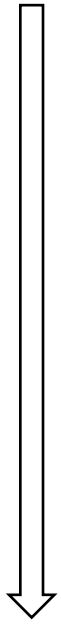
第七讲 贝叶斯分类器

高级机器学习

提纲

- 贝叶斯决策论
- 朴素贝叶斯（拉普拉斯修正）
- 半朴素贝叶斯
- 贝叶斯网



泛化性能越来越好
当然，模型也越来越精细

贝叶斯决策论（Bayesian decision theory）

概率框架下实施决策的基本理论

给定 N 个类别，令 λ_{ij} 代表将第 j 类样本误分类为第 i 类所产生的损失，则基于后验概率将样本 **x** 分到第 i 类的条件风险为：

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

贝叶斯判定准则 (Bayes decision rule)：

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

- h* 称为贝叶斯最优分类器(Bayes optimal classifier)，其总体风险称为贝叶斯风险 (Bayes risk)
- 反映了学习性能的理论上限

贝叶斯分类器

从这个角度来看，机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率

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

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

Thomas Bayes
(1701?-1761)

联合概率的计算将遭遇：

- 组合爆炸
- 样本稀疏
- ……

无法直接求解，需要设计有效/高效的算法

→ 机器学习

（考虑计算复杂度、样本复杂度……的数据分析）

判别式 vs 生成式

$P(c \mid \boldsymbol{x})$ 在现实中通常难以直接获得

两种基本策略：

判别式 (discriminative) 模型

思路：直接对 $P(c \mid \boldsymbol{x})$ 建模

代表：

- 决策树
- BP 神经网络
- SVM

生成式 (generative) 模型

思路：先对联合概率分布 $P(\boldsymbol{x}, c)$ 建模，再由此获得 $P(c \mid \boldsymbol{x})$

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

代表：贝叶斯分类器

注意：贝叶斯分类器 ≠ 贝叶斯学习 (Bayesian learning)

朴素贝叶斯分类器 (naïve Bayes classifier)

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

主要障碍：所有属性上的联合概率难以从有限训练样本估计获得

组合爆炸；样本稀疏

基本思路：假定属性相互独立？

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

d 为属性数，x_i 为 **x** 在第 i 个属性上的取值

$P(\boldsymbol{x})$ 对所有类别相同，于是

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

朴素贝叶斯分类器

- 估计 P(c): $P(c) = \frac{|D_c|}{|D|}$
- 估计 P(**x**|c):
  - 对离散属性，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

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

  - 对连续属性，考虑概率密度函数，假定 $p(x_i \mid c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$

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

拉普拉斯修正 (Laplacian Correction, LC)

若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题，因为概率连乘将“抹去”其他属性提供的信息

例如，训练时未出现“敲声=清脆”的好瓜，
测试时遇到“敲声=清脆”的样本……

令 N 表示训练集 D 中可能的类别数，N_i 表示第 i 个属性可能的取值数

假设属性值随类别呈均匀分布

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

理论上其它合理假设都可以

朴素贝叶斯分类器的使用

- 若对预测速度要求高

  → 预计算所有概率估值，使用时“查表”

- 若数据更替频繁

  → 不进行任何训练，收到预测请求时再估值

  (懒惰学习, lazy learning)

- 若数据不断增加

  → 基于现有估值，对新样本涉及的概率估值进行修正

  (增量学习, incremental learning)

半朴素贝叶斯分类器

朴素贝叶斯分类器的“属性独立性假设”在现实中往往难以成立

半朴素贝叶斯分类器 (semi-naïve Bayes classifier)

基本思路：适当考虑一部分属性间的相互依赖信息

最常用策略：独依赖估计 (One-Dependent Estimator, ODE)

假设每个属性在类别之外最多仅依赖一个其他属性

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

x_i 的“父属性”

关键是如何确定父属性

两种常见方法

- SPODE (Super-Parent ODE)：

  假设所有属性都依赖于同一属性，称为“超父” (Super-Parent)，然后通过交叉验证等模型选择方法来确定超父属性

- TAN (Tree Augmented naïve Bayes)：

  以属性间的条件”互信息”(mutual information)为边的权重，构建完全图，再利用最大带权生成树算法，仅保留强相关属性间的依赖性

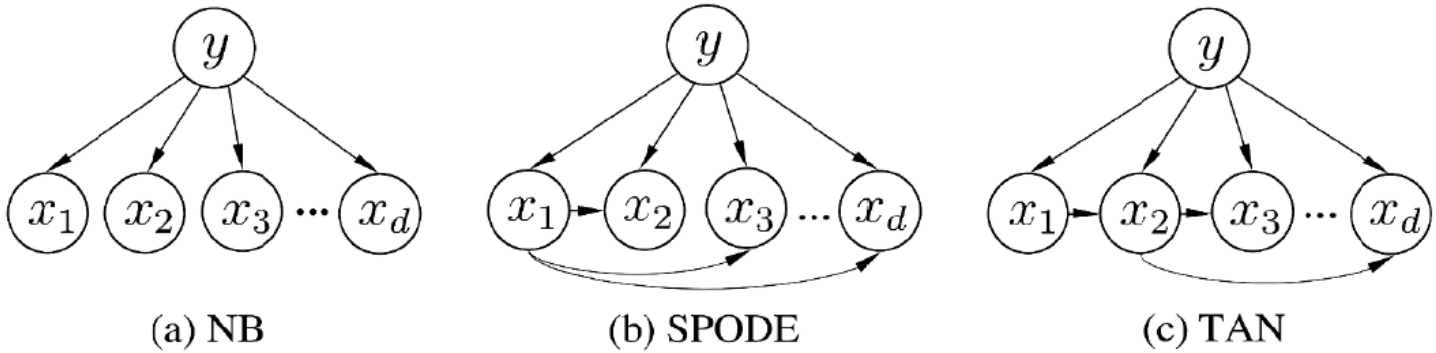


图 7.1 朴素贝叶斯与两种半朴素贝叶斯分类器所考虑的属性依赖关系

AODE (Averaged One-Dependent Estimator)

Geoff Webb
澳大利亚Monash大学

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

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

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

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

D_{c,x_i,x_j} 表示类别为 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

高阶依赖

能否通过考虑属性间的高阶依赖来进一步提升泛化性能？

最简单的做法：ODE → kDE

将父属性 pa_i 替换为包含 k 个属性的集合 $\mathbf{pa}_i$

明显障碍：随着 k 的增加，估计 $P(x_i \mid y, \mathbf{pa}_i)$ 所需的样本数将以指数级增加

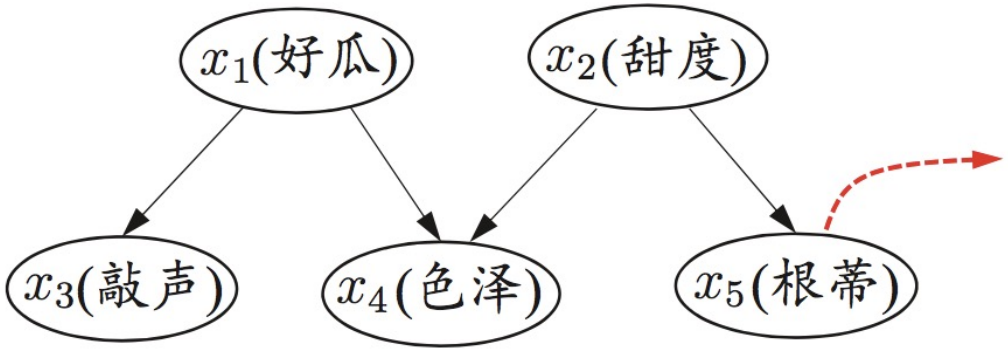
- 训练样本非常充分 → 性能可能提升
- 有限训练样本 → 高阶联合概率估计困难

如何高效且有效地利用属性间的高阶依赖？

贝叶斯网（Bayesian network）

亦称“信念网”(brief network)

有向无环图(DAG, Directed Acyclic Graph)



条件概率表（CPT, Conditional Probability Table）

		根蒂：硬挺	根蒂：蜷缩
甜度	高	0.1	0.9
甜度	低	0.7	0.3

贝叶斯网 $B = \langle G, \Theta \rangle$

G：结构　Θ：参数

1985年 J. Pearl 命名为贝叶斯网，为了强调：

- 输入信息的主观本质
- 对贝叶斯条件的依赖性
- 因果与证据推理的区别

Judea Pearl (1936 -)
2011 图灵奖

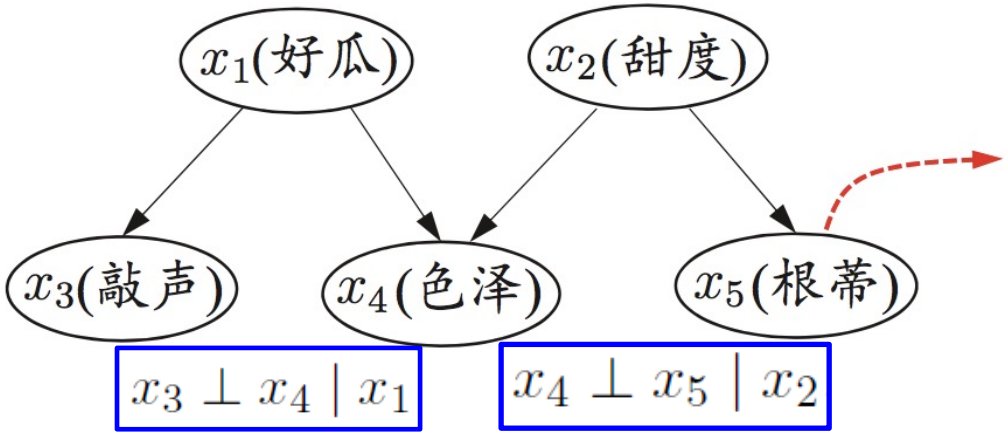
概率图模型(Probabilistic graphical model)

- 有向图模型 → 贝叶斯网
- 无向图模型 → 马尔可夫网

→ 第14章

贝叶斯网（Bayesian network）

有向无环图（DAG, Directed Acyclic Graph）



条件概率表（CPT, Conditional Probability Table）

		根蒂：硬挺	根蒂：蜷缩
甜度	高	0.1	0.9
甜度	低	0.7	0.3

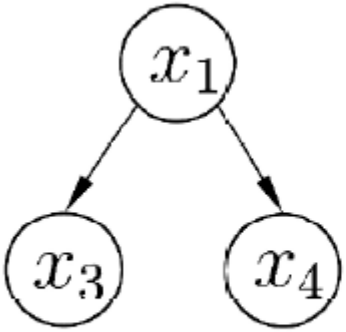
给定父结点集，贝叶斯网假设每个属性与其非后裔属性独立

$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$

π_i：父结点集

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3 \mid x_1)P(x_4 \mid x_1, x_2)P(x_5 \mid x_2)$$

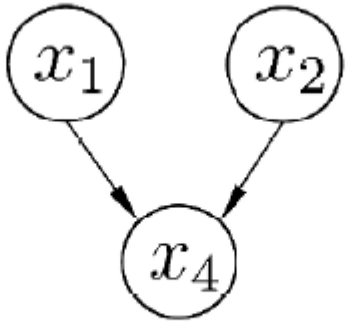
三变量间的典型依赖关系



同父结构

条件独立性

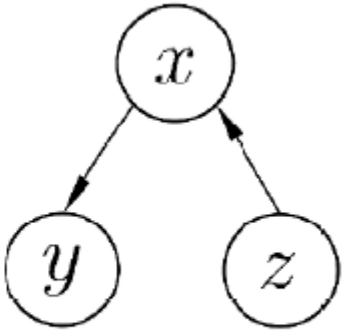
$x_3 \perp x_4 \mid x_1$



V型结构

边际独立性

$x_1 \perp\!\!\!\perp x_2$



顺序结构

条件独立性

$y \perp z \mid x$

- 给定 x_4， x_1 与 x_2 必不独立
- 若 x_4 未知，则 x_1 与 x_2 独立

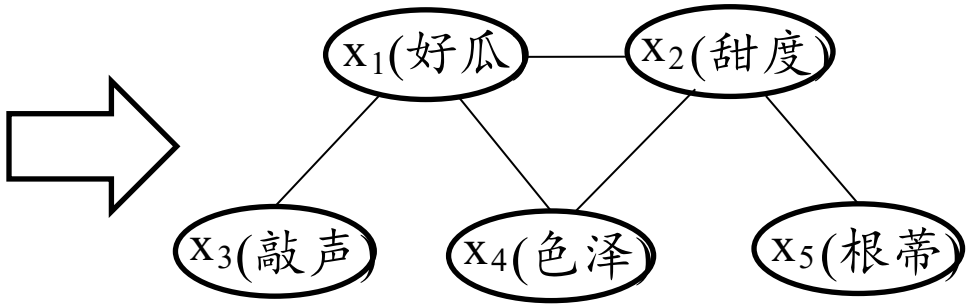
分析条件独立性

“有向分离”（D-separation）

先将有向图转变为无向图

- V 型结构父结点相连
- 有向边变成无向边

道德图
(moral graph)



若 x 和 y 能在图上被 **z** 分入两个连通分支，则有

$$x \perp y \mid \mathbf{z}$$

由图可得：

$$x_3 \perp x_4 \mid x_1$$
$$x_4 \perp x_5 \mid x_2$$
$$x_3 \perp x_2 \mid x_1$$
$$x_3 \perp x_5 \mid x_1$$
$$x_3 \perp x_5 \mid x_2$$

得到条件独立性关系之后，估计出条件概率表，就得到了最终网络

小结

- 掌握贝叶斯决策论
- 熟悉朴素贝叶斯（拉普拉斯修正）
- 掌握半朴素贝叶斯
- 了解贝叶斯网