_n$，由贝叶斯公式

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

其中，α是一个依赖于样本集$\mathcal{D}$，而独立于μ的归一化系数。

因为$p(x_k|\mu)\sim N(\mu,\sigma^2)$, 和 $p(\mu)\sim N(\mu_0,\sigma_0^2)$ ，有

$$\begin{aligned}
p(\mu|\mathcal{D}) &= \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)} \\
&= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right] \\
&= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],
\end{aligned}$$

$p(\mu|\mathcal{D})$是一个指数函数，其指数部分是二次函数，故 $p(\mu|\mathcal{D})$是正态分布，且在样本数增加时仍保持正态分布，我们称这样的$p(\mu|\mathcal{D})$ 为复制密度函数，把其先验密度$p(\mu)$称为**共轭先验(conjugate prior)**。

定义3.1 设θ是总体分布中的参数，$p(\theta)$是θ的先验密度函数，假如由抽样信息算得的后验密度函数与$p(\theta)$有相同的函数形式，则称$p(\theta)$是θ的**共轭先验分布**。

若写成如下形式：$p(\mu|\mathcal{D}) \sim N(\mu_n, \sigma_n^2)$ ，也就是

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

令上面两式对应项相等，可求得 μ_n, σ_n^2

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

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

先验知识和观测样本共同影响的$p(\mu|\mathcal{D})$分布。两者贡献之间的平衡取决于σ^2/σ_0^2，这个比值称为**决断因子**(dogmatism)

其中$\bar{x}_n$是样本均值。当$\sigma^2/\sigma_0^2 \neq \infty$，且获取足够的样本后，$\mu_0$，$\sigma_0^2$的精确假定就无关紧要了， μ_n 收敛于$\bar{x}_n$。

2. 计算类条件密度：$p(x|\mathcal{D})$

由得到的后验均值$p(\mu|\mathcal{D})$，可计算类条件密度如下

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu$$
$$= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu$$
$$= \frac{1}{2\pi\sigma\sigma_n}exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n)$$

其中，$f(\sigma,\sigma_n)=\int exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu$

$$p(x|\mathcal{D}) \propto \quad exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right],$$

因此，$p(x|\mathcal{D})$ 是一个正态分布 $p(x|\mathcal{D})\sim N(\mu_n,\sigma^2+\sigma_n^2)$

最后，$p(x|\mathcal{D})$就是类条件密度$p(x|\omega_j,\mathcal{D})$，结合先验概率$P(\omega_j)$，设计贝叶斯分类器所需全部概率信息已具备。

贝叶斯参数估计：一般理论

假设**前提**，

- 已知类条件密度的参数式$p(\mathbf{x}|\theta)$；
- 参数θ的先验密度$p(\theta)$包含关于θ全部先验知识；
- 其余关于θ的信息都包含在独立观察样本$\mathbf{x}_1,\cdots,\mathbf{x}_n$中,这些样本服从未知的概率密度函数$p(\mathbf{x})$.

贝叶斯估计的**最基本问题**就是：

1. 计算后验密度函数$p(\theta|\mathcal{D})$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$ ，其中 $p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\theta)$

2. 计算类条件密度$p(\mathbf{x}|\mathcal{D})$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})d\theta$$

最大似然估计 与 贝叶斯估计

区别	最大似然估计	贝叶斯估计
计算复杂度	微分运算 √	多重积分运算
可解释性(interpretability)	• 是设计者提供的模型集合中的一个最佳模型; • 易于解释和理解	• 是多模型的加权平均，以反映对各种可行解答的不确定程度; • 过于复杂而难于理解
对先验信息的信任程度，如，初始假设的密度函数$p(x\|\theta)$的参数形式	结果的形式$p(x\|\hat{\theta})$与初始假设的参数式必然一致;	• 结果的形式可能与初始假设可能不同; • 利用全部 $p(\theta\|\mathcal{D})$ 分布信息，若信息可靠，则结果较ML更准确；在无特别先验信息情况下(如均匀分布)，二者效果类似。
所设计的贝叶斯分类器的分类误差	• 贝叶斯误差：无法消除 • 模型误差：选正确的模型 • 估计误差：增加训练样本数	• 贝叶斯误差：无法消除 • 模型误差：选正确的模型 • 估计误差：增加训练样本数

联系：最大似然估计可解释为具有均匀先验的最大后验概率估计。

当训练样本数趋于无穷大时，两者效果一致。

贝叶斯估计方法有很强的理论和算法基础。但在实际应用中，最大似然估计更简便，且设计出的分类器的性能几乎与贝叶斯方法得到的结果相差无几。

参数估计方法小结

VS

经典参数估计：最大似然估计(Maximum-likelihood Estimation,简称ML)

- 观点：
  - 将参数$\boldsymbol{\theta}$视为未知确定量。
  - 最好的估计$\hat{\boldsymbol{\theta}}$是使得观测值 $\boldsymbol{x}$ 最可能出现的估计。
- 参数的点估计

贝叶斯参数估计：

- 观点：
  - 将参数$\boldsymbol{\theta}$视为随机变量，其先验分布已知。利用观测值$\boldsymbol{x}$的信息，来修正先验认识，得到后验分布。
  - 好的估计应使得$\boldsymbol{\theta}$后验分布的峰值在$\boldsymbol{\theta}$的真值处
- 参数的分布估计


COFFEE
BREAK

3.4 无监督参数估计

背景：

- 把**参数估计方法推广**到概率模型中含有**隐变量**(如样本的未知类别)或允许样本存在**缺失特征**的情况。
- **样本类别未知(无监督)**情况下的类条件概密参数估计问题，被称为**无监督参数估计**。
  - 最大似然估计
  - 贝叶斯估计

无监督情况下的参数估计

问题描述：给定混合样本集 $\mathcal{D}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$，其类别数已知(c)，样本的标签未知。每个类别的类条件概率密度 $p(\boldsymbol{X}|\omega_i,\theta_i)$ 函数形式已知，$P(\omega_i)$未知。

目标：估计各类的分布参数θ_i 和$P(\omega_i)$，i=1,⋯,c 。令

$\theta=\{\theta_1,\cdots,\theta_c\}$，$P=\big(P(\omega_1),\cdots,P(\omega_c)\big)$，$\Theta=(\theta,P)$.

求解方法：**不分彼此，混合作战！**

混合样本集合 $\mathcal{D}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$

混合概率密度 $p(X)=\sum_{i=1}^{c} p(\boldsymbol{X}|\omega_i,\boldsymbol{\theta}_i)\,P(\omega_i)$

混合参数

分量概率密度

混合概率密度的最大似然估计

混合样本集$\boldsymbol{\mathcal{D}}$的似然函数：

$$p(\mathcal{D}|\boldsymbol{\theta}) = p(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{X}_k|\boldsymbol{\theta})$$

混合样本集$\boldsymbol{\mathcal{D}}$关于$\boldsymbol{\theta}$的对数似然函数：

$$L(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\left[\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i) P(\omega_i)\right]$$

求混合概率密度的最大似然估计，分两种情况：

(1) 假定各混合参数 $P(\omega_i)$, i=1,...,c 已知

(2) 假定 $P(\omega_i)$, i=1,...,c 未知

(1) $P(\omega_i)$已知

$$\boldsymbol{L}(\boldsymbol{\theta}) = \ln \boldsymbol{p}(\mathcal{D}|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln \boldsymbol{p}(\boldsymbol{X}_k|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\left[\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)\, P(\omega_i)\right]$$

$$\nabla_{\boldsymbol{\theta}_i}\boldsymbol{L}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}_i} \sum_{k=1}^{n} \ln \boldsymbol{p}(\boldsymbol{X}_k) = \sum_{k=1}^{n} \frac{1}{\boldsymbol{p}(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i}\boldsymbol{p}(\boldsymbol{X}_k)$$

$$= \sum_{k=1}^{n} \frac{1}{\boldsymbol{p}(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i}\left(\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)\, P(\omega_i)\right)$$

$$= \sum_{k=1}^{n} \frac{P(\omega_i)}{\boldsymbol{p}(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)$$

若 $i \neq j$时，$\boldsymbol{\theta}_i$和$\boldsymbol{\theta}_i$是独立的

$$\because \nabla_{\boldsymbol{\theta}_i} \ln p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)} \nabla_{\boldsymbol{\theta}_i} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)$$

$$\therefore \nabla_{\boldsymbol{\theta}_i} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i) = p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i) \nabla_{\boldsymbol{\theta}_i} \ln p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)$$

$$\therefore \nabla_{\boldsymbol{\theta}_i}\boldsymbol{L}(\boldsymbol{\theta}) = \sum_{k=1}^{n} \frac{p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i) P(\omega_i)}{\boldsymbol{p}(\boldsymbol{X}_k)} \nabla_{\boldsymbol{\theta}_i} \ln p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)$$

$$= \sum_{k=1}^{n} \boldsymbol{P}(\omega_i|\boldsymbol{X}_k, \boldsymbol{\theta}_i) \nabla_{\boldsymbol{\theta}_i} \ln p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)$$

后验概率公式

令 $\boldsymbol{\nabla}_{\boldsymbol{\theta}_i} \boldsymbol{L}(\boldsymbol{\theta}) = \sum_{k=1}^{n} \boldsymbol{P}(\omega_i|\boldsymbol{X}_k,\boldsymbol{\theta}_i)\boldsymbol{\nabla}_{\boldsymbol{\theta}_i} \ln p(\boldsymbol{X}_k|\omega_i,\boldsymbol{\theta}_i) = 0$，

解之，得参数集的最大似然估计： $\boldsymbol{\theta}_i$， i=1,…,c 。

将其代入下式，得$\boldsymbol{P(\omega_i)}$已知时混合概率密度$p(X)$的最大似然估计：

$$p(\boldsymbol{X}) = \sum_{i=1}^{c} p(\boldsymbol{X}|\omega_i,\boldsymbol{\theta}_i)\, P(\omega_i)$$

(2) $P(\omega_i)$未知

用条件极值法

约束条件为：
$$\begin{cases} P(\omega_i) \geq 0, i = 1,2,\cdots,c \\ \sum_{i=1}^{c} P(\omega_i) = 1 \end{cases}$$

构造目标函数：

$$J = \sum_{k=1}^{n} \ln\left(\sum_{i=1}^{c} p(\boldsymbol{X}_k | \omega_i, \boldsymbol{\theta}_i) P(\omega_i)\right) + \lambda\left(\sum_{i=1}^{c} P(\omega_i) - 1\right)$$

其中，λ为待定的Lagrange乘子。

对上式分别求关于$P(\omega_i)$、 $\boldsymbol{\theta}_i$ 的偏导,并令之为0。

求条件极值(一) $\nabla_{P(\omega_i)} J = 0$

$$\nabla_{P(\omega_i)} J = \frac{\partial \sum_{k=1}^{n} \ln\left(\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i, \boldsymbol{\theta}_i)\, P(\omega_i)\right)}{\partial P(\omega_i)} + \lambda$$

$$= \sum_{k=1}^{n} \frac{p(\boldsymbol{X}_k|\omega_i,\boldsymbol{\theta}_i)}{\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i,\boldsymbol{\theta}_i) P(\omega_i)} + \lambda = 0$$

$$\sum_{k=1}^{n} \frac{p(\boldsymbol{X}_k|\omega_i,\widehat{\boldsymbol{\theta}}_i)}{\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i,\widehat{\boldsymbol{\theta}}_i)\widehat{P}(\omega_i)} = -\lambda\ ,\ i = 1,2,\cdots,c$$

$$\sum_{k=1}^{n} \frac{p(\boldsymbol{X}_k|\omega_i,\widehat{\boldsymbol{\theta}}_i)\widehat{P}(\omega_i)}{\sum_{i=1}^{c} p(\boldsymbol{X}_k|\omega_i,\widehat{\boldsymbol{\theta}}_i)\widehat{P}(\omega_i)} = -\lambda\widehat{P}(\omega_i)\ ,\ i = 1,2,\cdots,c$$

$$\sum_{k=1}^{n} \widehat{P}(\omega_i|\boldsymbol{X}_k, \widehat{\boldsymbol{\theta}}_i) = -\lambda\widehat{P}(\omega_i)\ ,\ i = 1,2,\cdots,c$$

$$\sum_{k=1}^{n}\hat{P}(\omega_i|\boldsymbol{X}_k,\hat{\boldsymbol{\theta}}_i) = -\lambda\hat{P}(\omega_i)\ ,\ i = 1,2,\cdots,c \qquad (1)$$

上述各式相加

$$\sum_{i=1}^{c}\sum_{k=1}^{n}\hat{P}(\omega_i|\boldsymbol{X}_k,\hat{\boldsymbol{\theta}}_i) = -\lambda\sum_{i=1}^{c}\hat{P}(\omega_i)$$

$$\sum_{k=1}^{n}\sum_{i=1}^{c}\hat{P}(\omega_i|X_k,\hat{\theta}_i) = -\lambda$$

等于1

$\lambda = -n$,代入(1)中，得

$$\boxed{\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n}\hat{P}\left(\omega_i\middle|\boldsymbol{X}_k,\hat{\boldsymbol{\theta}}_i\right),\ \ i = 1,2,\cdots,c \qquad (2)}$$

同济大学 TONGJI UNIVERSITY

求条件极值(二) $\nabla_{\boldsymbol{\theta}_i} J=\mathbf{0}$, i=1,…,c

$$J=\sum_{k=1}^{n}\ln\left(\sum_{i=1}^{c}p(\boldsymbol{X}_k|\omega_i,\boldsymbol{\theta}_i)P(\boldsymbol{\omega}_i)\right)+\lambda\left(\sum_{i=1}^{c}P(\omega_i)-1\right)$$

$$\boxed{\sum_{k=1}^{n}\hat{\boldsymbol{P}}(\omega_i|\boldsymbol{X}_k,\hat{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}\ln p\left(\boldsymbol{X}_k\middle|\omega_i,\hat{\boldsymbol{\theta}}_i\right)=0,\quad i=1,\ldots,c\qquad(3)}$$

其中，

$$\hat{P}\left(\omega_i\middle|\boldsymbol{X}_k,\hat{\boldsymbol{\theta}}_i\right)=\frac{p\left(\boldsymbol{X}_k\middle|\omega_i,\hat{\boldsymbol{\theta}}_i\right)\hat{P}(\omega_i)}{\boldsymbol{p}(\boldsymbol{X}_k)}=\frac{p\left(\boldsymbol{X}_k\middle|\omega_i,\hat{\boldsymbol{\theta}}_i\right)\hat{P}(\omega_i)}{\sum_{j=1}^{c}p\left(\boldsymbol{X}_k\middle|\omega_j,\hat{\boldsymbol{\theta}}_j\right)\hat{P}(\omega_j)},\quad i=1,2,\cdots,c$$

原则上，可通过(2)、(3)式联立求解得到$\theta_i, P(\omega_i), i=1,\cdots,c$的最大似然估计。但得到闭式解困难，通常通过迭代算法，如EM算法，进行求解。

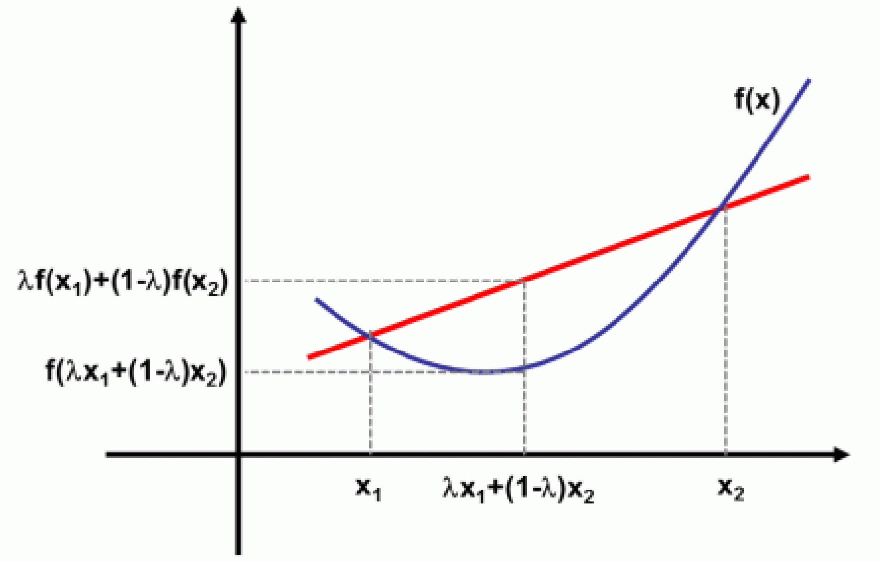
3.5 EM算法

- **期望最大 (Expectation Maximization,EM)算法：** 解决在概率模型中含有无法观测的隐含变量情况下的参数估计问题。
- **应用场合：**
  - 数据不完整，有缺失特征；
  - 存在隐变量，如样本的类别未知。
- **核心思想：**
  - 根据已有的、不完整数据，利用<u>对数似然函数期望</u>，<u>迭代地估计</u>分布函数的未知参数。

背景知识(1)凸函数(Convex Functions)

定义3.2 令 f 为区间 $I=[a,b]$ 上的实值函数。f 被称为凸函数，如果对 $\forall x_1,x_2 \in I,\lambda \in [0,1]$ 下式成立，

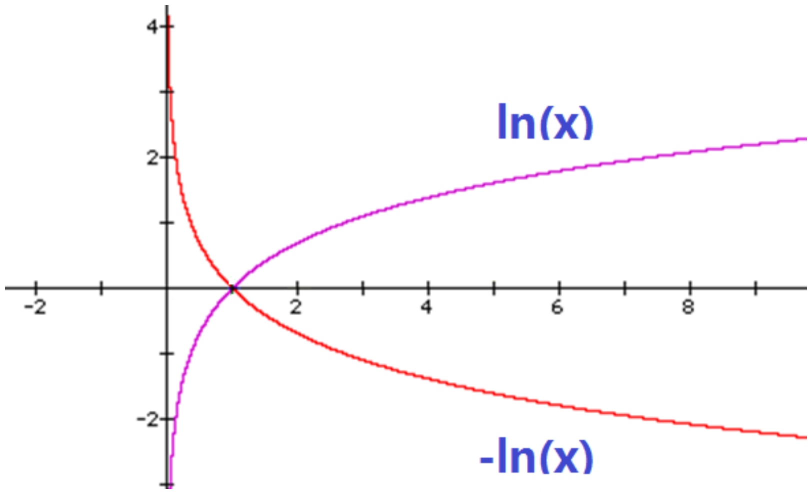
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$



弦在弧上

定理3.2 如果$f(x)$在[a,b]上是二阶可微的且 $f''(x) \geq 0$,则$f(x)$在[a,b]上是凸函数。

推论 $-\ln(x)$在$(0,\infty)$上是严格凸函数。



背景知识(2)詹森不等式(Jensen's inequality)

定理3.3 令 f 是一定义在区间 I 上的凸函数。如果 $x_1, x_2, \cdots, x_n \in I$ 且 $\lambda_1, \lambda_2, \cdots \lambda_n \geq 0, \sum_{i=1}^{n} \lambda_i = 1$,

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

因 $-\log(x)$ 是凸函数，利用詹森不等式，得下式:

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i)$$

完全数据 & 不完全数据

- Y：观测随机变量的数据，Z：隐随机变量的数据。
- Y和Z连在一起称为完全数据，观测数据Y又称为不完全数据。
- 设观测数据Y的概率分布是$P(\boldsymbol{Y}|\boldsymbol{\theta})$，其中$\boldsymbol{\theta}$是待估计参数，则不完全数据Y的对数似然函数$\boldsymbol{L}(\boldsymbol{\theta}) = \log P(\boldsymbol{Y}|\boldsymbol{\theta})$。
- 设Y和Z的联合概率分布是$P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})$，则完全数据的对数似然函数为$\log P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})$。
- **EM算法通过迭代求$\boldsymbol{L}(\boldsymbol{\theta}) = \log P(\boldsymbol{Y}|\boldsymbol{\theta})$的极大似然估计。**

EM算法原理

一个含隐变量Z的概率模型，目标是极大化观测数据Y关于参数θ的对数似然函数，

$$L(\boldsymbol{\theta}) = \log P(\boldsymbol{Y}|\boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} P(\boldsymbol{Y},\boldsymbol{Z}\,|\boldsymbol{\theta})$$

$$= \log \textstyle\sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{\theta}\,)P(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta})$$

EM算法通过迭代逐步近似极大化$\boldsymbol{L}(\boldsymbol{\theta})$。设第$\boldsymbol{i}$次迭代后$\boldsymbol{\theta}$的估计值是$\boldsymbol{\theta}^{(i)}$。希望新估计值$\boldsymbol{\theta}$能使$L(\boldsymbol{\theta})$增加，即$L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}^{(i)})$,并逐步达到极大值。

考虑两者差：

$$\boxed{\boldsymbol{L}(\boldsymbol{\theta}) - \boldsymbol{L}\left(\boldsymbol{\theta}^{(i)}\right) = \log\left(\sum_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{\theta}\,)P(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta})\right) - \log P(\boldsymbol{Y}|\boldsymbol{\theta}^{(i)})}$$

因$P(Z|Y,\theta^{(i)}) \geq 0$, 且 $\sum_Z P(Z|Y,\theta^{(i)}) = 1$。利用Jensen不等式，得两者差的下界：

分子分母同乘

$$L(\theta) - L(\theta^{(i)}) = \log\left(\sum_Z P(Z|Y,\theta^{(i)}) \frac{P(Z|\theta)P(Y|Z,\theta)}{P(Z|Y,\theta^{(i)})}\right) - \log P(Y|\theta^{(i)})$$

$$\geq \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Z|\theta)P(Y|Z,\theta)}{P(Z|Y,\theta^{(i)})} - \log P(Y|\theta^{(i)})$$

$$= \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Z|\theta)P(Y|Z,\theta)}{P(Z|Y,\theta^{(i)}) \cancel{\log} P(Y|\theta^{(i)})}$$

$$L(\theta) \geq L(\theta^{(i)}) + \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Z|\theta)P(Y|Z,\theta)}{P(Z|Y,\theta^{(i)}) \log P(Y|\theta^{(i)})}$$

- 不等式右侧函数记为$B(\theta,\theta^{(i)})$，是$L(\theta)$的一个下界。
- 任何使$B(\theta,\theta^{(i)})$增大的θ，也使$L(\theta)$增大。

为使$L(\boldsymbol{\theta})$尽可能增大，选择$\boldsymbol{\theta}^{(i+1)}$使得$B(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)})$达到极大，即

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\theta}\left(L(\boldsymbol{\theta}^{(i)}) + \sum_{Z} P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(i)})\log\frac{P(\boldsymbol{Z}|\boldsymbol{\theta})P(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta})}{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(i)})\log P(\boldsymbol{Y}|\boldsymbol{\theta}^{(i)})}\right)$$

$$= \arg\max_{\theta}\left(\sum_{Z} P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(i)})\log\frac{P(\boldsymbol{Z}|\boldsymbol{\theta})P(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta})}{\boxed{P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(i)})}}\right)$$

省略对θ极大化而言是常数的项

$$= \arg\max_{\theta}\left(\boxed{\sum_{Z} P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta}^{(i)})\log P(\boldsymbol{Z}|\boldsymbol{\theta})P(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta})}\right)$$

$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)})$函数

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\theta}\left(Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)})\right)$$

上式是EM算法的一次迭代，即求Q函数及其极大化。

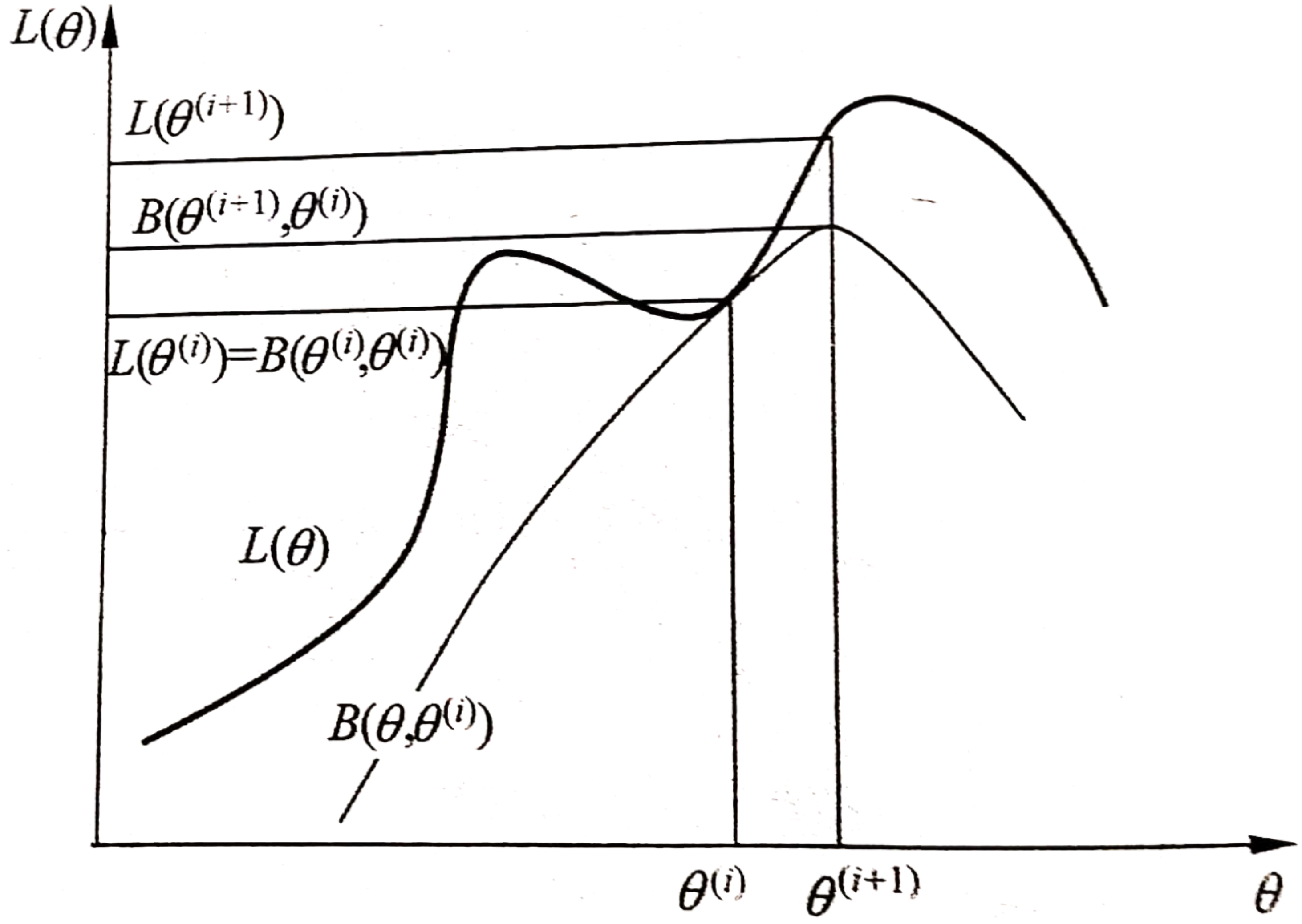


图3.4 EM算法的直观解释

EM算法是通过不断地使下界极大化，去逼近求解“对数似然函数极大化”。

定义3.3　（Q函数）

完全数据的对数似然函数$\log P(Y,Z|\theta)$，关于隐变量Z在给定观测数据$\boldsymbol{Y}$和当前参数$\boldsymbol{\theta}^{(i)}$条件下的期望，称为Q函数，即

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right)=E_Z\left[\log P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})\middle|\boldsymbol{Y},\boldsymbol{\theta}^{(i)}\right]$$

$$=\sum_Z P\left(\boldsymbol{Z}\middle|\boldsymbol{Y},\boldsymbol{\theta}^{(i)}\right)\log P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})$$

这里，$P\left(\boldsymbol{Z}\middle|\boldsymbol{Y},\boldsymbol{\theta}^{(i)}\right)$是在给定观测数据$\boldsymbol{Y}$和当前参数$\boldsymbol{\theta}^{(i)}$下隐变量$\boldsymbol{Z}$的条件概率分布。

EM算法：

输入：观测数据Y, 隐变量Z, 联合分布$P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})$，条件分布$P(\boldsymbol{Z}|\boldsymbol{Y},\boldsymbol{\theta})$;

输出：模型参数$\boldsymbol{\theta}$。

(1) 选择参数的初始值$\boldsymbol{\theta}^{(0)}$，开始迭代；

(2) E步：记$\boldsymbol{\theta}^{(i)}$为第$\boldsymbol{i}$次迭代参数$\boldsymbol{\theta}$的估计值，在第$\boldsymbol{i}$+1次迭代的E步，计算

待极大化

$$Q\left(\boxed{\boldsymbol{\theta}},\boxed{\boldsymbol{\theta}^{(i)}}\right)=\sum_{Z}P\left(\boldsymbol{Z}\middle|\boldsymbol{Y},\boldsymbol{\theta}^{(i)}\right)\log P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})$$

(3) M步：求使$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right)$极大化的$\boldsymbol{\theta}$，确定第$\boldsymbol{i}$+1次迭代的参数估计值$\boldsymbol{\theta}^{(i+1)}$

$$\boldsymbol{\theta}^{(i+1)}=\arg\max_{\theta}\left(Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}\right)\right)$$

(4) 重复第(2)步和第(3)步，直到收敛。

关于EM算法的两点说明

- **参数的初值**：可任选，但EM算法对初值敏感。
- **迭代终止条件**：一般是对较小的正数ε_1 或 ε_2，若满足

$$\left\|\boldsymbol{\theta}^{(i+1)}-\boldsymbol{\theta}^{(i)}\right\|<\varepsilon_1$$

或

$$\left\|Q\left(\boldsymbol{\theta}^{(i+1)},\boldsymbol{\theta}^{(i)}\right)-Q(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}^{(i-1)})\right\|<\varepsilon_2$$

则停止迭代。

EM算法的收敛性

定理3.4 设$P(Y|\theta)$为观测数据的似然函数，$\theta^{(i)}$(i=1,2,...)为EM算法得到的参数估计序列，$P(Y|\theta^{(i)})$为对应的似然函数序列，则$P(Y|\theta^{(i)})$是单调递增的，即

$$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)})$$

定理3.5 设$L(\theta)=logP(Y|\theta)$为好数据的对数似然函数，$\theta^{(i)}$(i=1,2,⋯)为EM算法得到的参数估计序列，L$(\theta^{(i)})$为对应的对数似然函数序列

(1)如果$P(Y|\theta)$有上界，则$L\left(\theta^{(i)}\right)=logP(Y|\theta^{(i)})$收敛到某一值$L^*$;

(2)在函数$Q\left(\theta|\theta^{'}\right)$与$L(\theta)$满足一定条件下，由EM算法得到的参数估计序列$\theta^{(i)}$的收敛值$\theta^*$是$L(\theta)$的稳定点。

同济大学 TONGJI UNIVERSITY

EM算法在无监督参数估计中的应用

问题描述：给定混合样本集$\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n\}$，其类别数已知(c)，各样本的类别标签未知。每个类别的类条件概率密度 $p(\boldsymbol{Y}|\omega_j, \theta_j)$ 函数形式已知，$P(\omega_j)$未知。

求解目标：估计各类的分布参数θ_j 和$P(\omega_j)$，j=1,⋯,c。

令$\theta = \{\theta_1, \cdots, \theta_c\}$，$P = \left(P(\omega_1), \cdots, P(\omega_c)\right)$，$\Theta = (\theta, P)$。

无监督参数估计，实质是混合模型的参数估计。

1. 明确隐变量，写出完全数据的对数似然函数

各样本的未知类别即是隐变量。

令(Y,z)是完全数据, 其中Y是观测样本, z是类标签。若$Y \in \omega_j$, 则$z=j$。其概密为

$$p(\boldsymbol{Y},z|\Theta)=p(\boldsymbol{Y}|z,\Theta)P(z|\Theta)=p(\boldsymbol{Y}|\omega_j,\theta_j)P(\omega_j)$$

完全数据的对数似然函数：

$$L(\theta)=\ln p(\boldsymbol{Y},z|\Theta)=\sum_{k=1}^{n}\ln\left[p\left(\boldsymbol{Y}_k|\omega_{kj},\theta_{kj}\right)P\left(\omega_{kj}\right)\right]$$

其中下标kj表示$\boldsymbol{Y}_k$的实际类标签，$\omega_{kj}\in\{\omega_1,\cdots,\omega_c\}$。

2. EM算法的E步：确定Q函数

令EM算法中第t次迭代已得到θ和P的估计,记为$\theta^{(t)}$和$P^{(t)}$,合记为$\Theta^{(t)}$。

构造对数似然函数关于类标签的条件期望:

$$Q(\Theta,\Theta^{(t)}) = E\{\sum_{k=1}^{n}\ln[p(\boldsymbol{Y}_k|\omega_{kj},\theta_{kj})P(\omega_{kj})]\}$$

$$= \sum_{k=1}^{n}E\{\ln[p(\boldsymbol{Y}_k|\omega_{kj},\theta_{kj})P(\omega_{kj})]\}$$

$$= \sum_{k=1}^{n}\sum_{kj=1}^{c}P(\omega_{kj}|\boldsymbol{Y}_k,\Theta^{(t)})\ln[p(\boldsymbol{Y}_k|\omega_{kj},\theta_{kj})P(\omega_{kj})]$$

从上式可看出,对每个观测样本$\boldsymbol{Y}_k$, 式中ω_{kj}取遍各类, 故上式可重写为:

$$Q(\Theta,\Theta^{(t)}) = \sum_{k=1}^{n}\sum_{j=1}^{c}P(\omega_j|\boldsymbol{Y}_k,\Theta^{(t)})\ln[p(\boldsymbol{Y}_k|\omega_j,\theta_j)P(\omega_j)]$$

3. EM算法的M步

求使Q函数取最大的θ和P，即 $\Theta^{(t+1)} = \underset{\Theta}{\arg\max}\left\{Q(\Theta, \Theta^{(t)})\right\}$

①估计θ_j。设各类的概密彼此独立，则由$\frac{\partial Q(\Theta,\Theta^{(t)})}{\partial \theta_j} = 0$ 可知 θ_j应满足

$$\sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\boldsymbol{Y}_k, \Theta^{(t)}) \frac{\partial \ln p\left(\boldsymbol{Y}_k\middle|\omega_j, \theta_j\right)}{\partial \theta_j} = 0$$

②估计$P(\omega_j)$。由于$P(\omega_j) \geq 0$，$\sum_{j=1}^{c} P\left(\omega_j\right) = 1$，采用条件极值法，目标函数

$$J = Q(\Theta, \Theta^{(t)}) + \lambda\left(\sum_{j=1}^{c} P\left(\omega_j\right) - 1\right)$$

求$\nabla_{P(\omega_i)} J = 0$，可得 $\boldsymbol{P\left(\omega_j\right) = \frac{1}{n}\sum_{i=1}^{n} P(\omega_j|Y_k, \Theta^{(t)})}$，j=1,⋯,c

其中，$P(\omega_j|X_k, \Theta^{(t)}) = \dfrac{p\left(\boldsymbol{Y}_i\middle|\omega_j, \theta^{(t)}\right)P^{(t)}(\omega_j)}{\sum_{j=1}^{c} p\left(\boldsymbol{Y}_i\middle|\omega_j, \theta^{(t)}\right)P^{(t)}(\omega_j)}$

无监督参数估计的EM算法:

输入： 混合样本集 $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n\}$，类别数c，混合概型，阈值$\varepsilon$

输出： 各类的分布参数θ_j 和$P(\omega_j)$， j=1,…,c

过程:

(1)选择初始估计值： $\theta = \theta^{(0)}$， $P = P^{(0)}$; 令t=0。

(2)重复执行下列步骤:

①计算 $P(\omega_j|\boldsymbol{Y}_k, \Theta^{(t)}) = \dfrac{p\left(\boldsymbol{Y}_k\middle|\omega_j, \theta^{(t)}\right)P^{(t)}(\omega_j)}{\sum_{j=1}^{c} p\left(\boldsymbol{Y}_k\middle|\omega_j, \theta^{(t)}\right)P^{(t)}(\omega_j)}$, k=1,2,…,n； j=1,2,…,c

②计算 $P^{(t+1)}(\omega_j) = \frac{1}{n}\sum_{k=1}^{n} P(\omega_j|\boldsymbol{Y}_k, \Theta^{(t)})$， j=1,2,…,c

③求解 $\theta_j^{(t+1)}$。 $\theta_j^{(t+1)}$是下面关于θ_j的方程的解,

$$\sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\boldsymbol{Y}_k, \Theta^{(t)}) \frac{\partial \ln P\left(\boldsymbol{Y}_k\middle|\omega_j, \theta_j\right)}{\partial \theta_j} = 0 ,\quad j=1,2,\dots,c$$

④检查是否收敛。若$\left\|\Theta^{(t+1)} - \Theta^{(t)}\right\| < \varepsilon$, 则停止, 否则t=t+1, 转(2) 。

同济大学 TONGJI UNIVERSITY

高斯混合模型

定义3.4 （高斯混合模型）

高斯混合模型是指具有如下形式的概率分布模型：

$$P(y)=\sum_{j=1}^{c}\pi_j\mathcal{N}(y|\mu_j,\Sigma_j)$$

其中， π_j是系数，$\pi_j\geq 0$，$\sum_{j=1}^{c}\pi_j=1$；

$\mathcal{N}(y|\mu_j,\Sigma_j)$ 是高斯密度，被称为第j个分模型。

同济大学
TONGJI UNIVERSITY

高斯混合模型的EM算法

设观测数据$\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_n$由高斯混合模型生成，

$$P(\boldsymbol{Y}|\theta)=\sum_{j=1}^{2}\pi_j f(Y|\theta_j)$$

- $\mathbf{Z}=(\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_n)$是隐变量，它确定观测数据来自混合分布中的哪部分。

目标是估计混合参数(来自两高斯分布的概率)和两高斯分布的均值和方差，因此，$\boldsymbol{\theta}=(\boldsymbol{P},\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2),\boldsymbol{P}=(\boldsymbol{P}_1,\boldsymbol{P}_2)^{\boldsymbol{T}}$

$$\boxed{\begin{array}{ll}\boldsymbol{Y}_i|(\mathbf{z}_i=\omega_1)\sim\mathcal{N}_d(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) & P(\mathbf{z}_i=\omega_1)=P_1\\ \boldsymbol{Y}_i|(\mathbf{z}_i=\omega_2)\sim\mathcal{N}_d(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) & P(\mathbf{z}_i=\omega_2)=P_2=1-P_1\end{array}}$$

高斯混合概型参数估计的EM算法:

输入： 混合样本集 $\mathcal{D}=\{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_{\boldsymbol{n}}\}$，类别数c，阈值$\varepsilon$

输出： 各类的分布参数μ_j、Σ_j和$P(\omega_j)$， j=1,2

过程：

(1)选择初始估计值： $\theta=\theta^{(0)}=\left(P^{(0)},\mu_1^{(0)},\mu_2^{(0)},\Sigma_1^{(0)},\Sigma_2^{(0)}\right), P=(P_1^{(0)},P_2^{(0)})^T$; 令t=0。

(2)重复执行下列步骤：

① 计算 $T_{j,i}^{(t)}:=p\left(\mathbf{z}_i=\omega_j\middle|\boldsymbol{Y}_i,\theta^{(t)}\right)=\dfrac{P_j^{(t)}f(\boldsymbol{Y}_i|\boldsymbol{\mu}_j^{(t)},\boldsymbol{\Sigma}_j^{(t)})}{P_1^{(t)}f\left(\boldsymbol{Y}_i|\boldsymbol{\mu}_1^{(t)},\boldsymbol{\Sigma}_1^{(t)}\right)+P_2^{t}f\left(\boldsymbol{Y}_i|\boldsymbol{\mu}_2^{(t)},\boldsymbol{\Sigma}_2^{(t)}\right)}$ ， $i=1,2,\cdots,n$

② 计算 $P_j^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}T_{j,i}^{(t)}$ ， $j=1,2$

③ 计算 $\mu_j^{(t+1)}=\dfrac{\sum_{i=1}^{n}T_{j,i}^{(t)}\boldsymbol{Y}_i}{\sum_{i=1}^{n}T_{j,i}^{(t)}}$， $\Sigma_j^{(t+1)}=\dfrac{\sum_{i=1}^{n}T_{j,i}^{(t)}(\boldsymbol{Y}_i-\mu_1^{(t+1)})\left(\boldsymbol{Y}_i-\mu_1^{(t+1)}\right)^T}{\sum_{i=1}^{n}T_{j,i}^{(t)}}$， $j=1,2$

④ 检查是否收敛.若$\left\|\Theta^{(t+1)}-\Theta^{(t)}\right\|<\varepsilon$,则停止,否则t=t+1, 转(2) 。

作业2

1）编码实现高斯混合概型参数估计的EM算法。

2）利用Sklearn中的make_blobs方法生成高斯混合样本集，用于参数估计。

3）调用1)中自编的参数估计函数，对2)中生成的混合样本集进行参数估计。

2019年10月27日交作业