第三章
概率密度函数的估计

2010-10-13

最小风险估计

- **损失函数**：把 θ 估计为 $\hat{\theta}$ 所造成的损失：$\lambda(\hat{\theta},\theta)$
- **期望风险**：

$$\begin{aligned} R &= \int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\mathbf{x},\theta)d\theta d\mathbf{x} \\ &= \int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})p(\mathbf{x})d\theta d\mathbf{x} \\ &= \int_{E^d}p(\mathbf{x})\int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta d\mathbf{x} \\ &= \int_{E^d}R(\hat{\theta}\,|\,\mathbf{x})p(\mathbf{x})d\mathbf{x}; \end{aligned}$$

- **条件风险**：

$$R(\hat{\theta}\,|\,\mathbf{x}) = \int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta.$$

最小风险估计

- 最小化条件风险 ⇒ 最小化期望风险 。
- 在有限样本集下，最小化经验风险

$$R(\hat{\theta}\,|\,\boldsymbol{K}) = \int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\boldsymbol{K})d\theta.$$

- **贝叶斯估计量**：（在样本集 $\boldsymbol{K}$ 下）是条件风险（经验风险）最小的估计量 $\hat{\theta}_{\mathrm{BE}}$，即

$$\hat{\theta}_{\mathrm{BE}} = \arg\min_{\hat{\theta}} R(\hat{\theta}\,|\,\boldsymbol{K}).$$

最小风险估计

- 定义**平方误差损失函数** $\lambda(\hat{\theta},\theta) = (\theta-\hat{\theta})^2.$

$$\begin{aligned} R(\hat{\theta}\,|\,\mathbf{x}) &= \int_{\Theta}\lambda(\hat{\theta},\theta)p(\theta\,|\,\mathbf{x})d\theta \\ &= \int_{\Theta}[\theta-E(\theta\,|\,\mathbf{x})]^2p(\theta\,|\,\mathbf{x})d\theta + \int_{\Theta}[E(\theta\,|\,\mathbf{x})-\hat{\theta}]^2p(\theta\,|\,\mathbf{x})d\theta; \end{aligned}$$

- **定理**：如采用平方损失函数，则有

$$\hat{\theta}_{\mathrm{BE}} = E[\theta\,|\,\mathbf{x}] = \int_{\Theta}\theta p(\theta\,|\,\mathbf{x})d\theta;$$

- 同理，在给定样本集 $\boldsymbol{K}$ 下，θ 的贝叶斯估计为

$$\hat{\theta}_{\mathrm{BE}} = E[\theta\,|\,\boldsymbol{K}] = \int_{\Theta}\theta\, p(\theta\,|\,\boldsymbol{K})\,d\theta;$$

最小风险估计的求解步骤

1. 确定 θ 的先验分布 $p(\theta)$;
2. 由样本集 $\boldsymbol{K}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ 求出其联合分布：

$$p(\boldsymbol{K}\,|\,\theta) = \prod_{k=1}^{N}p(\mathbf{x}_k\,|\,\theta);$$

3. 计算 θ 的后验分布

$$p(\theta\,|\,\boldsymbol{K}) = \frac{p(\boldsymbol{K}\,|\,\theta)p(\theta)}{\int_{\Theta}p(\boldsymbol{K}\,|\,\theta)p(\theta)d\theta};$$

4. 计算贝叶斯估计

$$\hat{\theta}_{\mathrm{BE}} = \int_{\Theta}\theta p(\theta\,|\,\boldsymbol{K})d\theta.$$

一元正态分布的贝叶斯估计

- 总体分布密度为：

$$p(x\,|\,\mu) \sim N(\mu,\sigma^2);$$

- 均值 μ 未知，μ 的先验分布为：

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

- 样本集： $\boldsymbol{K}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$
- 用贝叶斯估计方法求 μ 的估计量

一元正态分布的贝叶斯估计

7

- 计算 μ 的后验分布

$$p(\mu \mid \boldsymbol{K}) = \frac{p(\boldsymbol{K} \mid \mu)p(\mu)}{p(\boldsymbol{K})}$$

$$= \alpha \prod_{k=1}^{N} p(x_k \mid \mu)p(\mu) \sim N(\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \quad \sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2};$$

其中 $m_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$ 为样本均值;

一元正态分布的贝叶斯估计

8

- 计算 μ 的贝叶斯估计

$$\hat{\mu} = \int \mu p(\mu \mid \boldsymbol{K}) d\mu = \mu_N;$$

- 当 $N=0$ 时， $\hat{\mu}_{BE} = \mu_0$;
- 当 $N \to \infty$ 时， $\hat{\mu}_{BE} \to m_N$;
- 如 $\sigma_0^2 = 0$ ，则 $\hat{\mu}_{BE} \equiv \mu_0$; 即先验知识可靠，样本不起作用。
- 如 $\sigma_n >> \sigma$，则 $\hat{\mu}_{BE} = m_N$; 即先验知识十分不确定，完全依靠样本信息。

贝叶斯学习

基本思想

10

- 利用 $\boldsymbol{\theta}$ 的先验分布 $p(\boldsymbol{\theta})$ 及训练样本提供的信息 $p(\boldsymbol{K}|\boldsymbol{\theta})$，求 $\boldsymbol{\theta}$ 的后验分布 $p(\boldsymbol{\theta}|\boldsymbol{K})$; 然后直接求解总体分布。

$$p(\mathbf{x} \mid \boldsymbol{K}) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid \boldsymbol{K}) d\boldsymbol{\theta}$$

$$= \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{K}) d\boldsymbol{\theta};$$

- 将类条件概率密度（总体分布）$p(\mathbf{x}|\boldsymbol{K})$ 和未知参数的后验概率密度 $p(\boldsymbol{\theta}|\boldsymbol{K})$ 联系起来;
- 贝叶斯学习的结果与最大似然估计的结果近似:

$$p(\mathbf{x} \mid \boldsymbol{K}) \approx p(\mathbf{x} \mid \hat{\boldsymbol{\theta}}_{\mathrm{ML}}).$$

递推（序贯）后验概率

11

- 考虑 $N>1$ 个学习样本，记样本集 $\boldsymbol{K}^N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$;

$$p(\boldsymbol{K}^N \mid \boldsymbol{\theta}) = p(\mathbf{x}_N \mid \boldsymbol{\theta}) p(\boldsymbol{K}^{N-1} \mid \boldsymbol{\theta}),$$

$$p(\boldsymbol{\theta} \mid \boldsymbol{K}^N) = \frac{p(\boldsymbol{K}^N \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{K}^N \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$= \frac{p(\mathbf{x}_N \mid \boldsymbol{\theta}) p(\boldsymbol{K}^{N-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \cdot \frac{1}{p(\boldsymbol{K}^{N-1})}}{\int_{\Theta} p(\mathbf{x}_N \mid \boldsymbol{\theta}) p(\boldsymbol{K}^{N-1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \cdot \frac{1}{p(\boldsymbol{K}^{N-1})}}$$

$$= \frac{p(\mathbf{x}_N \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{K}^{N-1})}{\int_{\Theta} p(\mathbf{x}_N \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{K}^{N-1}) d\boldsymbol{\theta}}.$$

递推贝叶斯学习

12

- 设 $p(\boldsymbol{\theta} \mid \boldsymbol{K}^0) = p(\boldsymbol{\theta})$ ，当样本数目增多，可得到后验概率密度函数序列:

$$p(\boldsymbol{\theta}), p(\boldsymbol{\theta} \mid \mathbf{x}_1), p(\boldsymbol{\theta} \mid \mathbf{x}_1, \mathbf{x}_2), \cdots$$

- 如果此序列收敛予以真实数值为中心的δ函数，则称样本分布具有贝叶斯学习(Bayesian Learning)性质:

$$p(\boldsymbol{\theta} \mid \boldsymbol{K}^{N \to \infty}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0);$$

$$p(\mathbf{x} \mid \boldsymbol{K}^{N \to \infty}) = p(\mathbf{x} \mid \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0) = p(\mathbf{x}).$$

一元正态分布的贝叶斯学习

13

- 计算 μ 的后验分布

$$p(\mu \mid \boldsymbol{K}^N) = \frac{p(\boldsymbol{K}^N \mid \mu)\, p(\mu)}{p(\boldsymbol{K}^N)}$$

$$= \alpha \prod_{k=1}^{N} p(x_k \mid \mu)\, p(\mu) \sim N(\mu_N, \sigma_N^2);$$

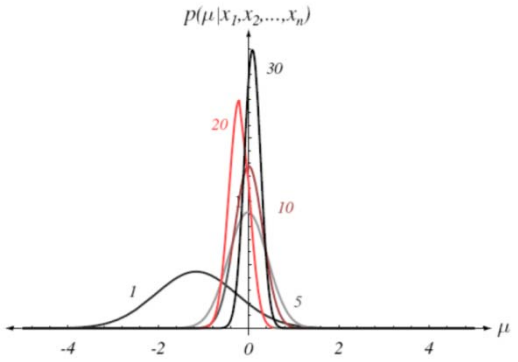
$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \quad \sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2};$$

其中 $m_N = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_k$ 为样本均值;

当N $\to \infty$时，$\sigma_N^2 \to 0$，$p(\mu|\boldsymbol{K}) \to \delta$函数。

一元正态分布的贝叶斯学习

14



一元正态分布的贝叶斯学习

15

- 直接计算总体密度：

$$p(x \mid \boldsymbol{K}^N) = \int p(x \mid \mu)\, p(\mu \mid \boldsymbol{K}^N)\, d\mu$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2 + \sigma_N^2}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu_N}{\sqrt{\sigma^2 + \sigma_N^2}}\right)^2\right\}$$

$$\sim N(\mu_N, \sigma^2 + \sigma_N^2).$$

均值 μ_N，

方差由 σ^2 增为 $\sigma^2 + \sigma_N^2$ ----- 由于用了 μ 的估计值而不确定性增加

非监督参数估计

简介最大似然法

假设条件

17

1. 样本集 $\boldsymbol{K}=\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ 中的样本分属于c个类别，但未知各样本所属类别；
2. 已知各类先验概率 $P(\omega_i)$，i=1, …, c；（有时也可未知，一起估计）
3. 已知类条件概率密度形式 $p(\boldsymbol{x}|\omega_i,\theta_i)$, i=1, …, c;
4. 需估计未知的 c 个参数向量 $\theta_1, \theta_2, \ldots, \theta_c$。

似然函数

18

- **混合密度函数**：分量密度的线性组合

$$p(\boldsymbol{x} \mid \theta) = \sum_{i=1}^{c} p(\boldsymbol{x} \mid \omega_i, \theta_i) P(\omega_i)$$

分量密度　混合参数

- 似然函数和对数似然函数：

$$l(\theta) = p(\boldsymbol{K} \mid \theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \theta)$$

$$H(\theta) = \ln[l(\theta)] = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i \mid \theta)$$

最大似然估计

19

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \prod_{k=1}^{N} p(x_k \mid \theta) = \arg\max_{\theta\in\Theta} \sum_{k=1}^{N} \ln p(x_k \mid \theta);$$

- **可识别性**
  - 设 $\theta \neq \theta'$，如对混合分布中每个 $\boldsymbol{x}$ 都有 $p(\boldsymbol{x}|\theta) \neq p(\boldsymbol{x}|\theta')$，则称密度 $p(\boldsymbol{x}|\theta)$ 是可识别的；
  - 大部分常见连续随机变量的分布密度函数都是可识别的；离散随机变量的混合概率函数往往是不可识别的。

最大似然估计

20

- **计算问题：**求解微分方程组

$$\nabla_{\theta_i} H(\theta) = \sum_{k=1}^{N} \frac{1}{p(x_k \mid \theta)} \nabla_{\theta_i} \left[\sum_{j=1}^{c} p(x_k \mid \omega_j, \theta_j) P(\omega_j) \right]$$

$$= \sum_{k=1}^{N} \frac{1}{p(x_k \mid \theta)} \nabla_{\theta_i} [p(x_k \mid \omega_i, \theta_i) P(\omega_i)] \quad （设 \theta_i, \theta_j 独立）$$

$$= \sum_{k=1}^{N} P(\omega_i \mid x_k, \theta_i) \nabla_{\theta_i} \ln p(x_k \mid \omega_i, \theta_i)$$

其中后验概率 $P(\omega_i \mid x_k, \theta_i) = \dfrac{p(x_k \mid \omega_i, \theta_i) P(\omega_i)}{p(x_k \mid \theta)}$

正态分布的非监督最大似然估计

21

- 均值向量 $\boldsymbol{\mu}_i$ 未知，$\boldsymbol{\Sigma}_i$，$P(\omega_i)$，c 已知
  - 最大似然估计满足方程组

$$\sum_{k=1}^{N} \hat{P}(\omega_i \mid x_k, \hat{\theta}_i) \nabla_{\theta_i} \ln p(x_k \mid \omega_i, \hat{\theta}_i) = 0, \quad i = 1, \cdots c$$

代入正态分布

$$\sum_{k=1}^{N} P(\omega_i \mid x_k, \hat{\mu}_i \quad \hat{\mu}_i(j+1) = \frac{\sum_{k=1}^{N} P(\omega_i \mid x_k, \hat{\mu}_i(j)) x_k}{\sum_{k=1}^{N} P(\omega_i \mid x_k, \hat{\mu}_i(j))} \quad \frac{\omega_i \mid x_k, \hat{\mu}_i) x_k}{'(\omega_i \mid x_k, \hat{\mu}_i)}$$

参数估计总结

22

- 最大似然估计

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p(\boldsymbol{K} \mid \theta) = \arg\max_{\theta} \prod_{k=1}^{N} p(\mathbf{x}_k \mid \theta);$$

- 最大后验概率估计

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\boldsymbol{K} \mid \theta) p(\theta) = \arg\max_{\theta} \sum_{k=1}^{N} \ln p(\mathbf{x}_k \mid \theta) + \ln p(\theta);$$

- 贝叶斯估计

$$p(\theta \mid \boldsymbol{K}) = \frac{p(\boldsymbol{K} \mid \theta) p(\theta)}{\int p(\boldsymbol{K} \mid \theta) p(\theta) d\theta}, \hat{\theta}_{\mathrm{BE}} = E(\theta \mid \boldsymbol{K}) = \int_{\Theta} \theta p(\theta \mid \boldsymbol{K}) d\theta;$$

- 贝叶斯学习

$$p(\mathbf{x} \mid \boldsymbol{K}) = \int p(\mathbf{x} \mid \theta) p(\theta \mid \boldsymbol{K}) d\theta.$$