贝叶斯决策论

- 显然，对每个样本$\mathbf{x}$，若 h 能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$ ，则总体风险 $R(h)$ 也将被最小化。
- 贝叶斯判定准则（Bayes decision rule）： 为最小化总体风险，只需在每个样本上选择那个能使条件风险 $R(c \mid \mathbf{x})$ 最小的类别标记，即

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c \mid x)$$

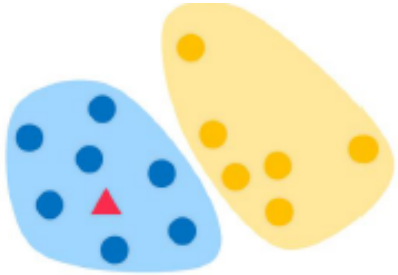
  - 此时，被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$ 称为贝叶斯风险 (Bayes risk)
  - $1 - R(h^*)$ 反映了分类器所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限。

贝叶斯决策论

- 不难看出，使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c \mid \mathbf{x})$ 。
- 然而，在现实中通常难以直接获得。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 $P(c \mid \mathbf{x})$。
- 主要有两种策略：
  - 判别式模型（discriminative models）
    - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$，来预测 c
    - 决策树，BP神经网络，支持向量机
  - 生成式模型（generative models）
    - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
    - 生成式模型考虑 $P(c \mid \mathbf{x}) = \dfrac{P(\mathbf{x}, c)}{P(\mathbf{x})}$

二者对比

● 生成式模型



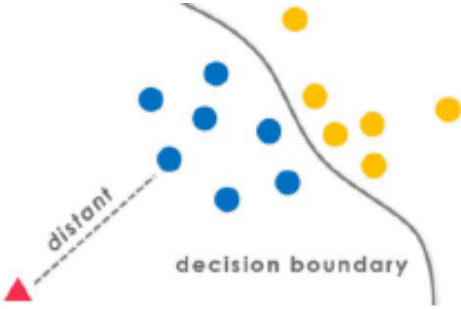
优点：

- 信息丰富
- 单类问题灵活性强
- 增量学习
- 合成缺失数据

缺点：

- 学习过程复杂
- 为分布牺牲分类性能

● 判别式模型



优点：

- 类间差异清晰
- 分类边界灵活
- 学习简单
- 性能较好

缺点：

- 不能反应数据特性
- 需要全部数据进行学习

由生成模型可以得到判别模型，
但由判别模型得不到生成模型。

朴素贝叶斯分类器

朴素贝叶斯分类器

- 估计后验概率 $P(c \mid \mathbf{x})$ 主要困难：类条件概率 $P(\mathbf{x} \mid c)$ 是所有属性上的联合概率难以从有限的训练样本估计获得。

- 朴素贝叶斯分类器(Naïve Bayes Classifier)采用了“属性条件独立性假设”(attribute conditional independence assumption)：每个属性独立地对分类结果发生影响。

- 基于属性条件独立性假设，后验概率可重写为

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

  - 其中 d 为属性数目，x_i 为 $\mathbf{x}$ 在第 i 个属性上的取值。

朴素贝叶斯分类器

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

由于对所有类别来说 $P(x)$ 相同，因此上式的贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

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

朴素贝叶斯分类器

- 令 D_c 表示训练集 D 中第 c 类样本组合的集合，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

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

- 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

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

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

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

举例

■ 训练样本

No.	天气	气温	湿度	风	*类别*
1	晴	热	高	无	*0*
2	晴	热	高	有	*0*
3	云	暖	高	无	*1*
4	雨	暖	高	无	*1*
5	雨	冷	正常	无	*1*
6	雨	冷	正常	有	*0*
7	云	冷	正常	有	*1*

No.	天气	气温	湿度	风	*类别*
8	晴	暖	高	无	*0*
9	晴	冷	正常	无	*1*
10	雨	暖	正常	无	*1*
11	晴	暖	正常	有	*1*
12	云	暖	高	有	*1*
13	云	热	正常	无	*1*
14	雨	暖	高	有	*0*

■ 测试

x_1	x_2	x_3	x_4	w
天气	温度	湿度	有风	打网球
晴	凉	高	是	?

举例

■ 训练样本

No.	天气	气温	湿度	风	类别
1	晴	热	高	无	0
2	晴	热	高	有	0
3	云	暖	高	无	1
4	雨	暖	高	无	1
5	雨	冷	正常	无	1
6	雨	冷	正常	有	0
7	云	冷	正常	有	1

No.	天气	气温	湿度	风	类别
8	晴	暖	高	无	0
9	晴	冷	正常	无	1
10	雨	暖	正常	无	1
11	晴	暖	正常	有	1
12	云	暖	高	有	1
13	云	热	正常	无	1
14	雨	暖	高	有	0

■ 统计结果

	打网球	
	1	0
$P(w)$	9/14	5/14

举例

■ 训练样本

No.	天气	气温	湿度	风	类别
1	晴	热	高	无	0
2	晴	热	高	有	0
3	云	暖	高	无	1
4	雨	暖	高	无	1
5	雨	冷	正常	无	1
6	雨	冷	正常	有	0
7	云	冷	正常	有	1

No.	天气	气温	湿度	风	类别
8	晴	暖	高	无	0
9	晴	冷	正常	无	1
10	雨	暖	正常	无	1
11	晴	暖	正常	有	1
12	云	暖	高	有	1
13	云	热	正常	无	1
14	雨	暖	高	有	0

■ 统计结果

天气	
1	0
晴 2/9	3/5
云 4/9	0/5
雨 3/9	2/5

$P(x_1 \mid w)$

举例

■ 模型(查询表)

$P(x_1 \mid w)$ $P(x_2 \mid w)$ $P(x_3 \mid w)$ $P(x_4 \mid w)$ $P(w)$

天气		温度		湿度		有风		打网球	
1	0	1	0	1	0	1	0	1	0
晴 2/9	3/5	热 1/9	2/5	高 3/9	4/5	否 6/9	2/5	9/14	5/14
云 4/9	0/5	暖 5/9	2/5	正常 6/9	1/5	是 3/9	3/5		
雨 3/9	2/5	凉 3/9	1/5						

■ 测试

x_1 x_2 x_3 x_4 w

天气	温度	湿度	有风	打网球
晴	凉	高	是	?

- $P(\mathbf{x} \mid w=1)P(w=1) = \left[\prod_{k=1}^{4} P(x_k \mid w=1)\right] P(w=1) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$
- $P(\mathbf{x} \mid w=0)P(w=0) = \left[\prod_{k=1}^{4} P(x_k \mid w=0)\right] P(w=0) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}$

举例

■ 模型(查询表)

$P(x_1 \mid w)$ $P(x_2 \mid w)$ $P(x_3 \mid w)$ $P(x_4 \mid w)$ $P(w)$

天气		温度		湿度		有风		打网球	
1	0	1	0	1	0	1	0	1	0
晴 2/9	3/5	热 1/9	2/5	高 3/9	4/5	否 6/9	2/5	9/14	5/14
云 4/9	0/5	暖 5/9	2/5	正常 6/9	1/5	是 3/9	3/5		
雨 3/9	2/5	凉 3/9	1/5						

■ 测试

x_1 x_2 x_3 x_4 w

天气	温度	湿度	有风	打网球
晴	凉	高	是	?

$$\frac{P(\mathbf{x} \mid w=1)P(w=1)}{P(\mathbf{x} \mid w=0)P(w=0)} = \frac{125}{486} < 1 \quad \Rightarrow \quad w=0$$

拉普拉斯修正

■ 模型(查询表)

$P(x_1 \mid w)$ $P(x_2 \mid w)$ $P(x_3 \mid w)$ $P(x_4 \mid w)$ $P(w)$

天气		温度		湿度		有风		打网球	
1	0	1	0	1	0	1	0	1	0
晴 2/9	3/5	热 1/9	2/5	高 3/9	4/5	否 6/9	2/5	9/14	5/14
云 4/9	0/5	暖 5/9	2/5	正常 6/9	1/5	是 3/9	3/5		
雨 3/9	2/5	凉 3/9	1/5						

■ 测试

x_1 x_2 x_3 x_4 w

天气	温度	湿度	有风	打网球
云	热	高	是	?

- $P(\mathbf{x} \mid w=1)P(w=1) = \left[\prod_{k=1}^{4} P(x_k \mid w=1)\right] P(w=1) = \frac{4}{9} \times \frac{1}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$
- $P(\mathbf{x} \mid w=0)P(w=0) = \left[\prod_{k=1}^{4} P(x_k \mid w=0)\right] P(w=0) = \frac{0}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}$?

拉普拉斯修正

- 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 $P(c)$ 和 $P(x_i \mid c)$ 分别修正为

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

- 现实任务中，朴素贝叶斯分类器的使用：
  - 速度要求高，“查表”；
  - 任务数据更替频繁，“懒惰学习”（lazy learning）；
  - 数据不断增加，增量学习等等。

半朴素贝叶斯分类器

半朴素贝叶斯分类器

- 朴素贝叶斯分类器：属性条件独立性假设
- 半朴素贝叶斯分类器（semi-naïve Bayes classifiers）：对属性条件独立假设记性一定程度的放松
- 最常用的一种策略：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，即

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

  - 其中 pa_i 为属性 x_i 所依赖的属性，称为 x_i 的父属性
- 对每个属性 x_i，若其父属性 pa_i 已知，则可估计概值 $P(x_i \mid c, pa_i)$，于是问题的关键转化为如何确定每个属性的父属性。

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。

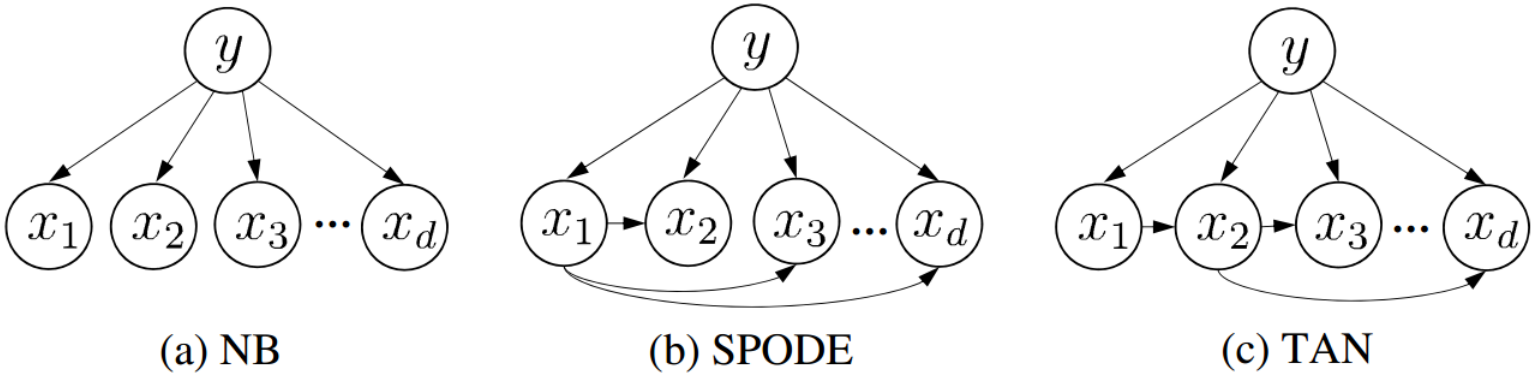


图7.1 朴素贝叶斯分类器与两种半朴素分类器所考虑的属性依赖关系

- 在图7.1 (b)中，x_1 是超父属性。

AODE

- AODE (Averaged One-Dependent Estimator) [Webb et al. 2005] 是一种基于集成学习机制、更为强大的分类器。
  - 尝试将每个属性作为超父构建 SPODE
  - 将具有足够训练数据支撑的SPODE集成起来作为最终结果

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

其中，D_{x_i} 是在第 i 个属性上取值 x_i 的样本的集合，m' 为阈值常数

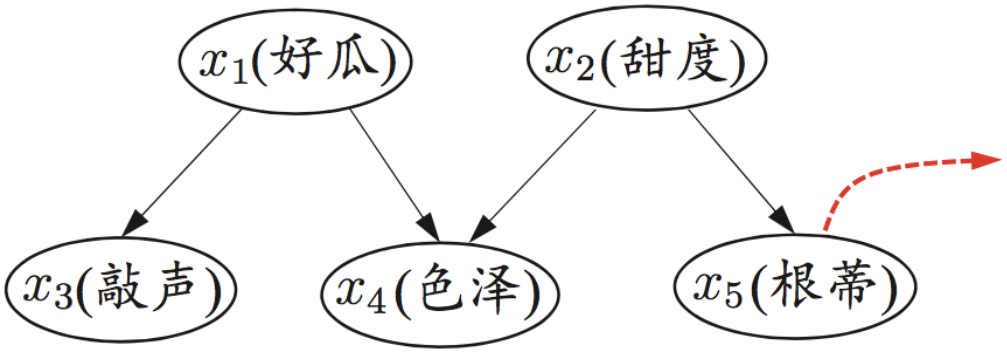
$$\hat{P}(x_i, c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \qquad \hat{P}(x_j \mid c, x_i) = \frac{|D_{c,x_i,x_j}| + 1}{|D_{c,x_i}| + N_j}$$

其中，N_i 是在第 i 个属性上取值数，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，

贝叶斯网

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。

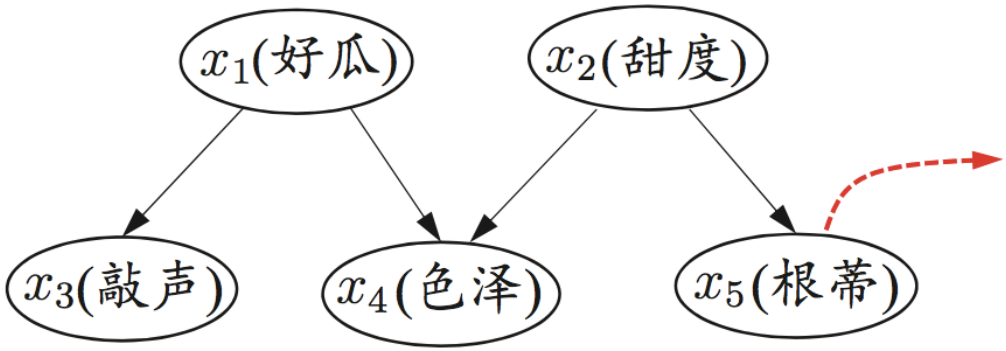


根蒂

甜度	硬挺	蜷缩
高	0.1	0.9
低	0.7	0.3

- 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 $P(\text{根蒂}=\text{硬挺}|\text{甜度}=\text{高})=0.1$

贝叶斯网



根蒂

甜度	硬挺	蜷缩
高	0.1	0.9
低	0.7	0.3

- 贝叶斯网B由结构和参数组成：$B = \langle G, \Theta \rangle$
  - G：有向无环图，每个节点对应一个属性，边表示属性间的依赖关系
  - Θ：包含每个属性的条件概率表 $\theta_{x_i|\pi_i}$
    - 假设属性x_i在G中的父节点集为π_i，则

$$\theta_{x_i|\pi_i} = P_B(x_i \mid \pi_i)$$

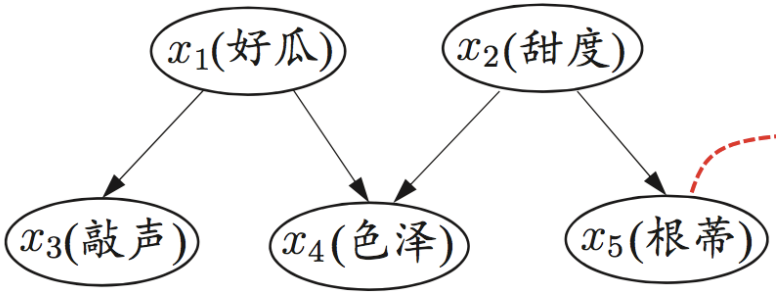
贝叶斯网：结构

- 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。

- $B = \langle G, \Theta \rangle$ 将属性 $x_1, x_2, \cdots, x_d$ 的联合概率分布定义为

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

右图的联合概率分布定义为：

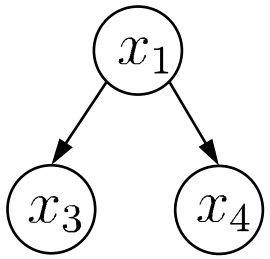


$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3 \mid x_1)P(x_4 \mid x_1, x_2)P(x_5 \mid x_2)$$

- 显然，x_3 和 x_4 在给定 x_1 的取值时独立，x_4 和 x_5 在给定 x_2 的取值时独立，记为 $x_3 \perp x_4 \mid x_1$ 和 $x_4 \perp x_5 \mid x_2$ 。

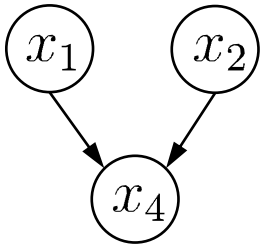
贝叶斯网：结构

- 贝叶斯网中三个变量之间的典型依赖关系：



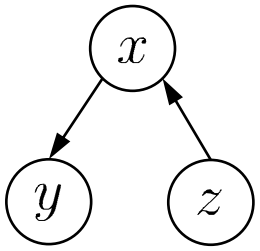
同父结构

$x_3 \perp x_4 \mid x_1$



V型结构

冲撞结构



顺序结构

$y \perp z \mid x$

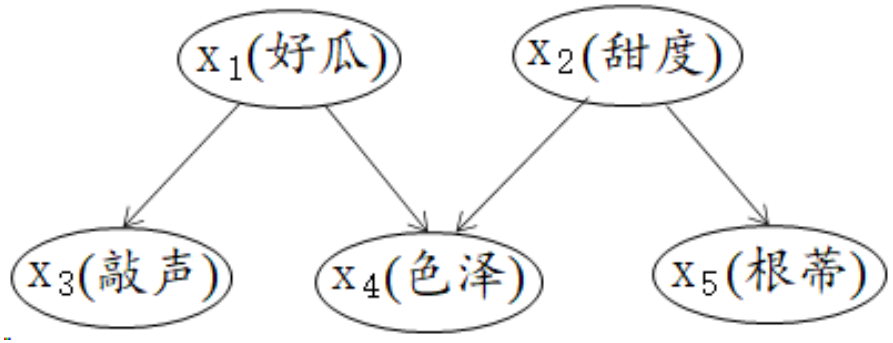
- 给定x_4的取值， x_1与x_2必不独立
- 若x_4的取值完全未知， x_1与x_2则是相互独立的

- 边际独立性，记为$x_1 \perp\!\!\!\perp x_2$

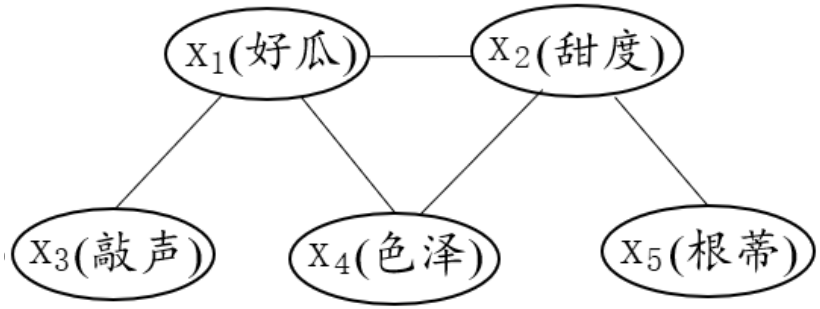
贝叶斯网：结构

- 分析有向图中变量间的条件独立性，可使用“有向分离”把有向图变为无向图
  - V型结构父结点相连
  - 有向边变成无向边

  由此产生的图称为道德图



- 若将x_1去掉，x_3和x_4分属两个连通分支，则称x_3和x_4被x_1有向分离，$x_3 \perp x_4 \mid x_1$ 成立。
- 从右图中找出所有条件独立关系。



EM算法

EM算法

- “不完整”的样本：西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”，则训练样本的“根蒂”属性变量值未知，如何计算？

- 未观测的变量称为“隐变量”（latent variable）。令 $\mathbf{X}$ 表示已观测变量集，$\mathbf{Z}$ 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

$$LL(\mathbf{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta})$$

- 由于 $\mathbf{Z}$ 是隐变量，上式无法直接求解。此时我们可以通过对 $\mathbf{Z}$ 计算期望，来最大化已观测数据的对数“边际似然”（marginal likelihood）

$$LL(\mathbf{\Theta} \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \mathbf{\Theta}) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta})$$

EM算法

EM (Expectation-Maximization)算法 [Dempster et al., 1977] 是常用的估计参数隐变量的利器。

- 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，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。

小结

- 贝叶斯分类原则：先验，似然，后验
  - 最小错误率贝叶斯：最大后验概率分类
  - 最小风险贝叶斯：最小期望风险分类
- 朴素贝叶斯分类器：条件独立假设，拉普拉斯平滑
- 半朴素贝叶斯分类器：对条件独立假设的放松
- 贝叶斯网：概率图模型，分析条件独立
- **EM**算法：估计参数隐变量，两步交替迭代