半朴素贝叶斯分类器
AODE
（Averaged One-Dependent Estimator）

判别式模型与生成式模型

- 判别式（ep. LR）：

$$P(c \mid \boldsymbol{x})$$

- 生成式：

$$P(c \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})} \Rightarrow P(c \mid \boldsymbol{x}) = \frac{P(c)\, P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})}$$

- 贝叶斯分类器属于生成式，假定样本由某种过程生成，不同的预设产生了不同的模型

回顾朴素贝叶斯

- 预设：样本各个属性互相独立

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

- P(c)：被分类为c的先验概率
- $P(x_i \mid c)$：被分类为c时，属性x_i的条件概率

- 训练方式：计数

朴素贝叶斯的训练

显然, 朴素贝叶斯分类器的训练过程就是基于训练集 D 来估计类先验概率 $P(c)$, 并为每个属性估计条件概率 $P(x_i \mid c)$.

令 D_c 表示训练集 D 中第 c 类样本组成的集合, 若有充足的独立同分布样本, 则可容易地估计出类先验概率

$$P(c) = \frac{|D_c|}{|D|} \ . \tag{7.16}$$

对离散属性而言, 令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合, 则条件概率 $P(x_i \mid c)$ 可估计为

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{|D_c|} \ . \tag{7.17}$$

对连续属性可考虑概率密度函数, 假定 $p(x_i \mid c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$, 其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的均值和方差, 则有

$$p(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) \ . \tag{7.18}$$

示例

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.46	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.36	0.37	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

- P(青绿|是) = 3/8
- P(青绿|否) = 3/9

示例

- P(密度=0.697|是)=$\frac{1}{\sqrt{2\pi}\cdot 0.129}\exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right)$

- 0.129 & 0.574：分类为是的样本，密度的均值与方差

$$P(\text{好瓜}=\text{是})\times P_{\text{青绿}|\text{是}}\times P_{\text{蜷缩}|\text{是}}\times P_{\text{浊响}|\text{是}}\times P_{\text{清晰}|\text{是}}\times P_{\text{凹陷}|\text{是}}$$
$$\times P_{\text{硬滑}|\text{是}}\times p_{\text{密度: 0.697}|\text{是}}\times p_{\text{含糖: 0.460}|\text{是}}\approx 0.038\ ,$$
$$P(\text{好瓜}=\text{否})\times P_{\text{青绿}|\text{否}}\times P_{\text{蜷缩}|\text{否}}\times P_{\text{浊响}|\text{否}}\times P_{\text{清晰}|\text{否}}\times P_{\text{凹陷}|\text{否}}$$
$$\times P_{\text{硬滑}|\text{否}}\times p_{\text{密度: 0.697}|\text{否}}\times p_{\text{含糖: 0.460}|\text{否}}\approx 6.80\times 10^{-5}\ .$$

朴素贝叶斯的问题

- 假设所有属性互相独立（不符合大多数场景）
- 使用朴素贝叶斯，往往代表「不需要考虑属性间独立性」

- 例：垃圾邮件分类问题，一半的邮件包含「博彩」，一半的邮件包含「发票」。朴素贝叶斯会认为一封邮件同时包含这两个词的概率是0.25（实际上是0）

从朴素贝叶斯到半朴素贝叶斯

- 区别：各维度之间是否有相互依赖信息

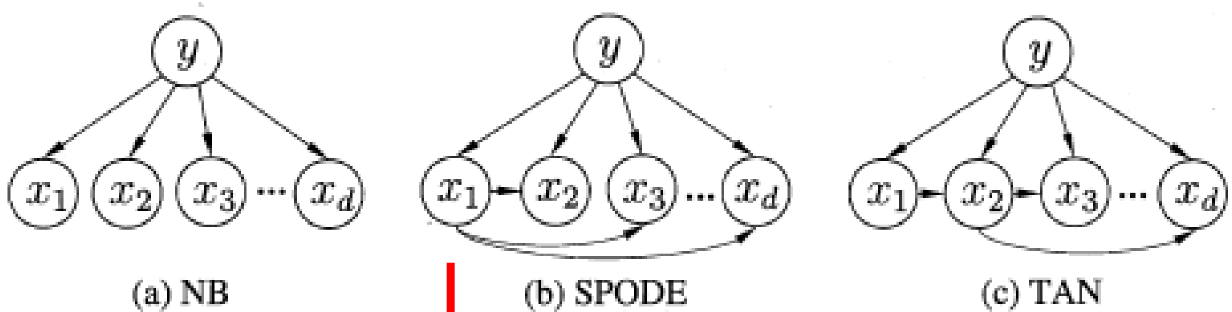


图 7.1　朴素贝叶斯与两种半朴素贝叶斯分类器所考虑的属性依赖关系

超父属性（super-parent）

独依赖估计（ODE）

- One-dependent Estimator
- 假设每个属性除了分类类别（c）以外，仅依赖一个其他属性

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

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

SPODE

- super-parent ODE

观看视频耗时	习题耗时	课程偏好	性别
短	长	理工	男
短	长	理工	男
长	短	文史	男
长	短	文史	女
短	长	理工	女
短	短	文史	女
长	短	文史	女

- 理工课程通常视频耗时短，习题耗时长，文史课程相反

SPODE

- 「课程偏好」这一属性作为超父（SP）
- 朴素贝叶斯：
  - P(习题耗时=长 | 男) = 2/3
- SPODE：
  - P(习题耗时=长 | 理工, 男) = 1

AODE

- Averaged One-Dependent Estimator
- Webb,G. 2005 “Not so naïve Bayes: Aggregating one-dependence Estimator”

- 尝试将每个属性作为超父构建SPODE，将有足够训练数据支撑的SPODE集成起来作为最终结果

$$P(c \mid \boldsymbol{x}) \propto \sum_{\substack{i=1 \\ |D_{x_i}| \geqslant m'}}^{d} P(c, x_i) \prod_{j=1}^{d} P(x_j \mid c, x_i) , \tag{7.23}$$

其中 D_{x_i} 是在第 i 个属性上取值为 x_i 的样本的集合, m' 为阈值常数. 显然, AODE 需估计 $P(c, x_i)$ 和 $P(x_j \mid c, x_i)$. 类似式(7.20), 有

$$\hat{P}(c, x_i) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} , \tag{7.24}$$

$$\hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j} , \tag{7.25}$$

其中 N_i 是第 i 个属性可能的取值数, D_{c,x_i} 是类别为 c 且在第 i 个属性上取值为 x_i 的样本集合, D_{c,x_i,x_j} 是类别为 c 且在第 i 和第 j 个属性上取值分别为 x_i 和 x_j 的样本集合. 例如, 对西瓜数据集 3.0 有

$$\hat{P}_{是,浊响} = \hat{P}(好瓜 = 是, 敲声 = 浊响) = \frac{6+1}{17+3} = 0.350 ,$$

$$\hat{P}_{凹陷|是,浊响} = \hat{P}(脐部 = 凹陷 \mid 好瓜 = 是, 敲声 = 浊响) = \frac{3+1}{6+3} = 0.444 .$$

拉普拉斯修正

- 用于平滑——某个属性值在样本中未出现，不进行平滑会将这一属性概率设为0

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

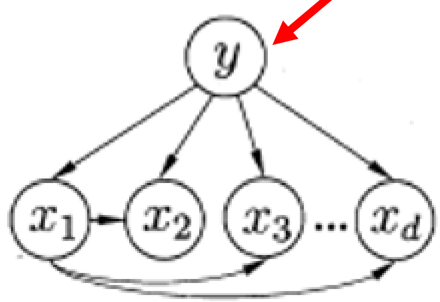
$$\hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j}$$

- 如果一个属性值没出现过，它对应的概率将被平滑处理成1/Ni（Ni为这个属性所有可能的取值）

AODE

- 每一项就是一个SPODE

$$P(c \mid \boldsymbol{x}) \propto \sum_{\substack{i=1 \\ |D_{x_i}| \geqslant m'}}^{d} \boxed{P(c, x_i) \prod_{j=1}^{d} P(x_j \mid c, x_i)}$$



(b) SPODE

编程任务

- 编写一个AODE（阈值m设为0）
- 支持离散属性、支持连续属性、支持增量学习
- 使用watermelon数据集训练与测试，输出每个样本在是和否两个属性上的估计概率值
- 提交代码与输出结果