第七章 贝叶斯分类器

王博：自动化（人工智能）学院
wangbo@hdu. edu. cn

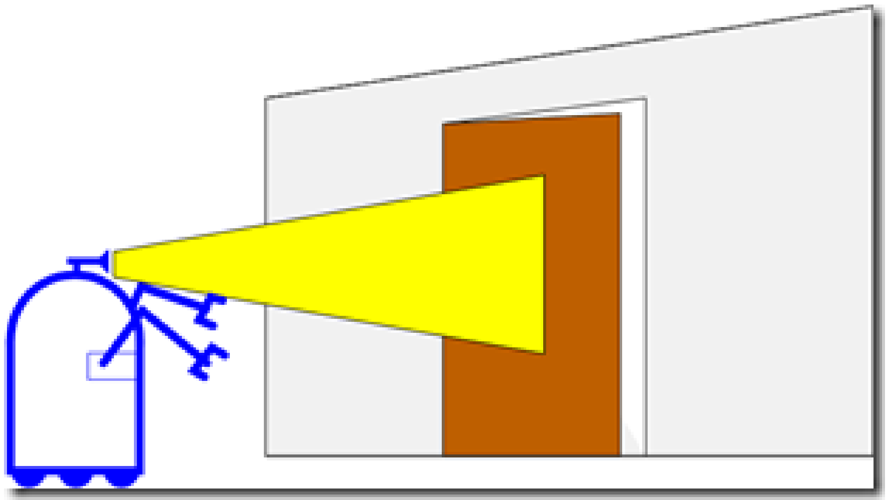
章节目录

- 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

章节目录

贝叶斯分类器

➤ 机器人根据观测的到门的距离，估算门开或关的概率，若测量到门的距离为z=0.6m，则可用条件概率描述门开着的概率。



$$P(\text{open}|z=0.6)=?$$

$$P(open|z=0.6)=\frac{P(z|open)P(open)}{P(z)} \quad <--- \text{贝叶斯公式}$$

$$=\frac{P(z|open)P(open)}{P(z|open)p(open)+P(z|\neg open)p(\neg open)} \quad <--- \text{全概率公式}$$

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）

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

- 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

- 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。

- 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

$$h^*(x) = \underset{c \in y}{\arg\min}\, R(c \mid x) \qquad (7.3)$$

  - 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

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

- 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

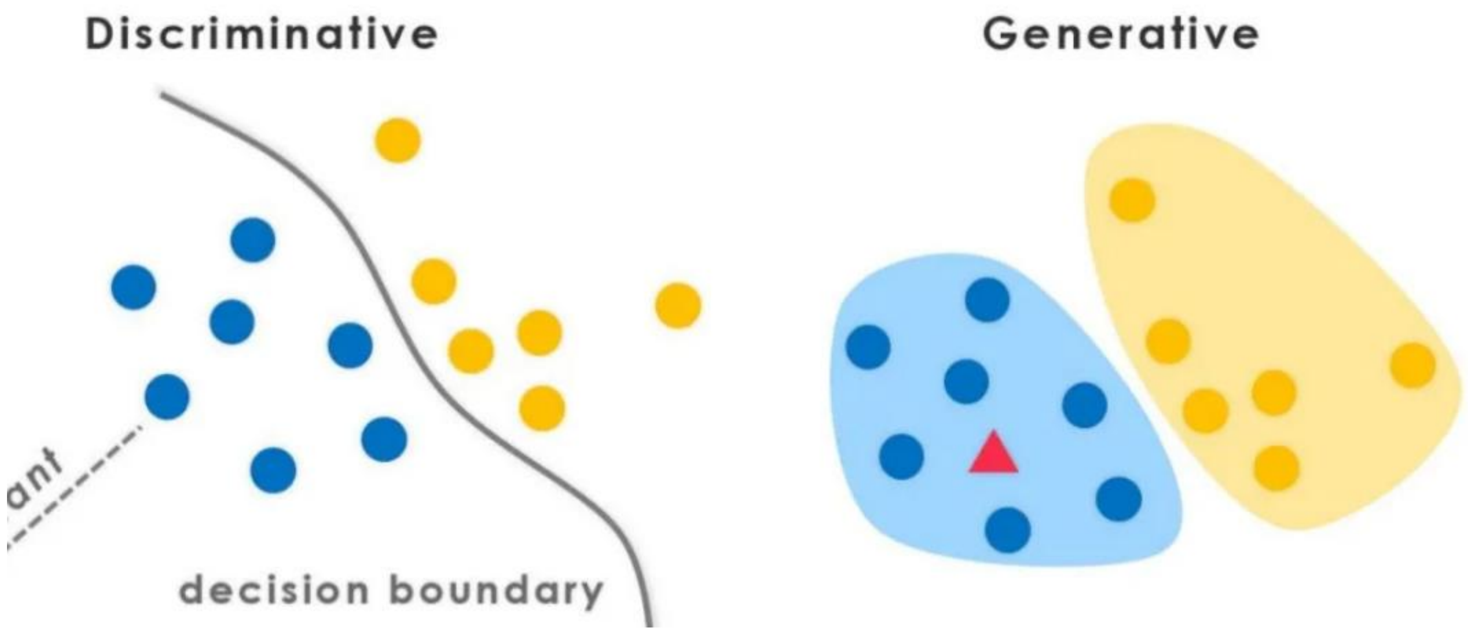
- 使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c \mid \mathbf{x})$。

- 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

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

贝叶斯决策论



判别模型之所以称为“判别”模型，是因为其根据X“判别” Y，或输入x得到y；

而生成模型之所以称为“生成”模型，是因为其预测的根据是联合概率P(X,Y)

贝叶斯决策论

- 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

章节目录

- 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

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

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

  - 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 $$\hat{\boldsymbol{\theta}}_{\boldsymbol{c}} = \underset{\boldsymbol{\theta}_c}{\arg\max}\, LL(\boldsymbol{\theta}_c) \qquad (7.11)$$

极大似然估计

- 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

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

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

- 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

- 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

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

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

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

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

- 朴素贝叶斯分类器的表达式

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 $P(x_i \mid c)$ 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

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

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

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \qquad (7.19)$$

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

- 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

- 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

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-parent），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

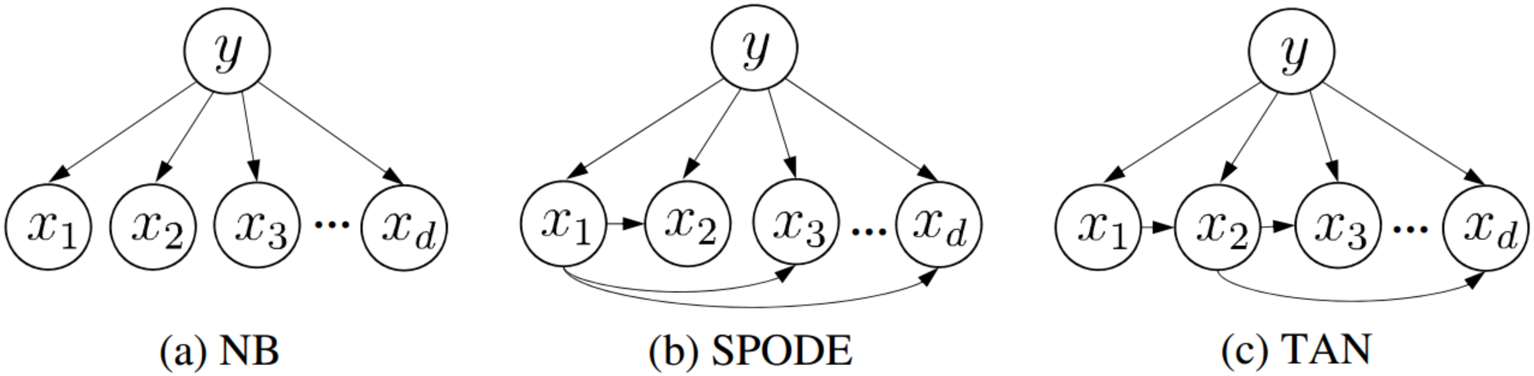


图7.1 朴素贝叶斯分类器与两种半朴素分类器所考虑的属性依赖关系

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

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。每个节点对应一个属性，属性直接有直接依赖关系则由一条边来连接。$B = \langle G, \Theta \rangle$

网络结构 G 是一个有向无环图, 其每个结点对应于一个属性, 若两个属性有直接依赖关系, 则它们由一条边连接起来; 参数 Θ 定量描述这种依赖关系, 假设属性 x_i 在 G 中的父结点集为 π_i, 则 Θ 包含了每个属性的条件概率表 $\theta_{x_i|\pi_i} = P_B(x_i \mid \pi_i)$.

贝叶斯网

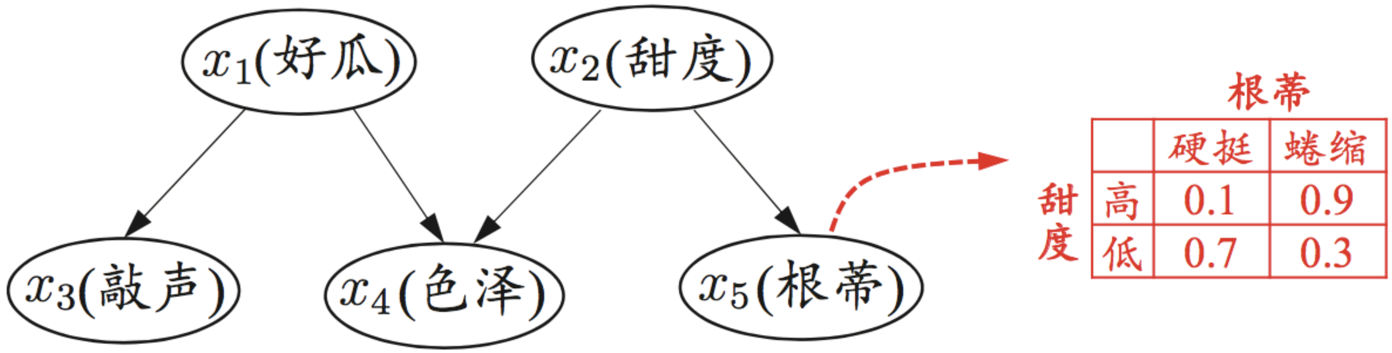


图7.2 西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

- 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(根蒂=硬挺|甜度=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

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

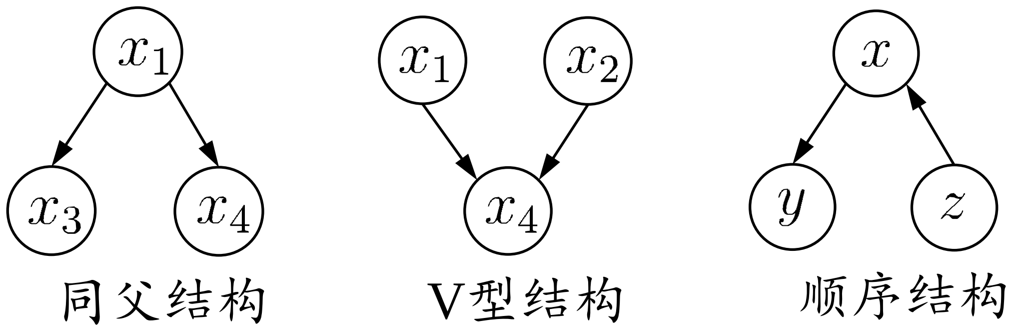
图7.2的联合概率分布定义为：

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

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：



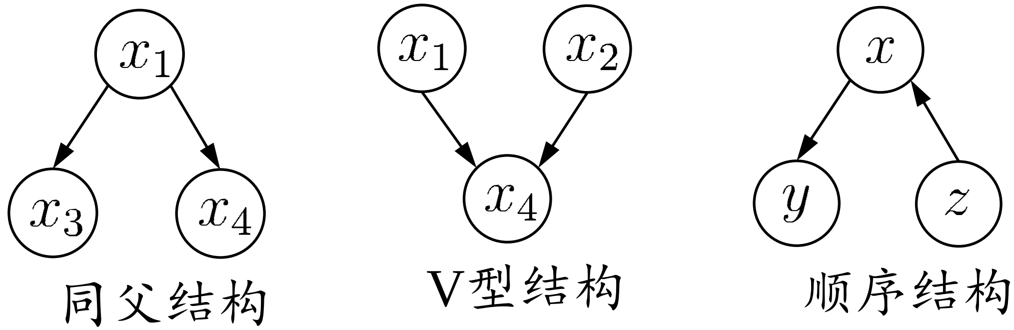
同父结构：x1给定x3与x4条件独立

V型结构：x4未知，x1和x2独立

顺序结构：x已知，y和z条件独立

贝叶斯网：结构

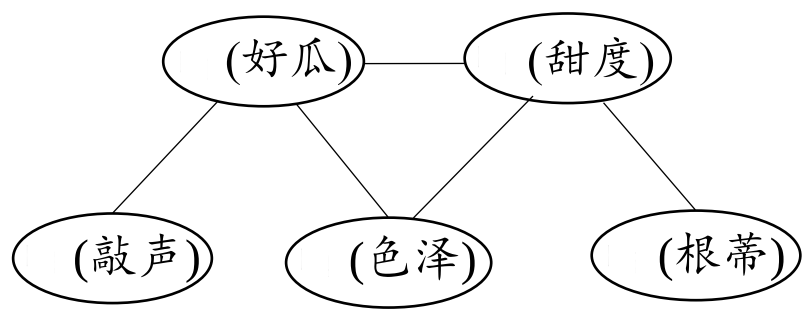
- 贝叶斯网中三个变量之间的典型依赖关系：



- 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)

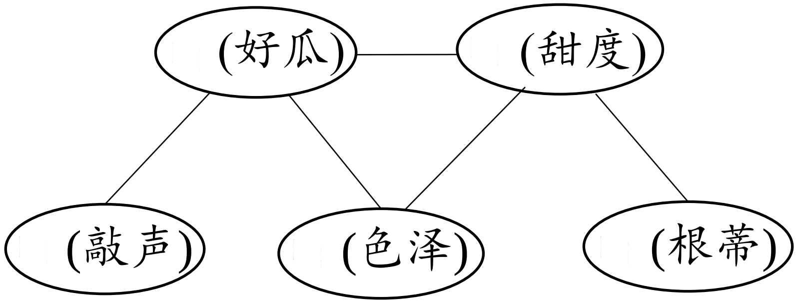


贝叶斯网：结构

- 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)



- 连通分支与点割集

所有的条件独立关系: $x_3 \perp x_4 \mid x_1$, $x_4 \perp x_5 \mid x_2$, $x_3 \perp x_2 \mid x_1$, $x_3 \perp x_5 \mid x_1$, $x_3 \perp x_5 \mid x_2$ 等.

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。
- 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。
- 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。

吉布斯采样算法

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

- 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

EM算法（机器学习方法）

例 9.1（**三硬币模型**）　假设有 3 枚硬币，分别记作 A，B，C。这些硬币正面出现的概率分别是 π，p 和 q。进行如下掷硬币试验：先掷硬币 A，根据其结果选出硬币 B 或硬币 C，正面选硬币 B，反面选硬币 C；然后掷选出的硬币，掷硬币的结果，出现正面记作 1，出现反面记作 0；独立地重复 n 次试验（这里，$n = 10$），观测结果如下：

$$1, 1, 0, 1, 0, 0, 1, 0, 1, 1$$

假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三硬币正面出现的概率，即三硬币模型的参数。

想要知道每种硬币正面向上的概率，我们要计算这个概率首先要知道每一轮用了哪一种硬币。如果我们想要推算每一次实验用了哪一种硬币又需要先知道硬币正面朝上的概率（**两个变量互相纠缠、互相依赖**）

EM算法引入

三硬币模型是一个生成模型，其中 $y \in \{0,1\}$ 是观测变量； z 是隐变量，表示硬币A的结果；$\theta=(\pi,p,q)$ 是模型参数，写作：

$$
\begin{aligned}
P(y \mid \theta) &= \sum_z P(y,z \mid \theta) = \sum_z P(z \mid \theta) P(y \mid z,\theta) \\
&= \pi p^y (1-p)^{1-y} + (1-\pi) q^y (1-q)^{1-y}
\end{aligned}
$$

则观测数据 Y 的似然函数为

$$
\begin{aligned}
P(Y \mid \theta) &= \sum_Z P(Z \mid \theta) P(Y \mid Z,\theta) \\
&= \prod_{j=1}^{n} \left[\pi p^{y_j} (1-p)^{1-y_j} + (1-\pi) q^{y_j} (1-q)^{1-y_j} \right]
\end{aligned}
$$

目标是求解模型参数 $\theta=(\pi,p,q)$ 的极大似然估计

$$\hat{\theta} = \arg\max_\theta \log P(Y \mid \theta)$$

该问题无解析解，可以通过EM算法迭代求解。

EM算法引入

1. 设置初值 $\theta^{(0)}=\left(\pi^{(0)}, p^{(0)}, q^{(0)}\right)$ ；

2. E步：计算在模型参数 $\pi^{(i)}, p^{(i)}, q^{(i)}$ 下观测数据 y_j 来自掷硬币 B 的概率

$$\mu_j^{(i+1)}=\frac{\pi^{(i)}\left(p^{(i)}\right)^{y_j}\left(1-p^{(i)}\right)^{1-y_j}}{\pi^{(i)}\left(p^{(i)}\right)^{y_j}\left(1-p^{(i)}\right)^{1-y_j}+\left(1-\pi^{(i)}\right)\left(q^{(i)}\right)^{y_j}\left(1-q^{(i)}\right)^{1-y_j}}$$

3. M步：计算模型参数的新估计值

$$\pi^{(i+1)}=\frac{1}{n} \sum_{j=1}^n \mu_j^{(i+1)}$$

$$p^{(i+1)}=\frac{\sum_{j=1}^n \mu_j^{(i+1)} y_j}{\sum_{j=1}^n \mu_j^{(i+1)}}$$

$$q^{(i+1)}=\frac{\sum_{j=1}^n\left(1-\mu_j^{(i+1)}\right) y_j}{\sum_{j=1}^n\left(1-\mu_j^{(i+1)}\right)}$$

EM算法

未观测的变量称为“隐变量”(latent variable)。令 Y 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

$$P(Y \mid \theta)=\sum_{Z} P(Z \mid \theta) P(Y \mid Z, \theta)$$

最大似然估计算法进阶版

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

EM算法导出

对于 m 个样本观察数据 $x=(x_1,x_2,...x_m)$ 中，找出样本的模型参数 θ，极大化模型分布的对数似然函数如下：

$$\theta = arg\max_{\theta}\sum_{i=1}^{m} logP(x_i|\theta)$$

如果得到的观察数据有未观察到的隐含数据 $z=(z_1,z_2,...z_m)$，此时我们的极大化模型分布的对数似然函数如下：

$$\theta = arg\max_{\theta}\sum_{i=1}^{m} logP(x_i|\theta) = arg\max_{\theta}\sum_{i=1}^{m} log\sum_{j=1}^{m} P(x_i,z_j|\theta)$$

EM算法导出

上面这个式子是没有办法直接求出 θ, 需要一些特殊的技巧, 首先对这个式子进行缩放如下

$$\begin{aligned}\sum_{i=1}^{m} log \sum_{j=1}^{m} P(x_i, z_j|\theta) &= \sum_{i=1}^{m} log \sum_{j=1}^{m} Q_i(z_j) \frac{P(x_i, z_j|\theta)}{Q_i(z_j)} \\ &\geq \sum_{i=1}^{m} \sum_{j=1}^{m} Q_i(z_j) log \frac{P(x_i, z_j|\theta)}{Q_i(z_j)}\end{aligned}$$

引入了一个未知的新的分布 $Q_i(z_j)$, 并结合 Jensen 不等式 $(f(E(x)) \geq E(f(x))$, 如果 $f(x)$ 是凹函数)

$$log \sum_j \lambda_j y_j \geq \sum_j \lambda_j log y_j, \lambda_j \geq 0, \sum_j \lambda_j = 1$$

EM算法导出

此时如果要满足 Jensen 不等式的等号, 则有: $\frac{P(x_i,z_j|\theta)}{Q_i(z_j)}=c$. 由于 $Q_i(z_j)$ 是一个分布: $\sum_{j=1}^{m}Q_i(z_j)=1$. 可以得到:

$$Q_i(z_j)=\frac{P(x_i,z_j|\theta)}{\sum_{j=1}^{m}P(x_i,z_j|\theta)}=\frac{P(x_i,z_j|\theta)}{P(x_i|\theta)}=P(z_j|x_i,\theta))$$

如果 $Q_i(z_j)=P(z_j|x_i,\theta))$, 则上面不等式包含隐藏数据的对数似然的一个下界。如果我们能极大化这个下界, 则也在尝试极大化我们的对数似然。即我们需要最大化下式:

$$arg\max_{\theta}\sum_{i=1}^{m}\sum_{j=1}^{m}Q_i(z_j)log\frac{P(x_i,z_j|\theta)}{Q_i(z_j)}$$

EM算法导出

去掉上式中为常数的部分, 则我们需要极大化的对数似然下界为:

$$\sum_{j=1}^{m} Q_i(z_j) log P(x_i, z_j|\theta)$$

上式也就是 EM 算法的 M 步, 那 E 步呢? 注意到上式中 $Q_i(z_j)$ 是一个分布, 因此 $\sum_{j=1}^{m} Q_i(z_j) log P(x_i, z_j|\theta)$ 可以理解为 $log P(x_i, z_j|\theta)$ 基于条件概率分布 $Q_i(z_j)$ 的期望

EM算法-步骤

输入：m 个样本观察数据 $x=(x_1,x_2,...x_m)$, 联合分布 $p(x,z|\theta)$, 条件分布 $p(z|x,\theta)$, 最大迭代次数 J.

1) 随机初始化模型参数 θ 的初值 θ^0。

2）for k from 1 to J 开始 EM 算法迭代：

a) E 步：计算联合分布的条件概率期望：

$$Q_i(z_i)=P(z_i|x_i,\theta^k))$$

$$L(\theta,\theta^k)=\sum_{i=1}^{m}\sum_{j=1}^{m}Q_i(z_j)logP(x_i,z_j|\theta)$$

b) M 步：极大化 $L(\theta,\theta^k)$, 得到 θ^{k+1}:

$$\theta^{k+1}=arg\max_{\theta}L(\theta,\theta^k)$$

c) 如果 θ^{k+1} 已收敛, 则算法结束。否则继续回到步骤 a) 进行 E 步迭代

输出：模型参数 θ。

EM算法-举例

我们目前有100个男生和100个女生的身高，但是我们不知道这200个数据中哪个是男生的身高，哪个是女生的身高，即抽取得到的每个样本都不知道是从哪个分布中抽取的。这个时候，对于每个样本，就有两个未知量需要估计：

（1）这个身高数据是来自于男生数据集合还是来自于女生?

（2）男生、女生身高数据集的正态分布的参数分别是多少?