模式识别的理论与方法

Pattern Recognition

裴继红

Chapter 3 (Part 1):
Maximum Likelihood and Bayesian Parameter Estimation

(Sections 3.1-3.2)

本讲内容:

- 最大似然估计

引言：贝叶斯理论的难点

- 贝叶斯分类理论中的数据效用
  - 设计一个贝叶斯分类器需要知道下面的信息:
    - $P(\omega_i)$ (每一个类的先验概率)
    - $P(x \mid \omega_i)$ (似然函数：类条件概率密度)

不幸的是，我们很少能够完整地知道这些信息!

需要从训练样本出发设计分类器

1. 对先验概率的估计一般不太困难
2. 类条件密度的估计比较困难，

主要问题是用于进行估计的样本数量通常太少 (特别是在特征空间的维数相对很大时!)

引言：参数估计

若已知待研究问题的一些先验信息，则可以降低估计的难度

- 若已知分布 $P(x \mid \omega_i)$的形式为正态情况，即

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

则只需要均值矢量和协方差矩阵这 2 个参数即可刻画其分布特征

- 若类条件概率密度函数的形式已知，表征概率密度函数的某些参数是未知的，则需要解决的问题转化为：使用样本估计这些未知的参数，因此称为**参数估计**

参数估计常使用的技术

- 最大似然估计（Maximum-Likelihood，ML）
- 贝叶斯估计（ Bayesian Estimation，BE ）

上述两种估计方法得到的结果几乎一样，但解决问题的思路不同

引言：MLE和BE

- 在**MLE**最大似然估计中，假设参数是未知的，但待估计的参数是确定的量
  - 最佳的估计是：训练样本集合在该参数决定的概率密度函数下，可以获得最大的概率值。称该参数为最优参数

- 在**BE**贝叶斯估计方法中，参数被看成是某个已知分布的随机变量
  - 利用样本的信息修正对参数的初始估计值，随着样本数量的增加，使**参数的后验概率密度函数**变成更加尖锐的函数。

在上述两种方法中, 均使用 $P(\omega_i \mid x)$ 作为分类规则!

最大似然估计
Maximum-Likelihood Estimation

- 最大似然估计的优点

1. 随着样本数量的增加，收敛性变好
2. 比任何其他的迭代技术都简单，适合实用

- 一般原理

假设有 c 个类，且

$$P(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)，即$$

$$P(x \mid \omega_j) \equiv P(x \mid \theta_j)$$

在正态分布时的参数有:

$$\theta_j = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, ..., \sigma_j^{11}, \sigma_j^{22}, \mathrm{cov}(x_j^m, x_j^n)...)$$

最大似然估计的一般原理

假设数据集合 D 划分成的类为 $\omega_1, \omega_2, \cdots, \omega_c$

我们可以将数据集合 D 划分成互不相交的样本子集合 $D_1, D_2, \ldots, D_c$，每一个子集合中的样本属于同一类

对每一个数据集合 D_i，单独估计自己的 $P(x|\omega_i)$

即，只需估计其参数即可得到分布函数，

若为正态分布, 则参数为 $\theta_i = \{\mu_i, \Sigma_i\}$

最大似然估计：独立抽样

- 使用信息
  - 由提供的训练样本进行估计

    $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, 每个 θ_i (i = 1, 2, …, c) 与一个特定类相关联

- 假设集合 D 中包含同一类的 n 个样本, $x_1, x_2, \ldots, x_n$, 且这些样本是独立抽样得到的，则

$$P(D \mid \theta) = \prod_{k=1}^{n} P(x_k \mid \theta) = F(\theta)$$

$P(D \mid \theta)$称为样本集合 D的似然函数

- 参数 θ 的ML估计是:
  - 通过使$P(D \mid \theta)$ 最大化的 θ 值，使得实际观测到的样本集合具有最大的似然概率

对数似然函数

- 在实际中，似然函数进行对数运算后，计算比较简单。
- 此时,称为对数似然函数，如下:

$$l(\theta) \equiv \ln p(D|\theta)$$

这样 $$l(\theta) \equiv \ln\left(\prod_{k=1}^{n} p(x_k|\theta)\right) = \sum_{k=1}^{n} \ln p(x_k|\theta)$$

由于自然对数函数是单调增函数，因此对数似然函数和原似然函数的极值点的位置相同

最优解的必要条件

似然函数的定义导数为:

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln P(x_k \mid \theta)$$

则最优解的必要条件是:

$$\nabla_\theta l = 0$$

若 θ 由 p 个参数组成，则上式代表 p 个方程组成的方程组

最大似然估计：求取似然函数的极值点

- 令 $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$，并令 ∇_θ 是梯度算子

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}\right]^t$$

$$l(\theta) \equiv \sum_{k=1}^{n} \ln p(x_k|\theta)$$

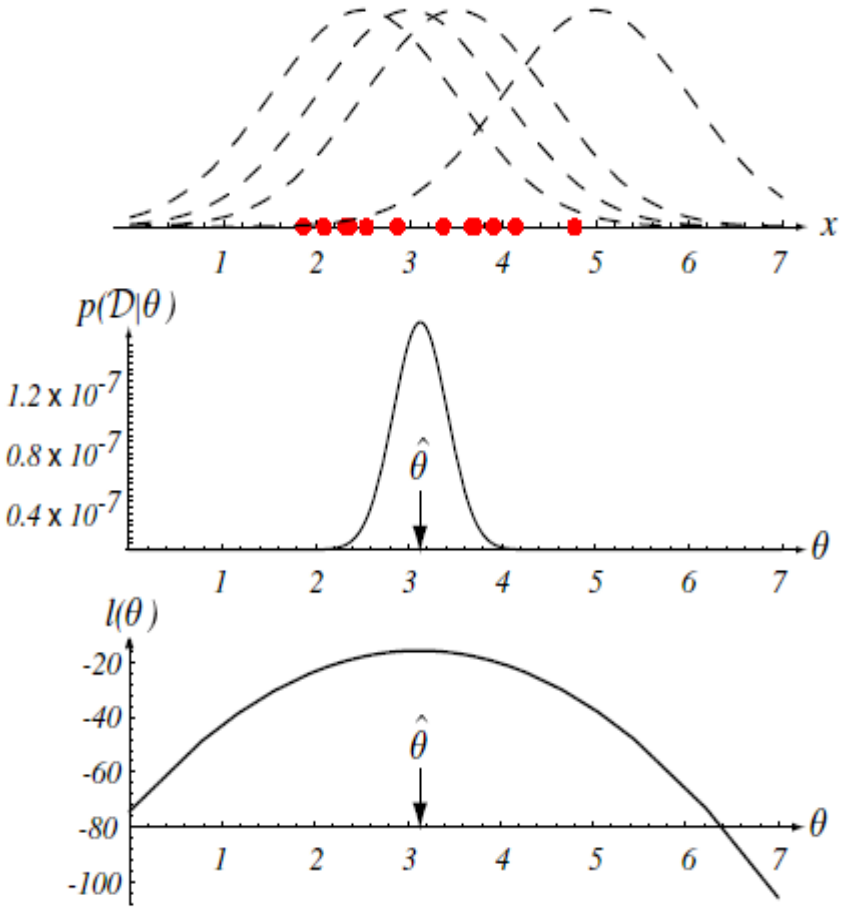
- 问题重新表述：求使对数似然函数取得最大值的参数 θ：

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

Figure 3.1

- 右上图示出了几个一维的训练样本点，假定它们是从一个方差已知，但均值未知的高斯分布抽样得到的。虚线表示了从源分布中的其中四种可能的分布。

- 右中间的图示出了以均值为变量的似然函数 $p(D|\theta)$ 。若有大量的训练样本点，该似然函数将非常窄。
- 使似然最大的值标记为 $\hat{\theta}$; 该参数也使得右下图中的对数似然 $l(\theta)$ 最大化。



注意：似然函数 $p(D|\theta)$和条件概率密度函数 $p(x|\theta)$很相似，但 $p(D|\theta)$是θ的函数，而 $p(x|\theta)$是以θ为参数的 x 的函数。

似然函数$p(D|\theta)$不是概率密度函数，其曲线下的面积没有实际意义

举例：正态分布，μ 未知

– $P(x_i \mid \mu) \sim N(\mu, \Sigma)$

训练样本由一个多元正态分布抽样得到：

$$\ln P(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\Sigma\right|\right] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu)$$

$$\nabla_\mu \ln P(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu)$$

此处，$\theta = \mu$ ，

即: 对μ的最大似然估计满足: $\sum_{k=1}^{n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$

举例：一元正态分布，μ 未知

等式两边乘以协方差矩阵 Σ，并重新整理公式，得到:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

μ 为所有训练样本的算术平均！

结论:

- 若已知 $P(x_k \mid \omega_j)$ (j = 1, 2, …, c) 的形式是*d*-维特征空间的高斯分布函数，则可以通过估计矢量$\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$，设计最优分类器。

举例：一元正态分布，μ 和 σ 未知

- 高斯分布: μ 和 σ 未知

  $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\partial}{\partial\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\partial}{\partial\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0 \qquad \begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

举例：一元正态分布， μ 和 σ未知

极值条件:

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

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

由 (1) 和 (2)式, 可得:

$$\theta_1=\mu=\sum_{k=1}^{n}\frac{x_k}{n} \quad ; \quad \theta_2=\sigma^2=\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2$$

偏差： Bias

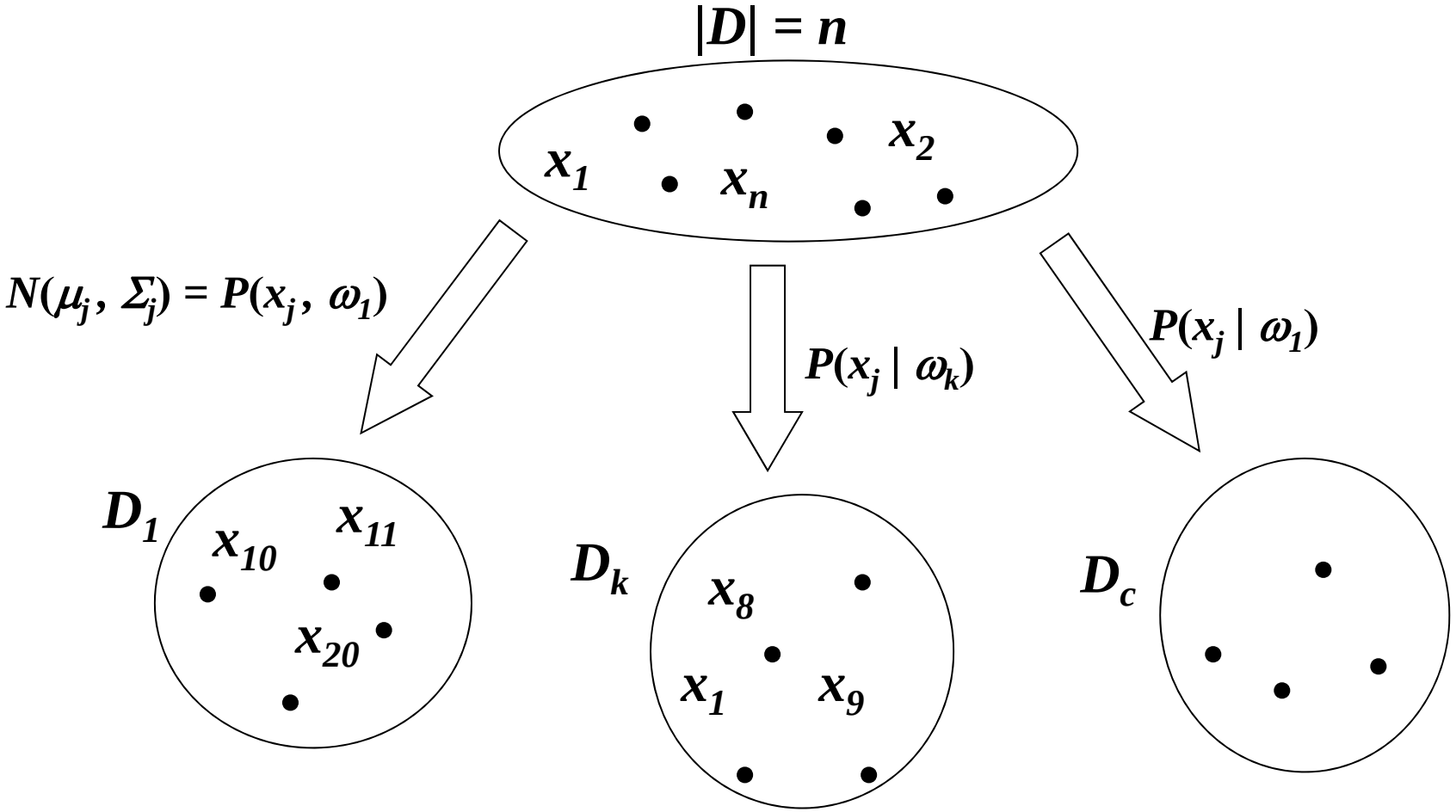
ML 估计 对方差 σ^2 的估计是有偏差的

$$E\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]=\frac{n-1}{n}\cdot\sigma^2\neq\sigma^2$$

对协方差矩阵 Σ 的无偏估计公式如下:

$$\underbrace{C=\frac{1}{n\text{-}1}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^t}_{\text{样本协方差矩阵}}$$

小结：最大似然估计图示



小结: 最大似然估计的最优解

$\theta = (\theta_1, \theta_2, \ldots, \theta_c)$,

问题: 寻找 θ 使得:

$$\max_{\theta}\{P(D\,|\,\theta)\} = \max_{\theta}\{P(x_1,\ldots,x_n\,|\,\theta)\}$$

$$\overset{\text{独立同分布}}{=} \max_{\theta}\left\{\prod_{k=1}^{n} P(x_k\,|\,\theta)\right\}$$

若 $p\left(D\,\middle|\,\hat{\theta}\right)$ 具有可微分性, 则可以通过梯度方法计算: θ