基本概念

贝叶斯公式:

$$P(\omega_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_j)P(\omega_j)}{P(\boldsymbol{x})}$$

- **样本** $\mathbf{x} \in \mathbb{R}^d$
- **类别/状态** ω_i
- **先验概率** $P(\omega_i)$：历史数据中类别ω_i的初始概率
- **样本分布密度** $p(\mathbf{x})$：所有样本在特征空间中的分布情况
- **类条件概率密度** $P(\mathbf{x} \mid \omega_i)$： 在类别 ω_i 下，样本 $\mathbf{x}$ 出现的概率
- **后验概率** $P(\omega_i \mid \mathbf{x})$：观察到样本**x**后修正的类别概率
- **错误概率** $P(e \mid \mathbf{x}) = \begin{cases} P(\omega_2 \mid \mathbf{x}) & 若判为\omega_1 \\ P(\omega_1 \mid \mathbf{x}) & 若判为\omega_2 \end{cases}$
- **平均错误率** $P(e) = \int P(e \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$
- **正确率** $P(c) = 1 - P(e)$

最小错误率贝叶斯决策

本质上是根据贝叶斯公式计算后验概率，基于 **最大后验概率** 进行判决，不考虑决策风险

- $x \in w_k \text{ iff } k = \arg\max_i \{P(w_i|X)\}$
- $P(w_i|X) = \frac{P(X|w_i)P(w_i)}{\sum_{j=1}^{c} P(X|w_j)P(w_j)}$

等价表达形式

- 后验概率形式

  $if\ P(\omega_i|\mathbf{x}) = \max_{j=1,\ldots,c} P(\omega_j|\mathbf{x}), then\ \mathbf{x} \in \omega_i$
- 似然与先验乘积形式

  $if\ P(\mathbf{x}|\omega_i)P(\omega_i) = \max_{j=1,\ldots,c} P(\mathbf{x}|\omega_j)P(\omega_j), then\ \mathbf{x} \in \omega_i$
- 似然比形式（二分类情况）

  $if\ l(x) = \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}, then\ \mathbf{x} \in \omega_1$

  $if\ l(x) = \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} < \frac{P(\omega_2)}{P(\omega_1)}, then\ \mathbf{x} \in \omega_2$
- 对数似然比形式（二分类情况）

  $if\ h(x) = \ln[l(x)] = \ln P(\mathbf{x}|\omega_1) - \ln P(\mathbf{x}|\omega_2) > \ln\left(\frac{P(\omega_2)}{P(\omega_1)}\right), then\ \mathbf{x} \in \omega_1$

  $if\ h(x) = \ln[l(x)] = \ln P(\mathbf{x}|\omega_1) - \ln P(\mathbf{x}|\omega_2) < \ln\left(\frac{P(\omega_2)}{P(\omega_1)}\right), then\ \mathbf{x} \in \omega_2$

其中：$l(x)$ 为似然比，$h(x)$ 为对数似然比，$\frac{P(\omega_2)}{P(\omega_1)}$ 为似然比阈值

最小风险贝叶斯决策

- 目标：最小化决策风险
- 损失函数：$\lambda(\alpha_i, \omega_j)$，表示将 ω_j 误判为 α_i 的代价
- 利用后验概率与损失函数，计算条件风险：
  $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|x), i = 1, 2, \ldots, a$
- 决策：$R(\alpha_k|x) = \min_{i=1,2,\ldots,a} R(\alpha_i|x)$
- 最小错误率贝叶斯决策就是在 0-1损失函数（即正确分类损失为0，错误分类损失为1）条件下的最小风险贝叶斯决策

朴素贝叶斯决策

相比于上面提到的贝叶斯决策，带分类的样本 $\mathbf{x}$ 从一维变成**多维**（可以认为有 $\{x_1, x_2, \ldots x_i, \ldots x_d\}$ 多重属性），现在我们要使用贝叶斯公式对这个多维样本进行分类

- **属性条件独立性假设**：在已知类别 ω_i 的条件下，假设所有属性 x_i 都是相互独立的。即如下公式：
  $P(\mathbf{x} \mid \omega) = P(x_1 x_2, \ldots, x_i, \ldots, x_d \mid \omega) = \prod_{i=1}^{d} P(x_i \mid \omega)$
- 好处：降低样本集大小需求；降低复杂度
- 贝叶斯公式 + 属性独立性条件
  $P(w|\mathbf{x}) = \frac{P(w)P(\mathbf{x}|w)}{P(\mathbf{x})} = \frac{P(w)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i|w)$
- 朴素贝叶斯决策
  $w_k = \arg\max_j P(w_j) \prod_{i=1}^{d} P(x_i|w_j)$

贝叶斯估计

把待估计参数 θ 看作是具有先验分布 $p(\theta)$ 的随机变量，其取值与样本集 $\mathbf{x}$ 有关，根据样本集 $\mathbf{x}$ 估计（利用样本将先验概率修正为后验概率 ）

概念：

- **损失函数** $\lambda(\hat{\theta}, \theta)$：将 θ 估计为 $\hat{\theta}$ 造成的损失
- **平方误差损失** $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$
- **期望风险** $R = \int_{E^d} R(\hat{\theta} \mid x) p(x) dx$
- **条件风险** $R(\hat{\theta} \mid x) = \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta \mid x) d\theta$
- **贝叶斯估计量** $\hat{\theta} = \arg\min \left(\int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta \mid x) d\theta \right)$
- **关键关系**：最小化期望风险 ⇨ 最小化所有x的条件风险
- **特例**：平方误差损失下，贝叶斯估计量=后验均值 $\hat{\theta} = E[\theta \mid x]$

贝叶斯估计定理（平方误差损失下）：

- 单样本形式：$\hat{\theta} = E(\theta \mid x) = \int_{\Theta} \theta p(\theta \mid x) d\theta$
- 样本集形式：$\hat{\theta} = E(\theta \mid X) = \int_{\Theta} \theta p(\theta \mid X) d\theta$

算法步骤（平方误差损失下）

1. 确定先验分布：$p(\theta)$
2. 计算样本联合分布：$p(X \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$
3. 求后验分布：$p(\theta \mid X) = \frac{p(X|\theta)p(\theta)}{\int_\Theta p(X|\theta)p(\theta)d\theta}$
4. 计算贝叶斯估计量：$\hat{\theta} = \int_\Theta \theta p(\theta \mid X)d\theta$

例题

【例题】细胞有正常(w_1)，异常(w_2)两类，其先验概率为$P(w_1) = 0.9, P(w_2) = 0.1$。有一个待识别的细胞，观测值为$x$，现在已知如果细胞是正常的，出现$x$的概率为0.2；如果细胞是异常的，出现$x$的概率是0.4。决策的损失表如下：

决策	w_1	w_2
w_1	0	6
w_2	1	0

1. 基于最小错误率原则对待识别细胞进行归类
2. 基于最小风险原则对待识别细胞进行归类

【解答】

1. 应用贝叶斯公式：

   已知x，细胞属于w_1的概率为：

$$\begin{aligned} P(w_1 \mid x) &= \frac{P(x \mid w_1)P(w_1)}{P(x \mid w_1)P(w_1) + P(x \mid w_2)P(w_2)} \\ &= \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.4 \times 0.1} \\ &= 0.818 \end{aligned}$$

   已知x，细胞属于w_2的概率为：

$$\begin{aligned} P(w_2 \mid x) &= \frac{P(x \mid w_2)P(w_2)}{P(x \mid w_1)P(w_1) + P(x \mid w_2)P(w_2)} \\ &= \frac{0.4 \times 0.1}{0.2 \times 0.9 + 0.4 \times 0.1} \\ &= 0.182 \end{aligned}$$

   所以，应该决策为w_1。

2. 决策为w_1的风险为：

$$R(w_1) = 0 \times P(w_1 \mid x) + 6 \times P(w_2 \mid x) = 1.092$$

   决策为w_2的风险为：

$$R(w_2) = 1 \times P(w_1 \mid x) + 0 \times P(w_2 \mid x) = 0.818$$

   所以，应该决策为w_2。

参考文献

[北航机器学习期末考试例题汇总 - 凉宫秋月的文艺部]