第七章：贝叶斯分类器

主讲：连德富 特任教授 | 博士生导师

邮箱：liandefu@ustc.edu.cn

手机：13739227137

主页：http://staff.ustc.edu.cn/~liandefu

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$\mathcal{Y} = \{c_1, \cdots, c_N\}$，$\lambda_{ij}$是将一个真实标记为$c_i$的样本误分类为$c_j$所产生的损失。基于后验概率$P(c_i|\boldsymbol{x})$可获得将样本$\boldsymbol{x}$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）

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

- 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

$$R(h) = \mathbb{E}_x[R(h(\boldsymbol{x})|\boldsymbol{x})]$$

贝叶斯决策论

- 寻找一个判定准则 $h: X \mapsto Y$以最小化总体风险

$$R(h) = \mathbb{E}_x[R(h(\boldsymbol{x})|\boldsymbol{x})] \qquad R(c_i|\boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\boldsymbol{x})$$

- 显然，对每个样本$\boldsymbol{x}$，若$\boldsymbol{x}$能最小化条件风险$R(h(\boldsymbol{x})|\boldsymbol{x})$，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 $R(c|\boldsymbol{x})$最小的类别标记，即

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

  - 被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险$R(h^\star)$称为贝叶斯风险 (Bayes risk)
  - 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

- 若目标是最小化分类错误率，则误判损失λ_{ij}可写为

$$\lambda_{ij} = \begin{cases} 0, \text{if } i = j \\ 1, \text{otherwise} \end{cases}$$

- 此时条件风险

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

- 最小化分类错误率的贝叶斯最有分类器为

$$h^{\star}(\boldsymbol{x}) = \arg\min_{c \in \mathcal{Y}} R(c|\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c|\boldsymbol{x})$$

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

贝叶斯决策论

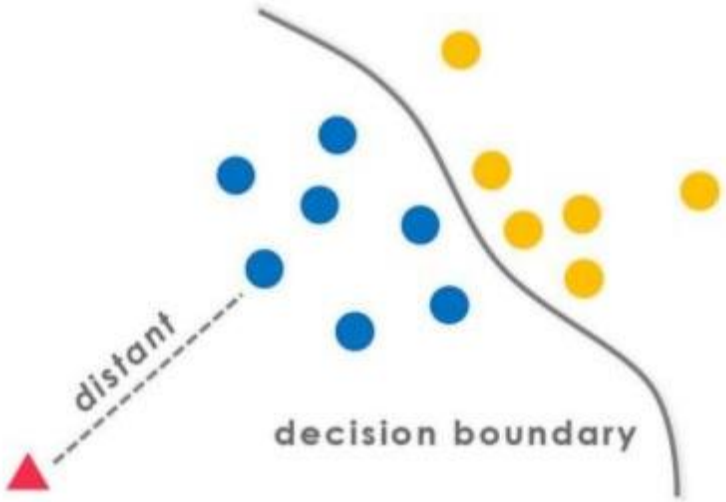
- 最小化分类错误率的贝叶斯最有分类器为

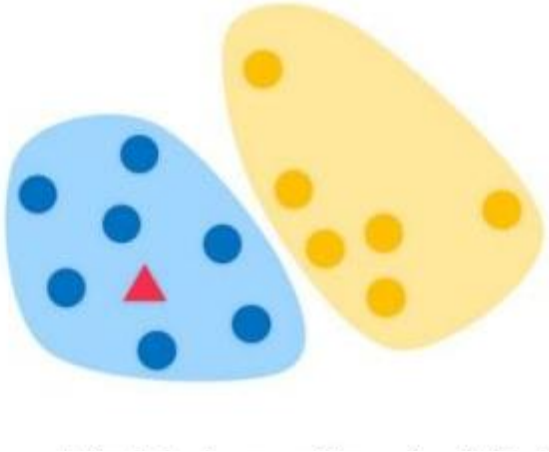
$$h^{\star}(\boldsymbol{x}) = \arg\min_{c\in\mathcal{Y}} R(c|\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c|\boldsymbol{x})$$

- 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$P(c|\boldsymbol{x})$
- 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 $P(c|\boldsymbol{x})$。

贝叶斯决策论

有监督学习目标：学习决策函数 $c = f(X)$ 或者条件概率 $P(c|X)$





- 判别模型
  - 直接学习$f(\boldsymbol{x})$或者$P(c|\boldsymbol{x})$
  - 直接面对预测，准确率较高，学习简单
  - 典型模型包括对率回归、SVM等

- 生成模型
  - 建模联合概率分布$P(\boldsymbol{x}, c)$，求出条件概率进行预测

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

  - 典型模型包括朴素贝叶斯

贝叶斯决策论

- 生成式模型

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

- 基于贝叶斯定理，$P(c|\boldsymbol{x})$可写成

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

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）

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

“证据”（evidence）因子，与类标记无关

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$P(\boldsymbol{x}|c)$，
  - 假设$P(\boldsymbol{x}|c)$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

$$P(D_c|\boldsymbol{\theta}_c) = \prod_{\boldsymbol{x}\in D_c} P(\boldsymbol{x}|\boldsymbol{\theta}_c)$$

- 对 $\boldsymbol{\theta}_c$进行极大似然估计，寻找能最大化似然$P(D_c|\boldsymbol{\theta}_c)$的参数值$\hat{\boldsymbol{\theta}}_c$。直观上看，极大似然估计是试图在$\boldsymbol{\theta}_c$所有可能的取值中，找到一个使数据出现的“可能性”最大值。
- 连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$LL(\theta_c) = \log P(D_c|\theta_c) = \sum_{x\in D_c} \log P(\boldsymbol{x}_c|\theta_c)$$

- 此时参数 $\boldsymbol{\theta}_c$的极大似然估计$\hat{\boldsymbol{\theta}}_c$为

$$\hat{\boldsymbol{\theta}}_c = \arg\max_{\boldsymbol{\theta}_c} LL(\boldsymbol{\theta}_c)$$

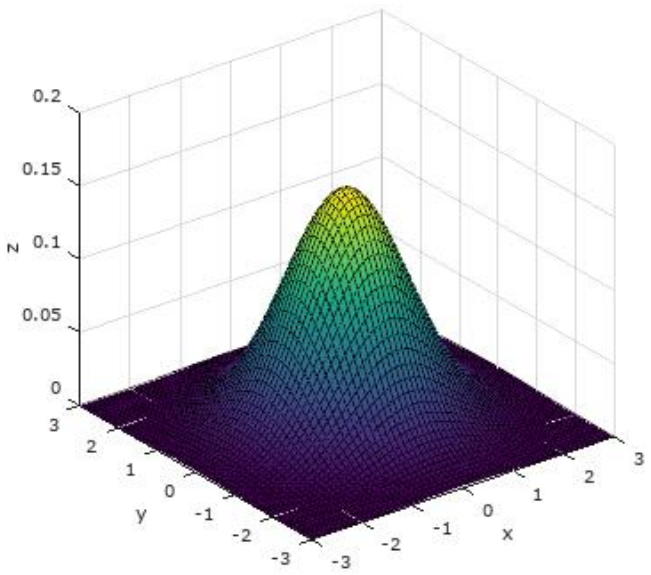
极大似然估计

- 在连续属性情形下，假设概率密度函数$P(\boldsymbol{x}|c)\sim\mathcal{N}(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$，

则参数$\boldsymbol{\mu}_c$和$\boldsymbol{\Sigma}_c$的极大似然估计为

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

$$\hat{\boldsymbol{\Sigma}}_c=\frac{1}{|D_c|}\sum_{\boldsymbol{x}\in D_c}(\boldsymbol{x}-\boldsymbol{\mu}_c)(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}$$



$$P(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)=\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_c|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

极大似然估计

$$P(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)=\frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_c|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

$$LL(\boldsymbol{\theta}_c)=\sum_{\boldsymbol{x}\in D_c}\log P(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$

$$=-\sum_{\boldsymbol{x}\in D_c}\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)-\frac{|D_c|}{2}\log|\boldsymbol{\Sigma}_c|-\frac{d|D_c|}{2}\log(2\pi)$$

$$\frac{\partial LL(\boldsymbol{\theta}_c)}{\partial\boldsymbol{\mu}_c}=\sum_{\boldsymbol{x}\in D_c}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)=0 \quad\Rightarrow\quad \hat{\boldsymbol{\mu}}_c=\frac{1}{|D_c|}\sum_{\boldsymbol{x}\in D_c}\boldsymbol{x}$$

$$\frac{\partial LL(\boldsymbol{\theta}_c)}{\partial\boldsymbol{\Sigma}_c}=-\frac{\partial}{\partial\boldsymbol{\Sigma}_c}\sum_{\boldsymbol{x}\in D_c}\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)-\frac{|D_c|}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}_c}\log|\boldsymbol{\Sigma}_c|$$

极大似然估计

$$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{|D_c|}\sum_{\boldsymbol{x}\in D_c}(\boldsymbol{x}-\boldsymbol{\mu}_c)(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}$$

$$\frac{\partial LL(\boldsymbol{\theta}_c)}{\partial \boldsymbol{\Sigma}_c} = -\frac{\partial}{\partial \boldsymbol{\Sigma}_c}\sum_{\boldsymbol{x}\in D_c}\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathsf{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) - \frac{|D_c|}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_c}\log|\boldsymbol{\Sigma}_c|$$

$$= -\frac{|D_c|}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_c}\mathrm{tr}\left(\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\right) - \frac{|D_c|}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_c}\log|\boldsymbol{\Sigma}_c| = \frac{|D_c|}{2}\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}_c^{-1} - \frac{|D_c|}{2}\boldsymbol{\Sigma}_c^{-1}$$

$$\left[\frac{\partial}{\partial \boldsymbol{\Sigma}_c}\mathrm{tr}\left(\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\right)\right]_{ij} = \frac{\partial}{\partial \Sigma_{ij}^c}\mathrm{tr}\left(\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\right)$$

$$= \mathrm{tr}\left(\frac{\partial}{\partial \Sigma_{ij}^c}\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\right)$$

$$= -\mathrm{tr}\left(\boldsymbol{\Sigma}_c^{-1}\frac{\partial \boldsymbol{\Sigma}_c}{\partial \Sigma_{ij}^c}\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\right)$$

$$= -\mathrm{tr}\left(\frac{\partial \boldsymbol{\Sigma}_c}{\partial \Sigma_{ij}^c}\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}_c^{-1}\right)$$

$$= -\left[\boldsymbol{\Sigma}_c^{-1}\hat{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}_c^{-1}\right]_{ij}$$

$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$

$$\frac{\partial \boldsymbol{A}^{-1}\boldsymbol{A}}{\partial x} = \frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{A} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x} = \boldsymbol{0}$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_c}\log|\boldsymbol{\Sigma}_c| = (\boldsymbol{\Sigma}_c^{-1})^{\mathsf{T}} = \boldsymbol{\Sigma}_c^{-1}$$

$$\frac{\partial LL(\boldsymbol{\theta}_c)}{\partial \boldsymbol{\Sigma}_c} = 0 \Rightarrow \hat{\boldsymbol{\Sigma}}_c\boldsymbol{\Sigma}_c^{-1} = I$$

$$\boldsymbol{\Sigma}_c = \hat{\boldsymbol{\Sigma}}_c$$

极大似然估计

- 在连续属性情形下，假设概率密度函数$P(\boldsymbol{x}|c)\sim\mathcal{N}(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$，则参数$\boldsymbol{\mu}_c$和$\boldsymbol{\Sigma}_c$的极大似然估计为

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

$$\hat{\boldsymbol{\Sigma}}_c=\frac{1}{|D_c|}\sum_{\boldsymbol{x}\in D_c}(\boldsymbol{x}-\boldsymbol{\mu}_c)(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}$$

- 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$(\boldsymbol{x}-\boldsymbol{\mu}_c)(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}$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。

朴素贝叶斯分类器

- 估计后验概率$P(c|\boldsymbol{x})$主要困难：类条件概率$P(\boldsymbol{x}|c)$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，

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

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

朴素贝叶斯分类器

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

由于对所有类别来说 $P(\boldsymbol{x})$相同，因此基于贝叶斯判定准则有

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

这就是朴素贝叶斯分类的表达式子

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$P(c)$

令 D_c表示训练集D中第c类样本组成的集合，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

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

  - 为每个属性估计条件概率$P(x_i|c)$

对离散属性，令D_{c,x_i}表示D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 $P(x_i|c)$可估计为

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

对连续属性，设$p(x_i|c) = \mathcal{N}(x_i|\mu_{c,i}, \sigma_{c,i}^2)$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

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

朴素贝叶斯分类器

- 用西瓜数据集3.0训练朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

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	?

朴素贝叶斯分类器

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	?

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

$$P(\text{密度} = 0.697|\text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi} \times 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \times 0.129^2}\right) \approx 1.959$$

$$P(\text{密度} = 0.697|\text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi} \times 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \times 0.195^2}\right) \approx 1.203$$

- 估计类先验概率$P(c)$,

$$P(\text{好瓜} = \text{是}) = \frac{8}{17}$$

- 为每个属性估计条件概率

$P(\text{色泽} = \text{青绿}|\text{好瓜} = \text{是}) = 3/8$

$P(\text{色泽} = \text{青绿}|\text{好瓜} = \text{否}) = 3/9$

$P(\text{根蒂} = \text{蜷缩}|\text{好瓜} = \text{是}) = 5/8$

$P(\text{根蒂} = \text{蜷缩}|\text{好瓜} = \text{否}) = 3/9$

$P(\text{敲声} = \text{浊响}|\text{好瓜} = \text{是}) = 6/8$

$P(\text{敲声} = \text{浊响}|\text{好瓜} = \text{否}) = 4/9$

$P(\text{纹理} = \text{清晰}|\text{好瓜} = \text{是}) = 7/8$

$P(\text{纹理} = \text{清晰}|\text{好瓜} = \text{否}) = 2/9$

$P(\text{脐部} = \text{凹陷}|\text{好瓜} = \text{是}) = 5/8$

$P(\text{脐部} = \text{凹陷}|\text{好瓜} = \text{否}) = 2/9$

朴素贝叶斯分类器

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	?

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

$$P(好瓜=是|\boldsymbol{x}) \propto P(好瓜=是) \times P_{青绿|是} \times P_{蜷缩|是} \times P_{浊响|是} \times P_{清晰|是} \times P_{凹陷|是} \times P_{硬滑|是} \approx 0.052$$

$$P(好瓜=否|\boldsymbol{x}) \propto P(好瓜=否) \times P_{青绿|否} \times P_{蜷缩|否} \times P_{浊响|否} \times P_{清晰|否} \times P_{凹陷|否} \times P_{硬滑|否} \approx 6.80 \times 10^{-5}$$

$$P(密度=0.697|好瓜=是) = \frac{1}{\sqrt{2\pi}\times 0.129} \exp\left(-\frac{(0.697-0.574)^2}{2\times 0.129^2}\right) \approx 1.959$$

测试例“测1”判别为好瓜

$$P(密度=0.697|好瓜=是) = \frac{1}{\sqrt{2\pi}\times 0.195} \exp\left(-\frac{(0.697-0.496)^2}{2\times 0.195^2}\right) \approx 1.203$$

- 估计类先验概率$P(c)$,

$$P(好瓜=是) = \frac{8}{17}$$

属性估计条件概率

$P(青绿|好瓜=是) = 3/8$

$P(青绿|好瓜=否) = 3/9$

$P(蜷缩|好瓜=是) = 5/8$

$P(蜷缩|好瓜=否) = 3/9$

$P(浊响|好瓜=是) = 6/8$

$P(敲声=浊响|好瓜=否) = 4/9$

$P(纹理=清晰|好瓜=是) = 7/8$

$P(清晰|好瓜=否) = 2/9$

$P(脐部=凹陷|好瓜=是) = 5/8$

$P(脐部=凹陷|好瓜=否) = 2/9$

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”
  - 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

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

- 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。

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

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

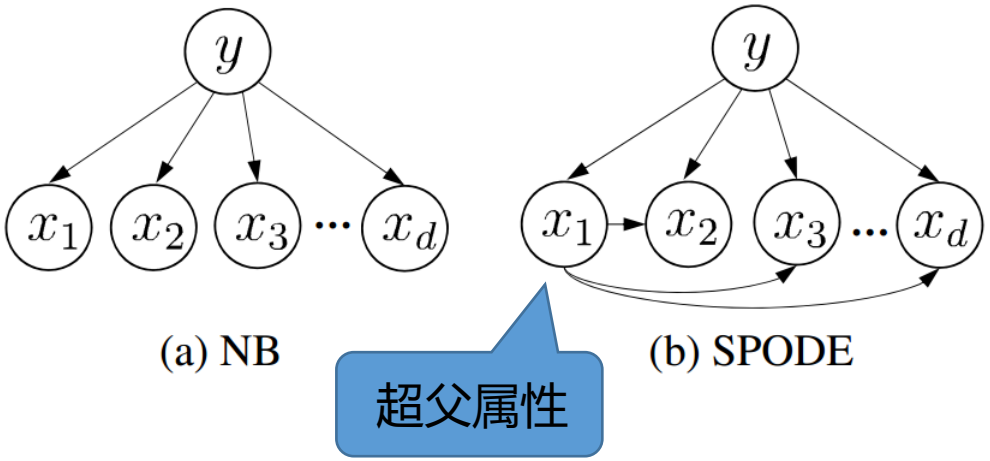
  - 其中pa_i为属性x_i所依赖的属性，称为x_i的父属性

- 对每个属性x_i，若其父属性pa_i已知，则可估计概值$P(x_i|c,pa_i)$，

问题的关键转化为如何确定每个属性的父属性

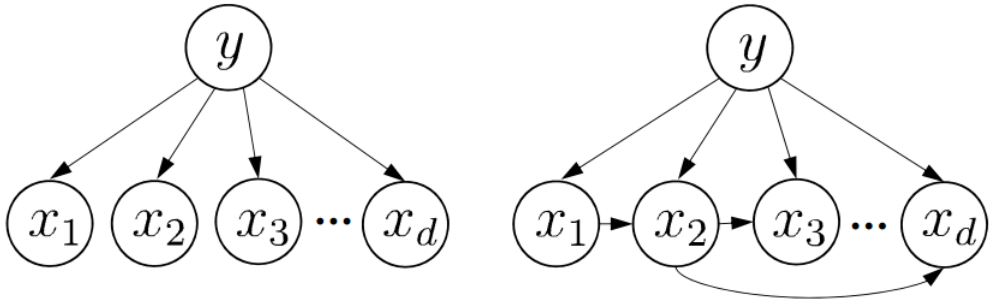
半朴素贝叶斯分类器

- 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。



半朴素贝叶斯分类器

- 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



(a) NB　　(c) TAN

  - 计算任意两个属性之间的条件互信息

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

  - 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互信息
  - 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。

半朴素贝叶斯分类器

- 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

$$P(c|\boldsymbol{x}) \propto \sum_{i;\left|D_{x_i}\right|\geq m'}^{d} P(c,x_i)\prod_{j}^{d} P\left(x_j|c,x_i\right)$$

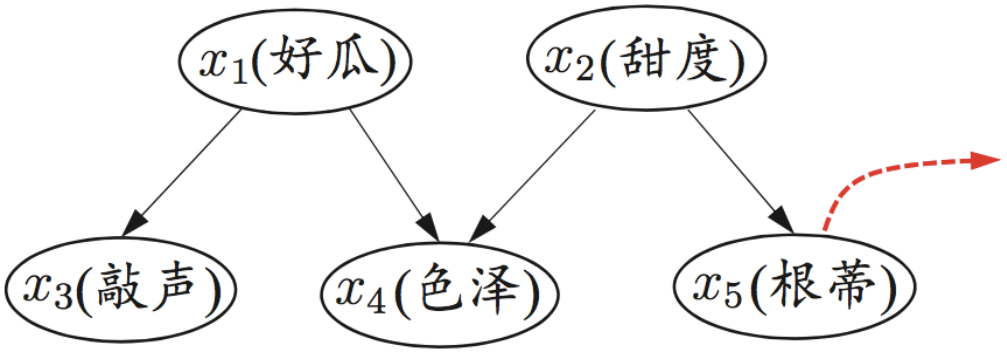
其中，D_{x_i}是在第i个属性上取值x_i的样本的集合，m'为阈值常数

$$P(x_i,c)=\frac{\left|D_{c,x_i}\right|+1}{|D|+N\times N_i} \qquad P\left(x_j|c,x_i\right)=\frac{\left|D_{c,x_i,x_j}\right|+1}{\left|D_{c,x_i}\right|+N_j}$$

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，D_{c,x_i,x_j}是类别为 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

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

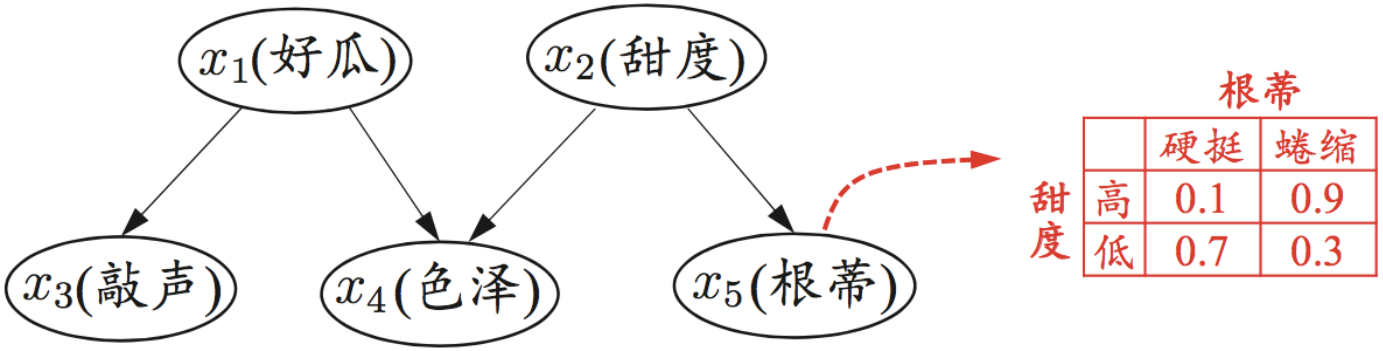
西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

- 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

- 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

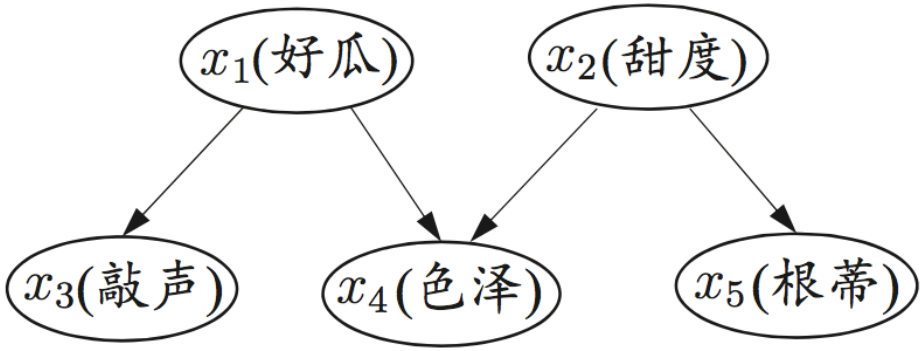
$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$



$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)$$

贝叶斯网

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)$$



x_4和x_5在给定x_2的取值时独立

$$x_5 \perp x_4 | x_2$$

$$P(x_3, x_4|x_1) = \frac{\sum_{x_2,x_5} P(x_1, x_2, x_3, x_4, x_5)}{\sum_{x_2,x_3,x_4,x_5} P(x_1, x_2, x_3, x_4, x_5)}$$

$$= \frac{\sum_{x_2,x_5} P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)}{\sum_{x_2,x_3,x_4,x_5} P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)}$$

$$= \frac{\sum_{x_2} P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)}{P(x_1)}$$

$$P(x_4|x_1, x_2)P(x_2) = P(x_4, x_2|x_1)$$

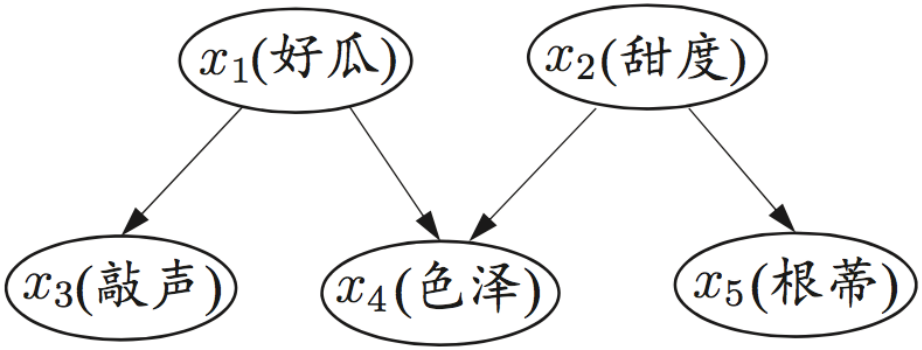
$$= \frac{P(x_1)P(x_3|x_1)P(x_4|x_1)}{P(x_1)} = P(x_3|x_1)P(x_4|x_1)$$

x_3和x_4在给定x_1的取值时独立

$$x_3 \perp x_4 | x_1$$

贝叶斯网

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)$$



$$P(x_1, x_2) = \sum_{x_3, x_4, x_5} P(x_1, x_2, x_3, x_4, x_5)$$

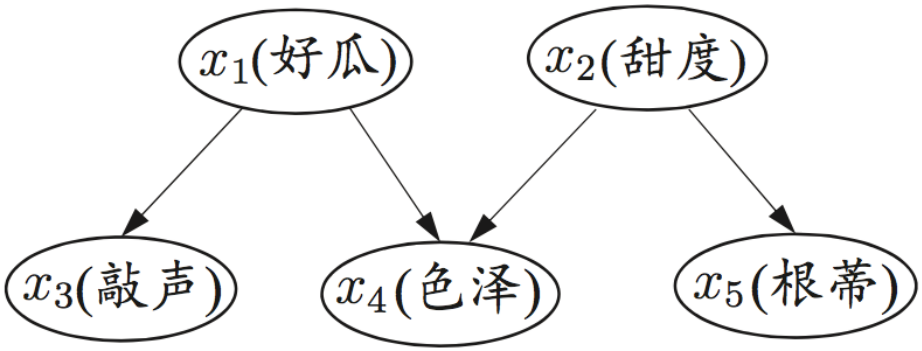
$$= \sum_{x_3, x_4, x_5} P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)$$

$$= P(x_1)P(x_2)$$

x_1和x_2相互独立　　$x_1 \perp x_2$　　称为边际独立性

贝叶斯网

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)$$



$$P(x_1, x_2|x_4) = \frac{\sum_{x_3,x_5} P(x_1, x_2, x_3, x_4, x_5)}{\sum_{x_1,x_2,x_3,x_5} P(x_1, x_2, x_3, x_4, x_5)}$$

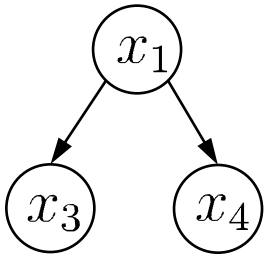
$$= \frac{\sum_{x_3,x_5} P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)}{\sum_{x_1,x_2,x_3,x_5} P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2)}$$

$$= \frac{P(x_1)P(x_2)P(x_4|x_1, x_2)}{\sum_{x_1,x_2} P(x_1)P(x_2)P(x_4|x_1, x_2)} = \frac{P(x_1, x_2, x_4)}{P(x_4)}$$

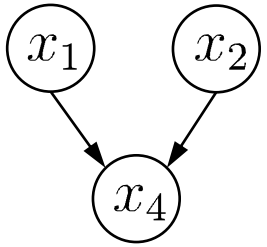
x_1和x_2在给定x_4的取值时不独立　　$x_1 \not\perp x_2|x_4$

贝叶斯网

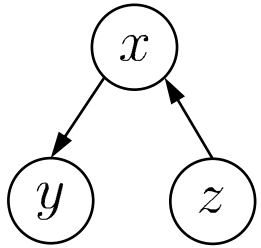
- 贝叶斯网中三个变量之间的典型依赖关系：



同父结构



V型结构



顺序结构

$$P(y,z|x) = \frac{P(x,y,z)}{P(x)}$$

$$= \frac{P(z)P(x|z)P(y|x)}{P(x)}$$

$$= \frac{P(x)P(z|x)P(y|x)}{P(x)}$$

$$= P(z|x)P(y|x)$$

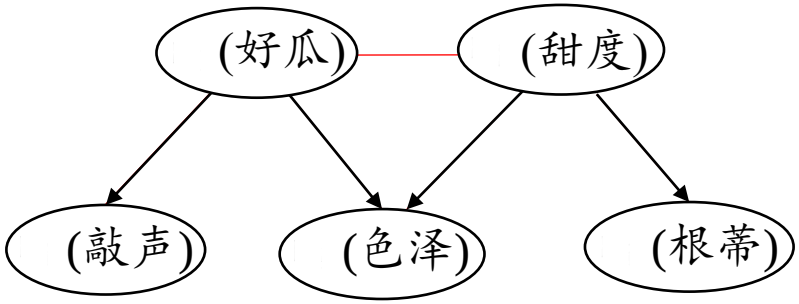
y和z在给定x的取值时独立

$$y \perp z|x$$

贝叶斯网

- 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”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$D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$，贝叶斯网$B = \langle G, \Theta \rangle$在$D$上的评价函数可以写为

$$s(B|D) = f(\theta)|B| - LL(B|D)$$

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

$$LL(B|D) = \sum_{i}^{m} \log P_B(\boldsymbol{x}_i)$$

是贝叶斯网的对数似然

贝叶斯网：学习

$$s(B|D) = f(\theta)|B| - LL(B|D)$$

- 若$f(\theta) = 1$，即每个参数用1位编码描述，则得到AIC (Akaike Information Criterion)

$$AIC(B|D) = |B| - LL(B|D)$$

- 若$f(\theta) = \frac{1}{2}\log m$，则得到BIC (Bayesian Information Criterion)

$$BIC(B|D) = \frac{\log m}{2}|B| - LL(B|D)$$

若贝叶斯网$B = \langle G, \Theta \rangle$网络结构$G$给定，$s(B|D)$最小化等价于参数$\Theta$的极大似然估计

参数$\theta_{x_i|\pi_i}$能直接在训练数据D上通过经验估计　　$\theta_{x_i|\pi_i} = \hat{P}_D(x_i|\pi_i)$

因此为了最小化评分函数$s(B|D)$，只需要对网络结构进行搜索

该问题是一个NP难问题

1. 贪心法：从某个网络结构出发，每次调整一条边（增加、删除、调整方向），直到评分函数不再降低为止

2. 通过给网络结构施加约束来消减搜索空间，例如将网络结构限定为树形结构

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。

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

$P(Q = q|E = e)$

{好瓜 = 是, 甜度 = 高}

{色泽 = 青绿，敲声 = 浊响，根蒂 = 蜷缩}

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

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

贝叶斯网：推断

- 吉布斯采样是在所有变量的联合状态空间与证据$E = e$一致的子空间中的随机游走，即每一步仅依赖于前一步的状态，这是一个“马尔可夫链”(Markov Chain)

- 在一定的条件下，无论从什么初始状态开始，马尔科夫链第t步的状态分布在$t \to \infty$时比收敛于一个平稳分布，即$P(Q|E = e)$

- 当T很大时，吉布斯采样相当于根据$P(Q|E = e)$采样

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，如何对模型参数进行估计?

- 未观测的变量称为“隐变量”(latent variable)。令$\boldsymbol{X}$表示已观测变量集，$\boldsymbol{Z}$表示隐变量集，若预对模型参数$\boldsymbol{\Theta}$做极大似然估计，则应最大化对数似然函数

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

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

$$LL(\boldsymbol{\Theta}|\boldsymbol{X}) = \ln P(\boldsymbol{X}|\boldsymbol{\Theta}) = \ln \sum_{Z} P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\Theta})$$

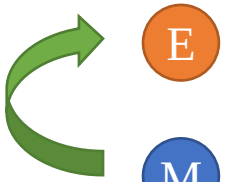
EM算法

- 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Z的值(E步)

当Z已知　　对Θ做极大似然估计(M步)

以初始值Θ^0为起点，可迭代执行以下步骤直至收敛：



E　基于Θ^t推断隐变量Z的期望，记为 Z^t;

M　基于已观测到变量X和Z^t对参数Θ做极大似然估计，记Θ^{t+1};

EM算法

若不是取$\boldsymbol{Z}$的期望，而是基于$\boldsymbol{\Theta}^t$计算隐变量Z的概率分布$P(\boldsymbol{Z}\,|\,\boldsymbol{X},\boldsymbol{\Theta}^t)$，则EM算法的两个步骤

以初始值$\boldsymbol{\Theta}^0$为起点，可迭代执行以下步骤直至收敛：

E 基于$\boldsymbol{\Theta}^t$推断隐变量$\boldsymbol{Z}$的分布$P(\boldsymbol{Z}\,|\,\boldsymbol{X},\boldsymbol{\Theta}^t)$，并计算对数似然$LL(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z})$关于$\boldsymbol{Z}$的期望；

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^t)=\mathbb{E}_{\boldsymbol{Z}\sim P(\boldsymbol{Z}\,|\,\boldsymbol{X},\boldsymbol{\Theta}^t)}LL(\boldsymbol{\Theta}|\boldsymbol{X},\boldsymbol{Z})$$

M 寻找参数最大化期望似然；

$$\boldsymbol{\Theta}^{t+1}=\arg\max_{\boldsymbol{\Theta}}Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^t)$$

EM算法

最大化已观测数据的对数“边际似然”

$$\arg\max_{\Theta} LL(\boldsymbol{\Theta}|\boldsymbol{X}) = \ln P(\boldsymbol{X}|\boldsymbol{\Theta}) = \ln \sum_{Z} P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\Theta})$$

$$\ln P(\boldsymbol{X}|\boldsymbol{\Theta}) = \ln \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \frac{P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\Theta})}{Q(\boldsymbol{Z})}$$

$$\geq \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \ln \frac{P(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\Theta})}{Q(\boldsymbol{Z})} \qquad = \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \ln \frac{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta})P(\boldsymbol{X}|\boldsymbol{\Theta})}{Q(\boldsymbol{Z})}$$

$$\mathcal{L}(Q,\boldsymbol{\Theta}) \qquad = \ln P(\boldsymbol{X}|\boldsymbol{\Theta}) + \sum_{Z} Q(\boldsymbol{Z}) \ln \frac{P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta})}{Q(\boldsymbol{Z})}$$

$$= \ln P(\boldsymbol{X}|\boldsymbol{\Theta}) - KL(Q||P)$$

$$\ln P(\boldsymbol{X}|\boldsymbol{\Theta}) = \mathcal{L}(Q,\boldsymbol{\Theta}) + KL(Q||P)$$

EM算法

$$\max_{\boldsymbol{\Theta}} \ln P(\boldsymbol{X}|\boldsymbol{\Theta}) = \max_{Q,\boldsymbol{\Theta}} \left(\mathcal{L}(Q, \boldsymbol{\Theta}) + KL(Q||P) \right)$$

因为$KL(Q||P) \geq 0$, 所以$\mathcal{L}(Q, \boldsymbol{\Theta}) \leq \ln P(\boldsymbol{X}|\boldsymbol{\Theta})$

$$\mathcal{L}(Q, \boldsymbol{\Theta}) = \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \ln \frac{P(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta})}{Q(\boldsymbol{Z})}$$

当$KL(Q||P) = 0$，即$Q(\boldsymbol{Z}) = P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta})$时，$\mathcal{L}(Q, \boldsymbol{\Theta})$取得最大值 $= \ln P(\boldsymbol{X}|\boldsymbol{\Theta})$

E 基于$\boldsymbol{\Theta}^t$推断隐变量$\boldsymbol{Z}$的分布$P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t)$，并计算对数似然$LL(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{Z})$关于$\boldsymbol{Z}$的期望;

$$\mathcal{L}(P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}), \boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{Z} \sim P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t)} LL(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^t)$$

M 寻找参数最大化期望似然;

$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^t)$$

作业

- 习题7.4
- 习题7.5
- 证明EM算法的收敛性