统计机器学习
Statistical Machine Learning

魏莱

上海海事大学信息工程学院

2019 年 4 月 17 日

第七章：贝叶斯分类

我们曾经谈过线性模型 $f(\mathbf{x}) = \mathbf{w}^t\mathbf{x}$ 的概率解释，理解从概率角度可以将预测误差——即真实值 y 与预测值 $f(\mathbf{x})$ 之间的差值，看成是满足正态分布的随机变量 ε，通过最大化似然函数的方法能够得到线性回归模型。进一步，当我们假设线性模型中的参数 $\mathbf{w}$ 也是满足某种分布的随机变量（Gauss 分布或 Laplace 分布），则我们可以得到正则线性回归模型。

实际上是否将模型参数 $\mathbf{w}$ 看做随机变量这是概率论中两个学派——频率学派和贝叶斯学派的最重要的区别。贝叶斯学派认为模型中所有的量都是随机变量，都有其隐含满足的分布，而频率学派则认为参数不论已知或者未知都是某种确定的值。

在机器学习发展的过程中，两个学派都起到了非常重要的推动作用，但随着数据量、数据维数的增加，贝叶斯学派起到越来越重要的作用。

1. 贝叶斯决策论

贝叶斯决策论 (Bayesian decision theory) 是概率框架下实施决策的基本方法。对分类任务来说，**在所有相关概率都已知**的理想情形下，贝叶斯决策论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**。

1. 贝叶斯决策论

假设数据集 $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$，包含 K 类别的样本，即 $y_i \in \{c_1, c_2, \cdots, c_K\}$. 按照上一章提到的，我们可以定义损失函数，当样本被错误分类时就产生一定的损失。假设一个样本 $\mathbf{x}$ 被分类成 c_j 类，定义其损失为 λ_j。则对于该样本，其可能的**期望损失（expected loss）**为

$$R(\mathbf{x}) = \sum_{j=1}^{K} \lambda_j P(c_j|\mathbf{x}), \tag{1}$$

其中 $P(c_j|\mathbf{x})$ 为样本 $\mathbf{x}$ 被分为 c_j 类的概率。按照贝叶斯理论，类别标签有其**先验分布（prior distribution）**，即具有先验概率 $P(c)$。这里 $P(c|\mathbf{x})$ 可以解释为：在给定样本 $\mathbf{x}$ 后的 c 的概率，因此称为**后验概率（posteriori distribution）**。

其中 $R(c_i|\mathbf{x})$ 称为样本 $\mathbf{x}$ 的**条件风险（conditional risk）**。于是，对于所有样本，一个最佳的预测模型 f 应该使得所有样本的条件风险之和（或者期望条件风险）最小：

$$\begin{aligned}
&\min \sum_{i=1}^{n} \sum_{j=1}^{K} \lambda_j P(c_j|\mathbf{x}_i) = \min \sum_{i=1}^{n} R(\mathbf{x}_i) \\
&\Leftrightarrow \min \sum_{i=1}^{n} R(\mathbf{x}_i) \frac{1}{n} \\
&\Leftrightarrow \min \sum_{i=1}^{n} R(\mathbf{x}_i) P(\mathbf{x}_i) \\
&= \min \mathbb{E}[R(\mathbf{x})|\mathbf{x}]
\end{aligned} \tag{2}$$

这里，$\mathbb{E}[R(\mathbf{x})|\mathbf{x}]$ 称为**总体风险**。而一旦数据集给出，那么 $P(\mathbf{x})$ 确定，由此可见，最小化 $\mathbb{E}[R(\mathbf{x})|\mathbf{x}]$，实际上就是最小化每一个 $R(\mathbf{x}_i)$ 的经验风险。

由于 $R(\mathbf{x}_i) = \sum_{j=1}^{K} \lambda_j P(c_j|\mathbf{x}_i)$，其中 λ_j 是被分成 c_j 类的损失，可以定义损失函数 $\lambda_j = g(f(\mathbf{x}_i), c_j)$。于是最小化总体风险可以具有如下形式：

$$\begin{aligned}
&\min \mathbb{E}[R(\mathbf{x})|\mathbf{x}] \\
&= \min \sum_{i=1}^{n} R(\mathbf{x}_i) P(\mathbf{x}_i) \\
&= \min \sum_{i=1}^{n} P(\mathbf{x}_i) \sum_{j=1}^{K} \lambda_j P(c_j|\mathbf{x}_i) \\
&= \min \sum_{i=1}^{n} P(\mathbf{x}_i) \sum_{j=1}^{K} g(f(\mathbf{x}_i), c_j) P(c_j|\mathbf{x}_i)
\end{aligned} \tag{3}$$

假设后验概率 $P(c_j|\mathbf{x}_i)$ 已知，要使得 $\mathbb{E}[R(\mathbf{x})|\mathbf{x}]$ 最小，实际上就是对每一样本使得损失 $g(f(\mathbf{x}_i), c_j)$ 达到最小。在分类任务中，满足这样条件的预测函数称为**贝叶斯最优分类器（Bayesian optimal classifier）**，对应的总体风险称为**贝叶斯风险（Bayesian risk）**。

现在假设 $\mathbf{x}_i$ 属于第 c_l 类，一个最优的分类器应该具有如下可能形式：

$$g(f(\mathbf{x}_i), c_j) = \begin{cases} 0, & \text{if } c_l = c_j; \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

那么，$R(\mathbf{x}_i) = \sum_{j=1}^{K} g(f(\mathbf{x}_i, c_j)) P(c_j|\mathbf{x}_i) = \sum_{j=1, j\neq l}^{K} P(c_j|\mathbf{x}_i) = 1 - P(c_l|\mathbf{x}_i)$，因此:

$$\min R(\mathbf{x}_i) = \max P(c_l|\mathbf{x}_i) \tag{5}$$

这说明如果一个分类器为贝叶斯最优分类器，那么对于每一个样本，应该选择**后验概率最大的类别作为其分类类别**。

那么现在问题已经转变成如何估算每一样本的后验概率。现有方法：

1. 判别式模型（discriminative models）：给定 $\mathbf{x}$，可通过直接建模 $P(c|\mathbf{x})$ 来预测其类别 c
2. 生成式模型（generative models）：先对联合概率 $P(\mathbf{x}, c)$ 建模，然后通过贝叶斯公式获得后验概率。
   a. 贝叶斯公式：

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})}, \tag{6}$$

其中，条件概率 $P(\mathbf{x}|c)$ 也称为**似然 (likelihood)**，$P(\mathbf{x})$ 是用于归一化的“证据”（evidence）因子，对给定样本 $\mathbf{x}$，证据因子与类标记无关。

于是，

$$P(c|\mathbf{x}) \propto P(\mathbf{x}|c)P(c). \tag{7}$$

因此估计后验概率的问题就转化为如何基于训练数据 $\mathbf{D}$ 来估计先验概率和似然。

- 类先验概率表达了样本空间中各类样本所占的比例。根据大数定律，当训练集包含充足的独立同分布样本时，$P(c)$ 可通过各类样本出现的频率来进行估计.
- 似然涉及了所有关于 $\mathbf{x}$ 的联合概率，直接使用频率来估计不可行。

2. 极大似然估计

先假定条件概率 $P(\mathbf{x}|c)$ 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的参数进行估计。即:

极大似然估计：是假设 $P(\mathbf{x}|c)$ 具有确定的形式并且被参数向量 θ_c 唯一确定，最大似然估计就是利用训练集 $\mathbf{D}$ 最大化条件概率 $P(\mathbf{x}|c)$，从而估计参数 θ_c。为了明确起见，将 $P(\mathbf{x}|c)$ 记为 $P(\mathbf{x}|\theta_c)$

令 $\mathbf{D}_k$ 表示训练集 $\mathbf{D}$ 中第 k 类样本组成的集合，假设这些样本 i.i.d，则参数 θ_c 对于数据集 $\mathbf{D}_c$ 的似然为：

$$L(\theta_c) = P(\mathbf{D}_c|\theta_c) = \prod_{\mathbf{x}\in\mathbf{D}_c} P(\mathbf{x}|\theta_c) \tag{8}$$

最大化 $P(\mathbf{D}_c|\theta_c)$，即可以求出 θ_c。最大化 $L(\theta_c)$ 等价于最大化 $\ln L(\theta)$，于是：

$$\ln L(\theta) = \ln P(\mathbf{D}_c|\theta_c) = \sum_{\mathbf{x}\in\mathbf{D}_c} \ln P(\mathbf{x}|\theta_c) \tag{9}$$

因此，$\theta_c = \arg\max \ln L(\theta)$。

现在假设 $p(\mathbf{x}|c) \sim \mathcal{N}(\mu_c, \sigma_c^2)$，这里参数 θ_c 实际上就包含了两个参数 μ, σ

$$\begin{aligned} \ln L(\theta) &= \sum_{\mathbf{x}\in\mathbf{D}_c} \ln P(\mathbf{x}|\theta_c) \\ &= \sum_{\mathbf{x}\in\mathbf{D}_c} \ln \left(\frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(\frac{-(\mathbf{x}-\mu_c)^2}{2\sigma_c^2} \right) \right) \\ &= \sum_{\mathbf{x}\in\mathbf{D}_c} \left(-\ln \sqrt{2\pi}\sigma_c - \frac{(\mathbf{x}-\mu_c)^2}{2\sigma_c^2} \right) \end{aligned} \tag{10}$$

然后，令 $\frac{\partial \ln L(\theta)}{\partial \mu_c} = 0, \frac{\partial \ln L(\theta)}{\partial \sigma_c^2} = 0$，可得

$$\mu_c = \frac{1}{|\mathbf{D}_c|} \sum_{\mathbf{x}\in\mathbf{D}_c} \mathbf{x} \tag{11}$$

$$\sigma_c^2 = \frac{1}{|\mathbf{D}_c|} \sum_{\mathbf{x}\in\mathbf{D}_c} (\mathbf{x}-\mu)(\mathbf{x}-\mu)^t \tag{12}$$

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，方差就是样本方差。

参数化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。在现实应用中，欲做出能较好地接近潜在真实分布的假设，往往需在一定程度上利用关于应用任务本身的经验知识。

3. 朴素贝叶斯分类器

根据前文所述，假定先验分布满足某一分布形式，可以通过最大化似然函数方法来估算先验分布。但前面估算的先验分布是样本 $\mathbf{x}$ 所有属性上的联合概率。一旦样本维数较高，难以从有限的样本准确估算出这个联合概率。

朴素贝叶斯分类器（naive Bayes classifier）采用了“属性条件独立性假设”(attribute conditional independence assumption): 对已知类别，假设所有属性相互独立。来简化似然估计，即：

$$P(\mathbf{x}|\theta_c) = P(x_1, x_2, \cdots, x_d|\theta_c) = \prod_{i=1}^{d} P(x_i|\theta_c) \tag{13}$$

d 为样本维数。

将 $P(x_i|\theta_c)$ 再记成 $p(x_i|c)$, 于是,

$$P(c|\mathbf{x}) \propto P(\mathbf{x}|c)P(c) = P(c)\prod_{i=1}^{d} P(x_i|c) \tag{14}$$

因此朴素贝叶斯最优分类器就是使得 $P(c)\prod_{i=1}^{d} P(x_i|c)$ 达到最大的分类器。

那么，如何估算 $P(c)$ 以及 $P(x_i|c)$，按照前文所述，类先验概率可以由每一类样本的频率估算，即

$$P(c) = \frac{|\mathbf{D}_c|}{|\mathbf{D}|} \tag{15}$$

而对于 $P(x_i|c)$，分为两种情况：离散属性及连续属性。

- 若 x_i 为离散属性，则条件概率 $P(x_i|c) = \frac{|\mathbf{D}_{c,x_i}|}{|\mathbf{D}_c|}$，其中 $\mathbf{D}_{c,x_i}$ 为第 i 个属性上取值等于 x_i 的样本组成的集合。
- 若 x_i 为连续属性，则条件概率 $P(x_i|c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$，其中 $\mu_{c,i}, \sigma_{c,i}^2$ 为第 c 类样本在第 i 个属性上的取值和方差，即 $P(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i-\mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$

3. 朴素贝叶斯分类器

例: 现在假设有数据如下：

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

尝试对下面这个样本进行分类

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

STEP 1: 估算类别先验概率 $P(c)$。假设好瓜属于 c_1，坏瓜属于 c_0 类，则 $P(c_1) = \frac{8}{17}, P(c_0) = \frac{9}{17}$。

STEP 2：估算每个属性的条件概率，现在 x_1=“青绿”，x_2=“蜷缩”，$\cdots$，$x_8 = 0.460$。于是：

$$P(x_1|c_1) = \frac{3}{8}, P(x_1|c_0) = \frac{3}{9}, \cdots,$$

$$p(x_8|c_1) = \frac{1}{\sqrt{2\pi}\sigma_8^1} \exp(-\frac{(0.460-\mu_8^1)^2}{2(\sigma_8^1)^2}),$$

$$p(x_8|c_0) = \frac{1}{\sqrt{2\pi}\sigma_8^0} \exp(-\frac{(0.460-\mu_8^0)^2}{2(\sigma_8^0)^2}),$$ 其中

$\mu_8^1 = 0.279, \sigma_8^1 = 0.101, \mu_8^0 = 0.154, \sigma_8^0 = 0.108,$

STEP 3.

$$P(c_1|\mathbf{x}) \propto P(c_1) \times P(x_1|c_1) \times P(x_2|c_1) \times \cdots \times p(x_8|c_1) = 0.038,$$

$$P(c_0|\mathbf{x}) \propto P(c_0) \times P(x_1|c_0) \times P(x_2|c_0) \times \cdots \times p(x_8|c_0) = 6.8 \times 10^{-5},$$

于是，该测试样本为 c_1 类。

注意到这样一种情况：若 x_3=“清脆”，则 $P(x_3|c_1) = 0$，这时根据朴素贝叶斯分类器方法，计算得到的后验概率为 0。因此，无论该样本的其他属性是什么，哪怕在其他属性上明显像好瓜，分类的结果都将是“坏瓜”，则显然不合理。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“平滑”(smoothing)，常用“拉普拉斯平滑”(Laplace smoothing)。

具体的，假设 N 表示训练集 $\mathbf{D}$ 中可能的类别数，N_i 表示第 i 个属性可能的取值数量，则先验概率和条件概率分别修改为：

$$P(c) = \frac{|\mathbf{D}_c + 1|}{\mathbf{D} + N} \tag{16}$$

$$P(x_i|c) = \frac{|\mathbf{D}_{c,x_i}| + 1}{|\mathbf{D}_c| + N_i} \tag{17}$$

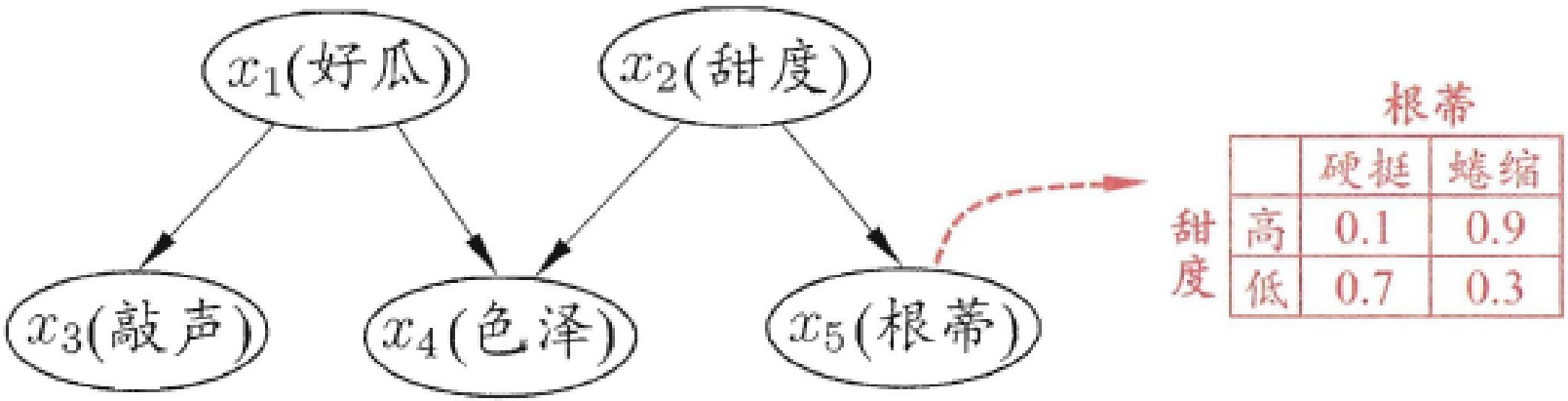
4. 贝叶斯网：结构

贝叶斯网 (Bayesian network) 借助有向无环图 (Directed Acyclic Graph，简称 DAG) 来刻画属性之间的依赖关系，并使用条件概率表 (Conditional Probability Table，简称 CPT) 来描述属性的联合概率分布。

一个贝叶斯网 B 由结构 G 和参数 Θ 两部分构成，即 $B = \langle G, \Theta \rangle$，$G$ 为一个有向无环图，每个结点对应一个属性，若两个属性之间有依赖关系，则存在一条边将其连接起来；参数 Θ 定理描述这种依赖关系，若 x_i 的父结点集合 π_i（结点 x_i 依赖的所有其他属性集合），则 Θ 包含了每个属性的条件概率表 $\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$。

4. 贝叶斯网：结构

下图给出了西瓜问题的一种贝叶斯两结构和属性“根蒂”的条件概率表。从图中网络结构可看出“色泽”直接依赖于“好瓜”和“甜度”，而“根蒂”则直接依赖于“甜度”，进一步从条件概率表能得到“根蒂”对“甜度”量化依赖关系，如 $P($ 根蒂 $=$ 硬挺 $|$ 甜度 $=$ 高 $)=0.1$ 等。



4. 贝叶斯网：结构

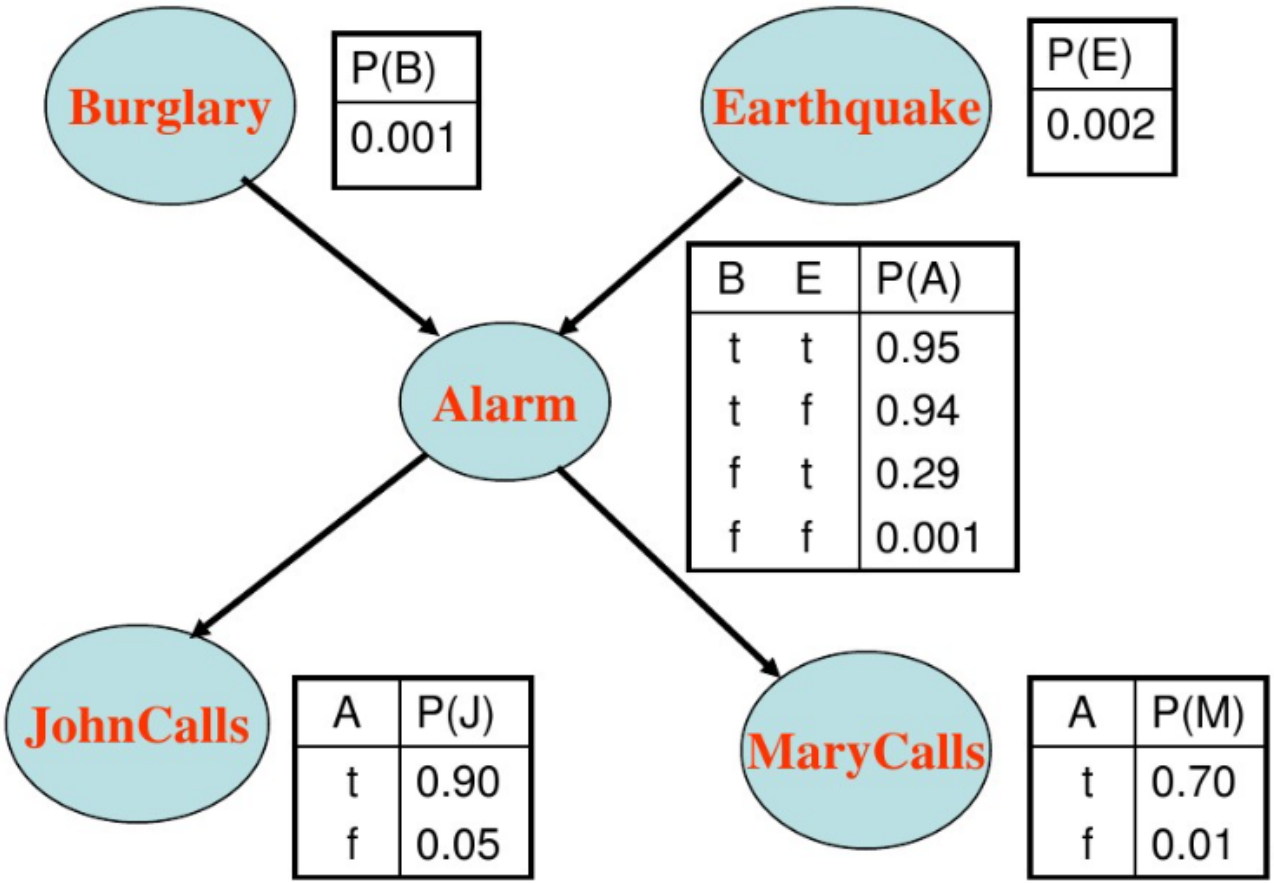
贝叶斯网结构有效地表达了属性间的条件独立性. 给定父结点集，贝叶斯网假设每个属性与它的非后裔属性独立。于是 $B = \langle G, \Theta \rangle$ 将属性 $x_1, \cdots, x_d$ 的联合概率分布定义为：

$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i} \tag{18}$$

根据上图，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 时独立，后两者分别记为 $x_3 \perp x_4|x_1, x_4 \perp x_5|x_2$，于是，联合概率

$$P_B(x_1, x_2, \cdots, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_2) \tag{19}$$

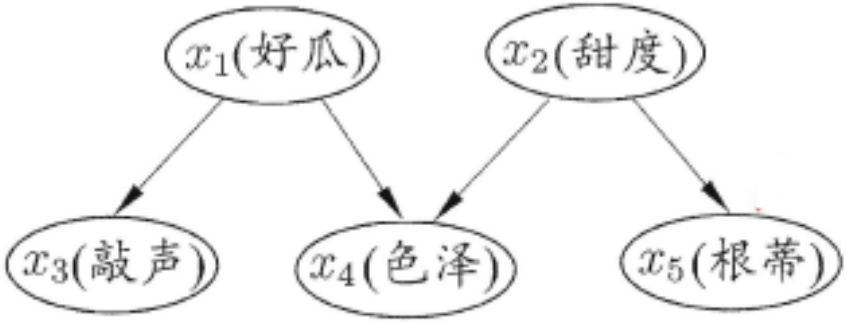
4. 贝叶斯网：结构



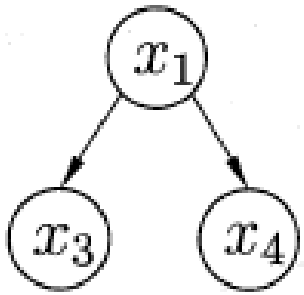
试计算，报警器响，但没有地震和盗贼，同时 John 和 Mary 打电话的概率？

4. 贝叶斯网：结构

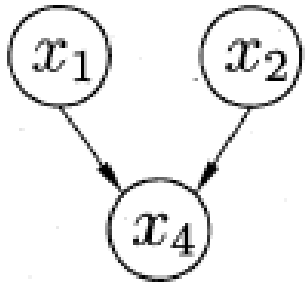
实际上，



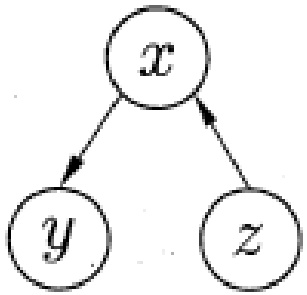
包含了结点（属性）之间的几种典型的依赖关系，如同父结构以及 V 型结构。实际上还有第三种结构：顺序结构。示意图为：



同父结构



V型结构



顺序结构

4. 贝叶斯网：结构

在同父结构中，给定父结点 x_1 的值，则 x_3, x_4 条件独立。在顺序结构中，给定 x 的值，y, z 条件独立。V 型结构称为“冲撞”结构，给定 x_4 的值，x_1, x_2 必定不独立，而在 x_4 未知时，x_1, x_2 却相互独立。这样的独立性，称为**边际独立 (marginal independence)**，记为 $x_1 \perp\!\!\!\perp x_2$

事实上，一个变量取值的确定与否，能对另两个变量间的独立性发生影响，这个现象并非 V 型结构所特有. 例如在同父结构中，条件独立性 $x_3 \perp x_4 | x_1$ 成立，但若 x_1 的取值未知，则 x_3, x_4 就不独立，即 $x_3 \perp\!\!\!\perp x_4$ 不成立; 在顺序结构中，y, z 也不是边际独立。

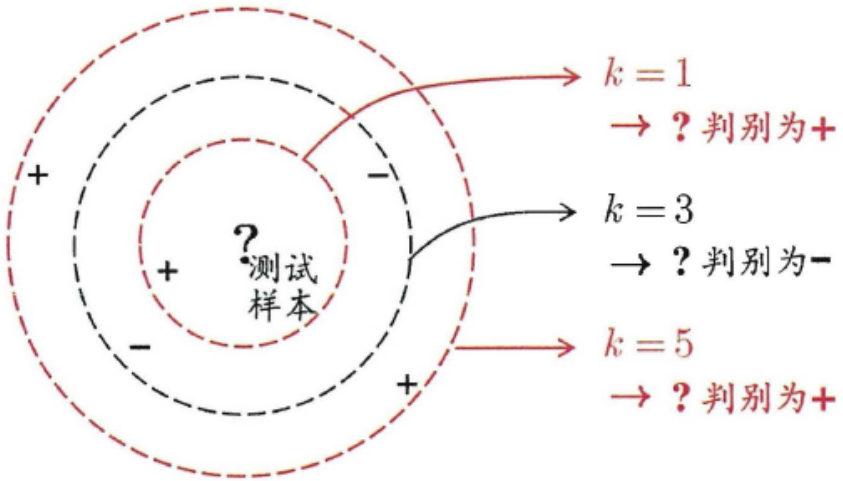
5. k 近邻学习

从上面几个算法中，可以发现，寻找一个样本的 k 个近邻是一个常用的操作。实际上,k 近邻是很多算法的基础，甚至可以直接来进行分类。

假设对于一样本 $\mathbf{x}$，能够**恰当地**找出 k 个近邻，我们可以设想这样一种方法：如果这 k 个近邻中包含某一类的样本数量占最多，那么 $\mathbf{x}$ 应该就能够被分到这一类中。这种分类方法就称为 k 近邻分类。

一个极端的情况，当 $k = 1$ 时，也即样本 $\mathbf{x}$ 的最近样本属于哪一类，就将 $\mathbf{x}$ 分到哪一类，此时称为最近邻分类。

5. k 近邻学习



我们来分析一下最近邻分类。给定测试样本 $\mathbf{x}$，若其最近邻样本为 $\mathbf{z}$，则最近邻分类器出错的概率就是 $\mathbf{x}$ 与 $\mathbf{z}$ 类别标记不同的概率。即

$$P(err) = 1 - \sum_{c} P(c|\mathbf{x})P(c|\mathbf{z}) \tag{20}$$

假设样本独立同分布，且在任意小的距离内，总能找到一个训练样本，那么

$$\begin{aligned} P(err) &= 1 - \sum_c P(c|\mathbf{x})P(c|\mathbf{z}) \\ &\cong 1 - \sum_c P(c|\mathbf{x})^2 \\ &\leq 1 - P(c^*|\mathbf{x})^2 \\ &\leq (1 + P(c^*|\mathbf{x}))(1 - P(c^*|\mathbf{x})) \\ &\leq 2 \times (1 - P(c^*|\mathbf{x})) \end{aligned} \tag{21}$$

其中 $c^* = \arg\max P(c|\mathbf{x})$ 表示贝叶斯最优分类结果。于是我们得到了有点令人惊讶的结论: 最近邻分类器虽简单，但它的泛化错误率不超过贝叶斯最优分类器的错误率的两倍!

在前面的讨论中，一直假设训练样本所有属性变量的值都已被观测到，即训练样本是“完整”的。但在现实应用中往往会遇到“不完整”的训练样本，例如由于西瓜的根蒂已脱落，无法看出是“蜷缩”还是“硬挺”，则训练样本的”根蒂“属性变量值未知。

未观测变量称为“隐变量”(latent variable)。令 $\mathbf{X}$ 表示观测变量集，$\mathbf{Z}$ 表示隐变量集，Θ 表示模型参数，对 Θ 做最大似然估计，则应该最大化对数似然 $\ln P(\mathbf{X}, \mathbf{Z}|\Theta)$，但 $\mathbf{Z}$ 是隐变量，无法求解。此时通过对 $\mathbf{Z}$ 计算期望，来最大化观测数据的**对数“边际似然”（marginal likelihood）**，即

$\ln P(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z}|\Theta)$。

5. EM 算法

EM(Expectation-Maximization) 算法是常用的估计参数隐变量的利器，它是一种迭代式的方法。其基本想法是: 若参数 Θ 已知，则可根据训练数据推断出最优隐变量 $\mathbf{Z}$ 的值 (E 步) ; 反之，若 $\mathbf{Z}$ 的值已知，则可方便地对参数 Θ 做极大似然估计 (M 步) 。即：以初始值 Θ^0 为起点，

E Step: 基于 Θ^t 推断隐变量 $\mathbf{Z}$ 的期望，记为 $\mathbf{Z}^t$;

M Step: 对于已观测变量 $\mathbf{X}$ 和 $\mathbf{Z}^t$，对参数 Θ 做极大似然估计，记为 Θ^{t+1}，跳回 E Step。

Thanks !