模式识别的理论与方法

Pattern Recognition

裴继红

Chapter 3 (Part 2):
Maximum Likelihood and Bayesian Parameter Estimation

(Sections 3.3-3.5)

本讲内容: 贝叶斯参数估计

- 贝叶斯估计 (BE)
- 贝叶斯估计: 高斯分布的情况
- 贝叶斯估计: 一般情况下的估计问题

回顾：贝叶斯决策

后验概率 $P(\omega_i \mid x)$ 的计算是贝叶斯分类的核心

- 目标：计算 $P(\omega_i \mid x, D)$

  给定训练样本集合 D，将贝叶斯公式重新写为

$$P(\omega_i \mid x, D) = \frac{P(x \mid \omega_i, D) P(\omega_i \mid D)}{\sum_{j=1}^{c} P(x \mid \omega_j, D) P(\omega_j \mid D)}$$

类条件密度

① 由贝叶斯公式，计算 $P(\omega_j \mid x, D)$

转化为计算 $P(x \mid \omega_j, D)$ 和 $P(\omega_j \mid D)$

② 对于有监督学习，事先知道每个样本的类属信息，因此

先验概率容易得到

假设样本类数 c，每一类的样本相互独立，则

$$P(\omega_i \mid D) \rightarrow P(\omega_i \mid D_1, D_2, \ldots D_i, \ldots D_c) \rightarrow P(\omega_i \mid D_i) \rightarrow P(\omega_i)$$

此时，后验概率 $P(\omega_i \mid x, D) \rightarrow P(\omega_i \mid x, D_i)$

在不引起混淆的情况下，为了书写方便常用 D 替代 D_i

类条件密度

贝叶斯公式的证明:

$$P(x,\omega_i \mid D_i) = P(x \mid D_i,\omega_i)P(\omega_i \mid D_i)$$

$$P(x \mid D) = \sum_j P(x,\omega_j \mid D_j)$$

$P(\omega_i) = P(\omega_i \mid D)$ (训练样本提供了这些信息!)

这样:

$$P(\omega_i \mid x, D) = \frac{P(x \mid \omega_i, D_i)P(\omega_i \mid D)}{P(x \mid D)} = \frac{P(x \mid \omega_i, D_i)P(\omega_i)}{\sum_{j=1}^{c} P(x \mid \omega_j, D_j)P(\omega_j)}$$

或记为: $$P(\omega_i \mid x, \mathrm{D}) = \frac{P(x \mid \omega_i, \mathrm{D})P(\omega_i)}{\sum_{j=1}^{c} P(x \mid \omega_j, \mathrm{D})P(\omega_j)}$$

贝叶斯参数学习

- 在**MLE**中，假设参数 θ 未知，但它是确定的量
- 在**BE**中，参数 θ 是一个随机变量

1. 条件概率 ***P*(*x*| *D*)** **[= *P*(*x* |ω_i, D_i)]** 的计算是贝叶斯学习的目的
2. 核心问题：给定样本集合***D***，计算 ***P*(θ| *D*)**

$$p(x|D)=\int p(x|\theta)\,p(\theta|D)\,d\theta$$

离散情况

$$p(x|D)=\sum_j p(x|\theta_j)\,p(\theta_j|D)$$

举例1：高斯分布，均值 μ 未知

目标: 使用后验概率密度 $P(\theta \mid D)$ 估计参数 θ

首先,考虑一元变量情况: $P(\mu \mid D)$

即这里, 只有均值 μ 是未知的参数

$$P(x \mid \mu) \sim N(\mu, \sigma^2) \qquad\qquad P(\mu) \sim N(\mu_0, \sigma_0^2)$$

(σ 、μ_0 和 σ_0 是已知的!)

举例1：高斯分布，均值 μ 未知

$$P(\mu \mid \mathrm{D}) = \frac{P(\mathrm{D} \mid \mu) P(\mu)}{\int P(\mathrm{D} \mid \mu) P(\mu) d\mu} \qquad (1)$$

$$= \alpha \prod_{k=1}^{n} P(x_k \mid \mu) P(\mu)$$

复制密度

$$P(\mu \mid \mathrm{D}) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

由式(1) 和(2)，可以推导出:

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \qquad \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

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

举例1：高斯分布，均值 μ 未知

- 单变量高斯分布 $P(x \mid D)$

计算出 $P(\mu \mid D)$

然后, 测试样本的条件概率密度 $P(x \mid D)$ 可以被计算出来！

$$P(x \mid \mathrm{D}) = \int P(x \mid \mu) P(\mu \mid \mathrm{D}) d\mu$$

在高斯分布情况下:

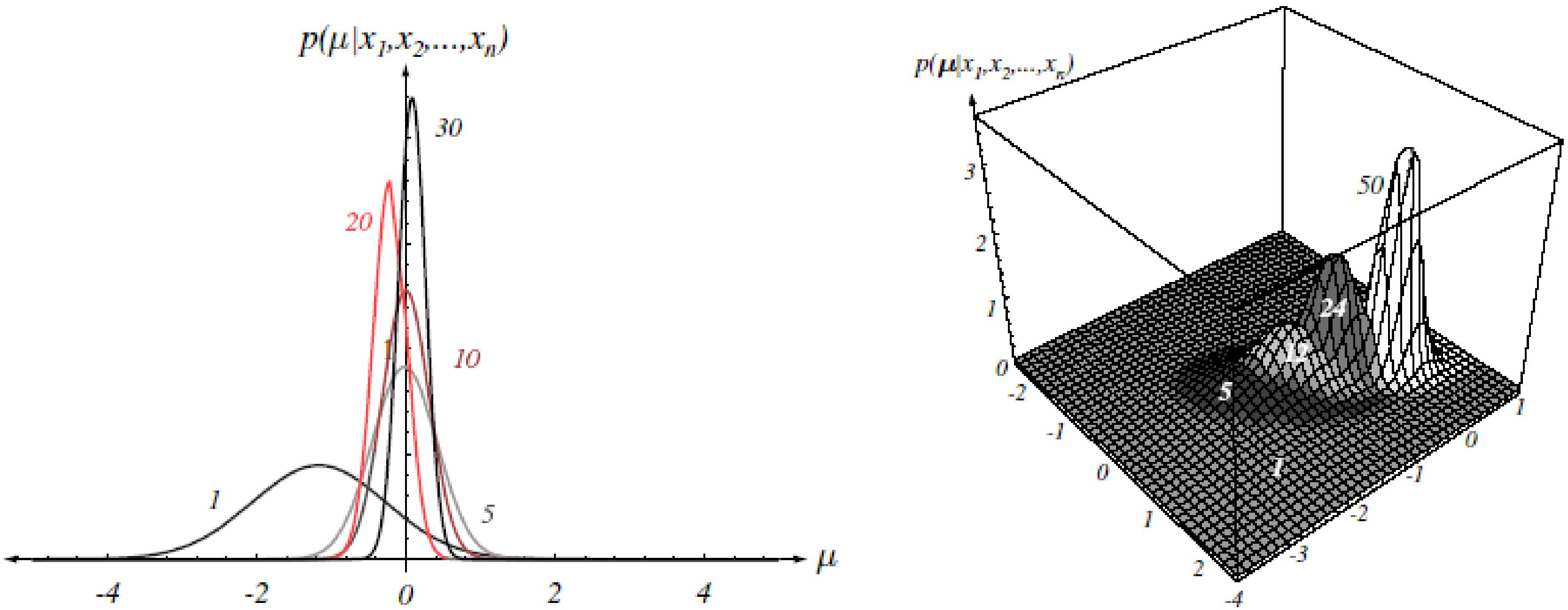
$$P(x \mid \mathrm{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

得到类条件概率密度 $P(x \mid D_j, \omega_j)$

- 因此: $P(x \mid D_j, \omega_j)$ 与 $P(\omega_j)$一起，利用贝叶斯公式，可得到贝叶斯分类规则如下：

$$\max_{\omega_j}\left[P(\omega_j \mid x, \mathrm{D}\right] \equiv \max_{\omega_j}\left[P(x \mid \omega_j, \mathrm{D}_j) P(\omega_j)\right]$$

Figure3.2



- 在1维和2维空间中正态分布时，均值参数的贝叶斯学习。
  图中后验分布的估计图上标注的数字是在估计中使用的样本数量

贝叶斯参数估计：一般情况

- $P(x \mid D)$ 的计算可以应用于任何的分布密度未知、但可以被参数化的情况。
- 基本假设如下:

1. 假设 $P(x \mid \theta)$ 的形式是已知的, 但参数 θ 的准确值未知；
2. 对 θ 的先验知识假定已经包含在一个已知的先验密度P(θ)之中。
3. 关于 θ 的其余的知识包含在：具有 n 个随机样本 $x_1, x_2, \ldots, x_n$ 的数据集合 D 中，该样本集合是由 $P(x)$ 经过独立抽样得到。

贝叶斯参数估计：核心问题

因此，贝叶斯参数估计的最本质的核心问题为:

① **计算出后验密度 P(θ | D)**

② **然后导出 条件概率P(x | D)**

$$p\left(x\middle|D\right)=\int p\left(x\middle|\theta\right)p\left(\theta\middle|D\right)d\theta$$

应用贝叶斯公式：

$$P(\theta\,|\,\mathrm{D})=\frac{P(\mathrm{D}\,|\,\theta)P(\theta)}{\int P(\mathrm{D}\,|\,\theta)P(\theta)d\theta}$$

由独立性假设有:

$$P(\mathrm{D}\,|\,\theta)=\prod_{k=1}^{n}P(x_k\,|\,\theta)$$

具体算法1：增量学习参数估计方法

- 假设集合 D^n 中包含同一类的 n 个样本,且这些样本是独立同分布抽样得到的, 记为 $\mathrm{D}^n = \{x_1, x_2, \ldots, x_n\}$， D^{n-1}为其前n-1个样本组成的集合。则

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

$$P(\theta \mid \mathrm{D}^n) = \frac{P(D^n \mid \theta) P(\theta)}{\int P(D^n \mid \theta) P(\theta) d\theta} = \frac{P(\mathrm{x}_n \mid \theta) P(\theta \mid D^{n-1})}{\int P(\mathrm{x}_n \mid \theta) P(\theta \mid D^{n-1}) d\theta}$$

$$P(\theta \mid \mathrm{D}^0) = P(\theta)$$

其他算法

- EM
- Gibbs
- ……