第7章 贝叶斯分类器

7.1 贝叶斯决策论

$Y=\{c_1, c_2 \cdots c_N\}$ λ_{ij}：真实c_j误分为c_i的损失

$P(c_i|x)$ 后验概率

样本分为c_i所产生的期望损失：

$$R(c_i|x)=\sum_{j=1}^{N}\lambda_{ij}P(c_j|x)$$

判定准则 $h: \mathcal{X}\to Y$ 最小化总体风险

$$R(h)=E_x[R(h(x)|x)]$$

贝叶斯判定准则：

最小化总体损失，只需在每个样本上选择那个能使条件风险最小的类别标记。

$$h^*(x)=\arg\min_{c\in Y} R(c|x)$$

若$\lambda_{ij}\Rightarrow\begin{cases}0 & \text{if } i=j\\ 1 & \text{otherwise}\end{cases}$

此时条件风险：

$$R(c|x)=1-P(c|x)$$

于是最小化分类错误率的贝叶斯分类：

$$h^*(x)=\arg\max_{c\in Y} P(c|x)$$

对每个样本，选择能使后验概率$P(c|x)$最大的类别标记

判别式模型：给定x，可通过直接建模$P(c|x)$来预测c（决策树，BP，SVM）

生成式模型：先对联合分布$P(x,c)$建模，然后再获得$P(c|x)$

生成式：$P(c|x)=\dfrac{P(x,c)}{P(x)}$

其中$P(c|x)=\dfrac{P(c)P(x|c)}{P(x)}$ → 似然：x相对于c的概率

$P(c)$：先验

$P(x)$：用于归一化，证据因子，给定样本x与类别无关

7.2 极大似然估计

假定$P(x|c)$具有确定的形式并且被参数向量θ_c唯一确定

用训练集D估计参数θ_c

概率模型训练 $\Longleftrightarrow$ 参数估计

令D_c表示训练集D第c类样本组成的集合，设样本独立同分布，则参数θ_c

$$P(D_c|\theta_c)=\prod_{x\in D_c}P(x|\theta_c)$$

对θ_c进行极大似然估计

$$LL(\theta_c)=\log P(D_c|\theta_c)$$
$$=\sum_{x\in D_c}\log P(x|\theta_c)$$

$$\hat{\theta}_c=\arg\max LL(\theta_c)$$

7.3 朴素贝叶斯分类器

“属性条件独立性假设”：假设每个属性独立地对分类结果产生影响（进行完整组合爆炸样本）

$$P(c|x)=\frac{P(c)P(x|c)}{P(x)}=\frac{P(c)}{P(x)}\prod_{i=1}^{d}P(x_i|c)$$

$$h_{nb}(x)=\arg\max_{c\in Y} P(c)\prod_{i=1}^{d}P(x_i|c)$$

朴素贝叶斯常需要"平滑"

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N}$$

$$\hat{P}(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

7.4 半朴素贝叶斯分类器

适当考虑一部分属性间相互依赖信息

"独依赖估计": ODE

$$P(c|x) \propto P(c) \prod_{i=1}^{d} P(x_i|c, pa_i)$$

pa_i 是 x_i 所依赖属性，父属性如何确定的？

① 假设所有属性都依赖于同一属性

SPODE 方法

通过交叉验证方法

② TAN

7.5 贝叶斯网