贝叶斯分类器

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法。在相关概率已知的情形下，基于概率和误判损失来选择最优的类别标记
- 条件风险（conditional risk）：
  - 假设有 N 种可能的类别标记，即 $y=\{c_1,c_2,\ldots,c_N\}$ ，λ_{ij} 是将一个真实标记为 c_j 的样本误分类为 c_i 所产生的损失。基于后验概率 $P\{c_i \mid \mathbf{x}\}$ 可获得将样本 $\mathbf{x}$ 分类为 c_i 所产生的期望损失（expected loss），即在样本上的“条件风险”（conditional risk）

    $$R\left(c_i \mid \mathbf{x}\right)=\sum_{j=1}^{N} \lambda_{ij} P\left(c_j \mid \mathbf{x}\right) \qquad (7.1)$$
  - 任务就是寻找一个判定准则$h: X \rightarrow Y$ 来最小化总体风险
    - $R(h)=E_x[R(h(x)|x)]$
- 贝叶斯判定准则（Bayes decision rule）：
  - $h^*(x)=argmin R(c|x) c \in y$
  - 以此方法确定的分类器是贝叶斯最优分类器（Bayes optimal classifier），$1-R(h^*)$ 称为反映了分类器所能达到的最好性能，即理论上限
  - 因为若目标是最小化分类错误率，则$\lambda_{i,j}$
  - $\lambda_[i,j]=\begin{cases}0 & \text{i=j}\\ 1 & \text{otherwise}\end{cases}$
  - 此时条件风险:$R(c|x)=1-P(c|x)$
- 使用此方法判定的关键就是得到后验概率$P（c|x)$ 但是很难在现实中直接获得，机器学习要做的就是基于有限的样本，尽可能准确的估计$P（c|x)$。主要采用的策略有两种：**判别式模型、生成式模型**
  - 判别式模型（discriminative models):
    - 给定x，直接建模P（c|x) 预测c
    - 决策树、BP神经网络、支持向量机
  - 生成式模型：
    - 生成式模型

      $$P(c \mid \mathbf{x})=\frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \qquad (7.7)$$

      基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

      类标记 c 相对于样本 $\mathbf{x}$ 的“类条件概率”（class-conditional probability），或称“似然”。

      $$P(c \mid \mathbf{x})=\frac{\boxed{P(c)}\boxed{P(\mathbf{x} \mid c)}}{\boxed{P(\mathbf{x})}} \qquad (7.8)$$

      先验概率
      样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

      “证据”（evidence）因子，与类标记无关

极大似然估计

极大似然估计就是在某种i情况出现时，认为事件A发生的概率和某一未知参数θ有关，而要做的就是求取使得$P(A|\theta)$ 达到最大值的θ 。此处想求取P(X|C)的值，而且默认P(X|C) 被参数θ 唯一确定，所以要用极大似然估计求取该θ

> 记关于类别 c 的类条件概率为 $P(\mathbf{x} \mid c)$ ，
> - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集 D 估计参数 $\boldsymbol{\theta}_c$
>
> 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同的方案：
> - 频率主义学派 (frequentist)认为参数虽然未知，但却存在客观值，因此可通过优化似然函数等准则来确定参数值
> - 贝叶斯学派 (Bayesian)认为参数是未观察到的随机变量、其本身也可由分布，因此可假定参数服从一个先验分布，然后基于观测到的数据计算参数的后验分布。

朴素贝叶斯分类器

采用：属性条件独立性假设；认为每个属性独立地对分类结果产生影响。

因此P(C|X) 重写为：

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.14)$$

拉普拉斯修正：

- 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，. 比如"敲声=清脆"测试例，训练集中没有该样例，因此连 乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是 "好瓜=否"，这显然不合理。为了避免其他属性携带的信息被训练集中未出现的属性值"抹去"， 在估计概率值时通常要进行"拉普拉斯修正"（Laplacian correction）
  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，则式 (7. 16)和 (7. 17)分别修正为
- $$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \quad (7.19) \qquad \hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \quad (7.20)$$

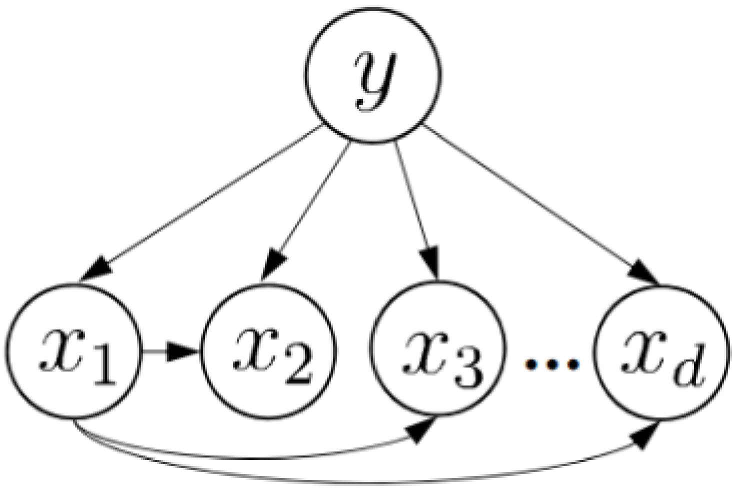
半朴素贝叶斯分类器

像朴素贝叶斯分类器，属性条件独立性假设太过于理想。半朴素贝叶斯分类器对于该假设的严格程度，进行了一些放松。最常用的策略是：独依赖估计（one dependent estimator）。假设每个属性在类别之外最多仅依赖一个其他属性即：

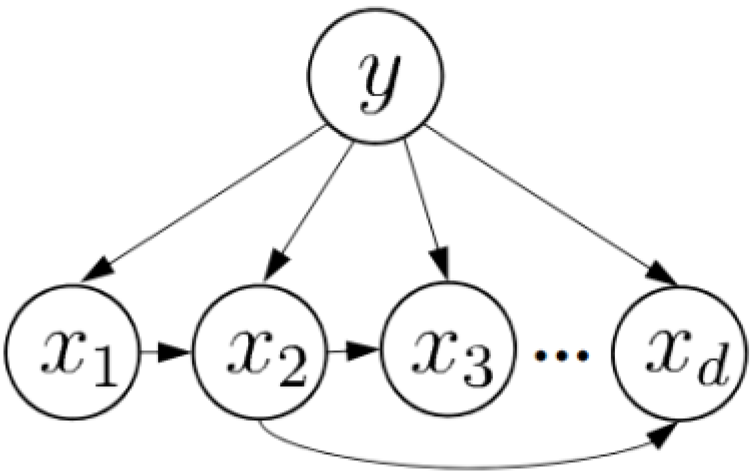
$$P(c \mid x) \propto P(c) \prod_{i=1}^{d} P(x_i \mid c, pa_i)$$

常见的方法有/：

- SPODE：最直接的做法是假设所有属性都依赖于同一属性，称为"超父" (super-parenet)，然后通过交叉验证等模型选择方法来确定超父属 性，由此形成了SPODE (Super-Parent ODE)方法。

- 

(b) SPODE

- TAN (Tree augmented Naïve Bayes) [Friedman et al., 1997] 则在最大带权生成树 (Maximum weighted spanning tree) 算法 [Chow and Liu, 1968] 的基础上，通过以下步骤将属性间依赖关系简约；
- 

(c) TAN

贝叶斯网

贝叶斯网 (Bayesian network)亦称“信念网”(brief network)， 它借助有向无环图 (Directed Acyclic Graph, DAG)来刻画属性 间的依赖关系，并使用条件概率表 (Conditional Probability Table, CPT)来表述属性的联合概率分布。

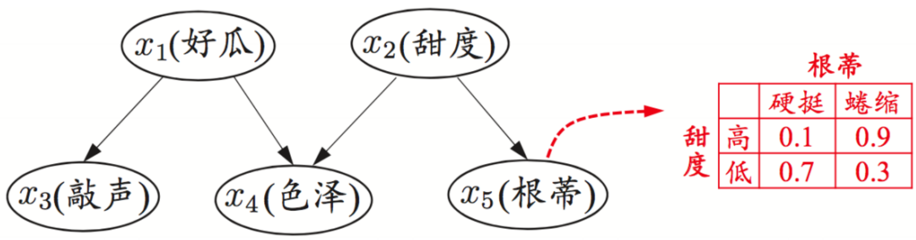
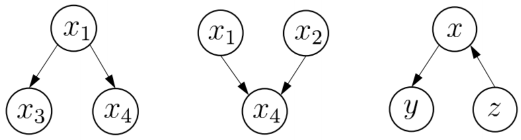


图7.2 西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

- 贝叶斯网有效地表达了属性间的条件独立性。给定父结集，贝叶斯 网假设每个属性与他的非后裔属性独立。

- 从网络图结构可以看出 -> “色泽”直接依赖于“好瓜”和“甜度”
- 从条件概率表可以得到 -> “根蒂”对“甜度”的量化依赖关系
  P(根蒂＝硬挺|甜度＝高)＝0.1

贝叶斯网中三个典型依赖关系：



同父结构　　V型结构　　顺序结构

贝叶斯网的首要任务是根据训练集找出结构最“恰当”的贝叶斯网。

- 通过已知变量观测值来推测待推测查询变量的过程称为“推断” (inference) ，已知变量观测值称为“证据” (evidence)。 最理想的是根据贝叶斯网络定义的联合概率分布来精确计算后验概 率，在现实应用中，贝叶斯网的近似推断常使用吉布斯采样(Gibbs sampling)来完成。

吉布斯采样随机产生一个与证据 $E=e$ 一致的样本 q^0 作为初始点，然后每步从当前样本出发产生下一个样本。假定经过 T 次采样的得到与 q 一致的样本共有 n_q 个，则可近似估算出后验概率

$$P(Q=q \mid E=e) \simeq \frac{n_q}{T} \qquad (7.33)$$

EM算法

EM算法针对不完整的样本提出；未观测的变量称为“隐变量”(latent variable)。令X表示已观 测变量集， Z表示隐变量集，若预对模型参数Θ做极大似然估计， 则应最大化对数似然函数:

$LL(\Theta|X,Z)=ln(P(X,Z)|\Theta)$

EM算法用于估计隐变量。

- 当参数Θ已知 -> 根据训练数据推断出最优隐变量 $\mathbf{Z}$ 的值(E步)
- 当$\mathbf{Z}$已知 -> 对Θ做极大似然估计(M步)

于是，以初始值Θ^0为起点，对式子(7.35)，可迭代执行以下步骤直至收敛：

- 基于Θ^t 推断隐变量 $\mathbf{Z}$ 的期望，记为 $\mathbf{Z}^t$ ；
- 基于已观测到变量 $\mathbf{X}$和 $\mathbf{Z}^t$ 对参数 Θ做极大似然估计，记为Θ^{t+1}；

- 这就是EM算法的原型。
