周志华 著

MACHINE LEARNING

机器学习

清华大学出版社

本章课件致谢：
霍轩

第七章：贝叶斯分类器

章节目录

章节目录

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。
  - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策考虑如何基于这些概率和误判损失来选择最优的类别标记。

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。
  - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策考虑如何基于这些概率和误判损失来选择最优的类别标记。

- 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）

$$R(c_i \mid \mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \mathbf{x}) \qquad (7.1)$$

- 我们的任务是寻找一个判定准则 $h : X \mapsto Y$ 以最小化总体风险

$$R(h) = \mathbf{E}_x \left[R(h(\mathbf{x}) \mid \mathbf{x}) \right] \qquad (7.2)$$

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。

- 这就产生了贝叶斯判定准则（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

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c \mid x) \qquad (7.3)$$

  - 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)
  - $1 - R(h^*)$ 反映了分类起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

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij}可写为

$$\lambda_{i,j}\begin{cases} 0, & \text{if } i=j; \\ 1, & \text{otherwise}, \end{cases} \tag{7.4}$$

贝叶斯决策论

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij}可写为

$$\lambda_{i,j} \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}, \end{cases} \tag{7.4}$$

- 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x}) \tag{7.5}$$

贝叶斯决策论

- 具体来说，若目标是最小化分类错误率，则误判损失 λ_{ij}可写为

$$\lambda_{i,j} \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}, \end{cases} \tag{7.4}$$

- 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x}) \tag{7.5}$$

- 于是，最小化分类错误率的贝叶斯最有分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c \mid x) \tag{7.6}$$

  - 即对每个样本 $\mathbf{x}$ ，选择能使后验概率 $P(c \mid \mathbf{x})$ 最大的类别标记。

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

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

贝叶斯决策论

- 生成式模型

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

贝叶斯决策论

- 生成式模型

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

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

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

贝叶斯决策论

- 生成式模型

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

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

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

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）

贝叶斯决策论

- 生成式模型

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

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

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

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

贝叶斯决策论

- 生成式模型

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

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

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

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

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

章节目录

极大似然估计

- 估计类条件概率的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计。

- 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 $\boldsymbol{\theta}_c$

极大似然估计

- 估计类条件概率的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计。

- 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。

极大似然估计

- 令 D_c 表示训练集中第 c 类样本的组合的集合，假设这些样本是独立的，则参数 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c}) \qquad (7.9)$$

  - 对 $\boldsymbol{\theta}_c$ 进行极大似然估计，寻找能最大化似然 $P(D_c \mid \theta_c)$ 的参数值 $\hat{\theta}_c$ 。直观上看，极大似然估计是试图在 θ_c 所有可能的取值中，找到一个使数据出现的“可能性”最大值。

极大似然估计

- 令 D_c 表示训练集中第 c 类样本的组合的集合，假设这些样本是独立的，则参数 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c}) \qquad (7.9)$$

  - 对 $\boldsymbol{\theta}_c$ 进行极大似然估计，寻找能最大化似然 $P(D_c \mid \theta_c)$ 的参数值 $\hat{\theta}_c$ 。直观上看，极大似然估计是试图在 θ_c 所有可能的取值中，找到一个使数据出现的“可能性”最大值。

- 式 (7.9) 的连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$\begin{aligned} LL(\boldsymbol{\theta_c}) &= \log P(D_c \mid \boldsymbol{\theta_c}) \\ &= \sum_{\mathbf{x} \in D_c} \log P(\mathbf{x} \mid \boldsymbol{\theta}_c) \end{aligned} \qquad (7.10)$$

- 此时参数 $\boldsymbol{\theta}_c$ 的极大似然估计 $\hat{\boldsymbol{\theta}}_c$ 为 $\hat{\boldsymbol{\theta}}_{\boldsymbol{c}} = \underset{\boldsymbol{\theta}_c}{\operatorname{argmax}} LL(\boldsymbol{\theta}_c)$ (7.11)

极大似然估计

- 例如，在连续属性情形下，假设概率密度函数 $p(\mathbf{x} \mid c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$，则参数 $\boldsymbol{\mu}_c$ 和 $\boldsymbol{\sigma}_c^2$ 的极大似然估计为

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} \mathbf{x} \tag{7.12}$$

$$\hat{\boldsymbol{\sigma}}_c^2 = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} (\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}} \tag{7.13}$$

- 也就是说，通过极大似然法得到的正态分布均值就是样本均值，方差就是 $(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$ 的均值，这显然是一个符合直觉的结果。

- 需注意的是，这种参数化的方法虽能使类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。

章节目录

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，(7.8)可重写为

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

  - 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

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

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，因此基于式 (7.6)的贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c) \qquad (7.15)$$

- 这就是朴素贝叶斯分类起的表达式子

朴素贝叶斯分类器

- 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 $P(x_i \mid c)$ 。
  - 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

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

  - 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

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{D} \qquad (7.17)$$

  - 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

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

朴素贝叶斯分类器

- 例子：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类（p151，西瓜数据集 p84 表4.3）

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

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。

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

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \quad (7.19) \qquad \hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \quad (7.20)$$

- 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。

章节目录

半朴素贝叶斯分类器

- 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用的属性条件独立性假设；对属性条件独立假设记性一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器”（semi-naïve Bayes classifiers）

半朴素贝叶斯分类器

- 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用的属性条件独立性假设；对属性条件独立假设记性一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器”（semi-naïve Bayes classifiers）
- 半朴素贝叶斯分类器最常用的一种策略：“独依赖估计”(One-Dependent Estimator,简称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。

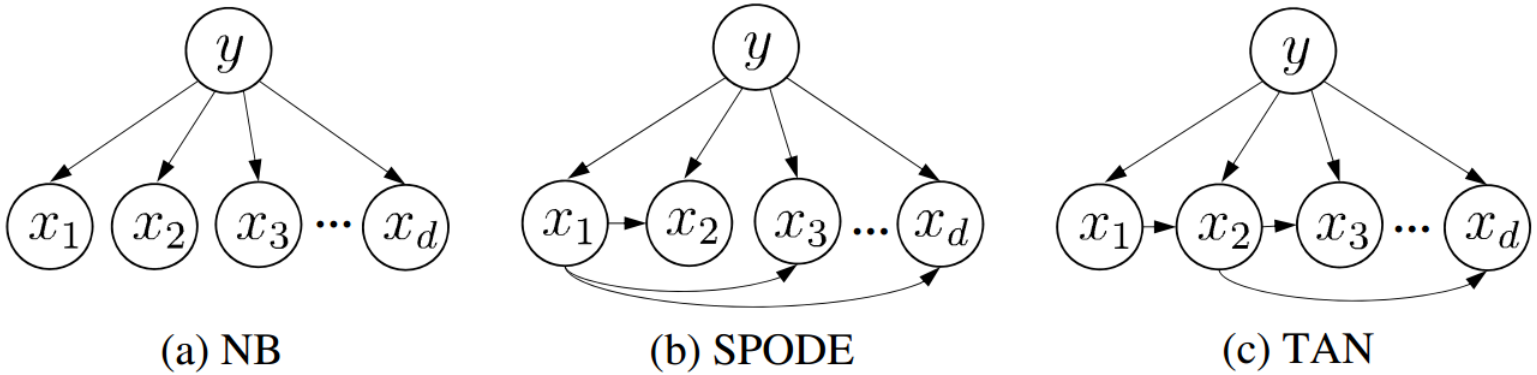


图7.1 朴素贝叶斯分类器与两种半朴素分类器所考虑的属性依赖关系

- 在图7.1（b）中，x_1 是超父属性。

TAN

- 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为图7.1 (c)。

  - 计算任意两个属性之间的条件互信息 (conditional mutual information)

$$I(x_i, x_j \mid y) = \sum_{x_i, x_j; c \in y} P(x_i, x_j \mid c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c) P(x_j \mid c)}$$

  - 以属性为结点构建完全图，任意两个结点之间边的权重设为 $I(x_i, x_j \mid y)$
  - 构建此完全图的最大带权生成树，挑选根变量，将边设为有向；
  - 加入类别节点y，增加从y到每个属性的有向边。

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集群起来作为最终结果

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

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，

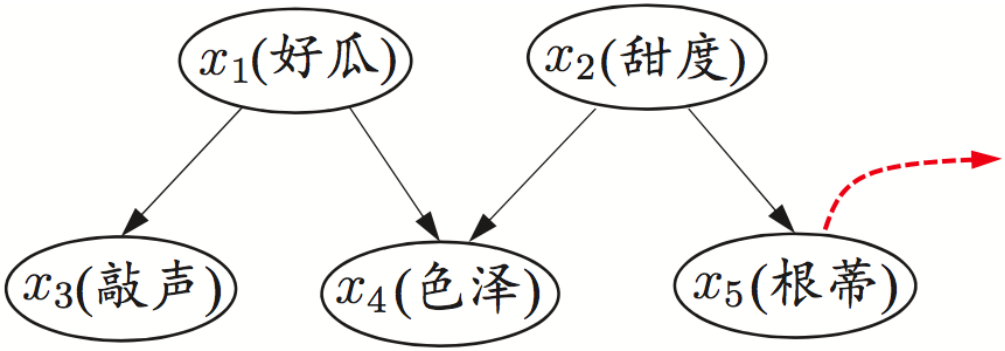
章节目录

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。

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

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

- 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：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

$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i} \qquad (7.26)$$

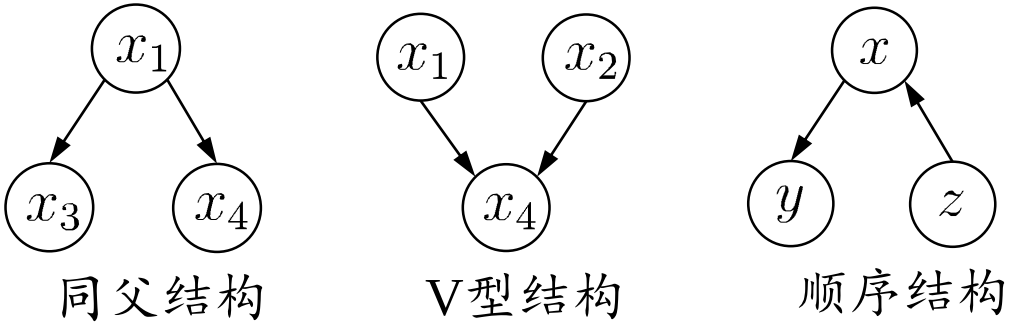
图7.2的联合概率分布定义为：

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

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 $x_4 \perp x_5 \mid x_2$。

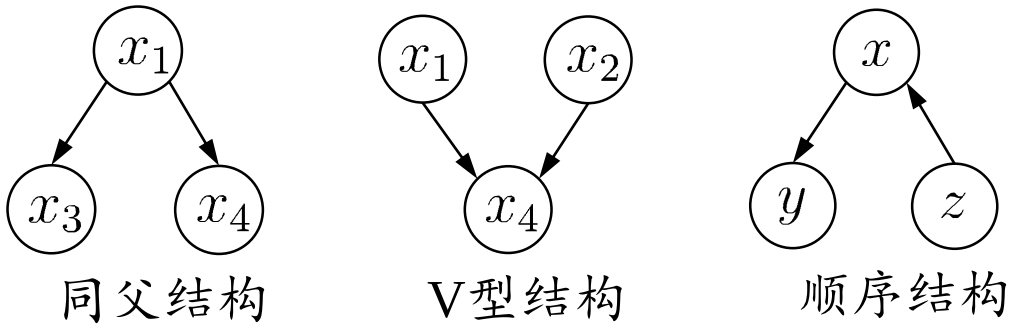
贝叶斯网：结构

- 贝叶斯网中三个变量之间的典型依赖关系：



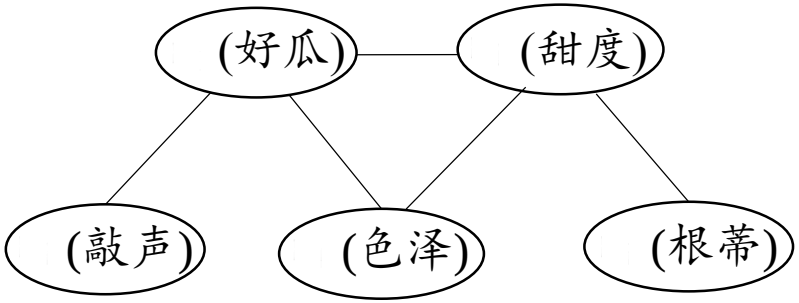
贝叶斯网：结构

- 贝叶斯网中三个变量之间的典型依赖关系：



- 分析有向图中变量间的条件独立性，可使用“有向分离”(D-separation)
  - 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(moral graph)



贝叶斯网：学习

- 贝叶斯网络首要任务：根据训练集找出结构最“恰当”的贝叶斯网。
- 我们用评分函数评估贝叶斯网与训练数据的契合程度。
  - “最小描述长度”（Minimal Description Length, MDL）综合编码长度（包括描述网路和编码数据）最短

贝叶斯网：学习

- 贝叶斯网络首要任务：根据训练集找出结构最“恰当”的贝叶斯网。
- 我们用评分函数评估贝叶斯网与训练数据的契合程度。
  - “最小描述长度”（Minimal Description Length, MDL）综合编码长度（包括描述网路和编码数据）最短

- 给定训练集 $D = \{x_1, x_2, \cdots, x_m\}$，贝叶斯网 $B = \langle G, \Theta \rangle$ 在 D 上的评价函数可以写为

$$S(B \mid D) = f(\theta)|B| - LL(B \mid D) \quad (7.28)$$

  - 其中，$|B|$ 是贝叶斯网的参数个数；$f(\theta)$表示描述每个参数；θ 所需的字节数，而

$$LL(B \mid D) = \sum_{i=1}^{m} \log P_B(x_i) \quad (7.29)$$

  - 是贝叶斯网的对数似然。

贝叶斯网：推断

- 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。

贝叶斯网：推断

- 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

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

- 吉布斯采样可以看做，每一步仅依赖于前一步的状态，这是一个“马尔可夫链”(Markov Chain)。更多马尔可夫链和吉布斯采样内容参见14.5章节。

输入: 贝叶斯网 $B = \langle G, \Theta \rangle$;
采样次数 T;
证据变量 $\mathbf{E}$ 及其取值 $\mathbf{e}$;
待查询变量 $\mathbf{Q}$ 及其取值 $\mathbf{q}$.

过程:

$n_q = 0$
$\mathbf{q}^0 =$ 对 $\mathbf{Q}$ 随机赋初值
for $t = 1, 2, \ldots, T$ **do**
  for $Q_i \in \mathbf{Q}$ **do**
    $\mathbf{Z} = \mathbf{E} \cup \mathbf{Q} \setminus \{Q_i\}$;
    $\mathbf{z} = \mathbf{e} \cup \mathbf{q}^{t-1} \setminus \{q_i^{t-1}\}$;
    根据 B 计算分布 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$;
    $q_i^t =$ 根据 $P_B(Q_i \mid \mathbf{Z} = \mathbf{z})$ 采样所获 Q_i 取值;
    $\mathbf{q}^t =$ 将 $\mathbf{q}^{t-1}$ 中的 q_i^{t-1} 用 q_i^t 替换
  end for
  if $\mathbf{q}^t = \mathbf{q}$ **then**
    $n_q = n_q + 1$
  end if
end for

输出: $P(\mathbf{Q} = \mathbf{q} \mid \mathbf{E} = \mathbf{e}) \simeq \frac{n_q}{T}$

图7.5 吉布斯采样算法

章节目录

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？

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

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \qquad (7.34)$$

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

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \qquad (7.34)$$

- 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）

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \boldsymbol{\Theta}) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \qquad (7.35)$$

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)

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，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。

EM算法

进一步，若我们不是取Z的期望，而是基于Θ^t 计算隐变量 $\mathbf{Z}$ 的概率分布 $P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$，则EM算法的两个步骤是：

- E步（Expectation）：以当前参数 Θ^t 推断隐变量分布 $LL(\Theta \mid \mathbf{X}, \mathbf{Z})$ ，并计算对数似然 $P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$ 关于 $\mathbf{Z}$ 的期望：

$$Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^t) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\Theta}^t} LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) \qquad (7.36)$$

- M步（Maximization）：寻找参数最大化期望似然，即

$$\boldsymbol{\Theta}^{t+1} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^t) \qquad (7.37)$$

- EM算法使用两个步骤交替计算：第一步计算期望（E步），利用当前估计的参数值计算对数似然的参数值；第二步最大化（M步），寻找能使E步产生的似然期望最大化的参数值……直至收敛到全局最优解。

小结

- 贝叶斯决策论
- 极大似然估计
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网
- **EM**算法