贝叶斯分类器

1.贝叶斯定理

假设X, Y 是一对随机变量，它们的**联合概率**$P(X=x, Y=y)$是指 X 取值 x 且 Y 取值 y 的概率，**条件概率**是指一随机变量在另一随机变量取值已知的情况下某一特定值的概率。例如，条件概率$P(Y=y|X=x)$是指在变量X取值x的情况下，变量 Y取值 y的概率。

X和Y的联合概率和条件概率满足如下关系：

$$P(X,Y)=P(Y|X)\cdot P(X)=P(X|Y)\cdot P(Y)$$

由上面的公式可以得到下面公式，称为**贝叶斯定理**：

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

2.朴素贝叶斯分类器

朴素贝叶斯的基本思想是：**对于给定的待分类项，求出在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。**

朴素贝叶斯分类器有一个假设：**给定类标号y_k，各属性之间条件独立。**条件独立假设表述如下：

$$P(X|Y=y_k)=\prod_{i=1}^{m}P(x_i|Y=y_k)$$

其中，每个属性集$X=x_1,x_2,\cdots,x_m$包含 m 个属性。

朴素贝叶斯分类器的工作流程如下：

> 1. 设$X=x_1,x_2,\cdots,x_m$是一个待分类项，而x_i为X的一个特征属性集；
> 2. 设类别集合$C=y_1,y_2,\cdots,y_n$;
> 3. 计算$P(y_1|X),P(y_2|X),\cdots,P(y_n|X)$;

4. 如果$P(y_k|X)=maxP(y_1|X),P(y_2|X),\cdots,P(y_n|X)$，则$X\in y_k$。

其中，最关键的是如何计算第三步中的各个条件概率，方法如下：

1. 给定一个训练样本集，统计在各类别下各个特征属性的条件概率，即

$$\begin{gathered}P(x_1|y_1),P(x_2|y_1),\cdots,P(x_m|y_1)\\P(x_1|y_2),P(x_2|y_2),\cdots,P(x_m|y_2)\\\vdots\\P(x_1|y_n),P(x_2|y_n),\cdots,P(x_m|y_n)\end{gathered}$$

2. 根据朴素贝叶斯分类器的假设，以及**贝叶斯定理**有：

$$P(y_k|X)=\frac{P(X|y_k)P(y_k)}{P(X)}$$

其中，

$$P(X|y_k)=\left[\prod_{i=1}^{m}P(x_i|y_k)\right]$$

由于对于所有的Y，$P(X)$是固定的，因此只要**找出使分子**$\prod_{i=1}^{m}P(x_i|y_k)$**最大的类**就够了。

由上文可以看出，计算在各个类别下特征属性的条件概率$P(X=x_i|Y=y_k)$是朴素贝叶斯分类的关键性步骤：

- 当特征属性为离散值时，可以通过统计训练样本中各个划分在每个类别中出现的频率来估计；
- 当特征属性为连续属性时，可以假设连续变量服从某种概率分布，然后使用训练数据估计分布的参数。高斯分布（也称正态分布）常被用来表示连续属性的类条件概率分布。该分布有两个参数，均值μ和方差σ^2。对每个类，属性x_i的类条件概率等于：

$$P(X=x_i|Y=y_k)=\frac{1}{\sqrt{2\pi}\sigma_{ik}}\exp\left[-\frac{(x_i-\mu_{ik})^2}{2\sigma_{ik}^2}\right]$$

参数μ_{ik}可以用类y_k的所有训练记录关于x_i的样本均值$\bar{x}$来估计，参数σ_{ik}^2可以用这些训练记录的样本方差s^2来估计

3.半朴素贝叶斯

1. 朴素贝叶斯法对条件概率做了特征的独立性假设：$p(\vec{x},y)=p(x_1,x_2,..,x_n|y)=\prod_{j=1}^{n}p(x_i|y)$。

   但是现实任务中这个假设有时候很难成立。若对特征独立性假设进行一定程度上的放松，这就是半朴素贝叶斯分类器`semi-naive Bayes classifiers`。

2. 半朴素贝叶斯分类器原理：适当考虑一部分特征之间的相互依赖信息，从而既不需要进行完全联合概率计算，又不至于彻底忽略了比较强的特征依赖关系。

3.1 独依赖估计 OED

1. 独依赖估计`One-Dependent Estimator:OED`是半朴素贝叶斯分类器最常用的一种策略。它假设每个特征在类别之外最多依赖于一个其他特征，即：

$$p(\vec{x},y)=p(x_1,x_2,..,x_n|y)=\prod_{j=1}^{n}p(x_j|y,x_j^p)$$

其中 x_j^p为特征 x_i所依赖的特征，称作的 x_j父特征。

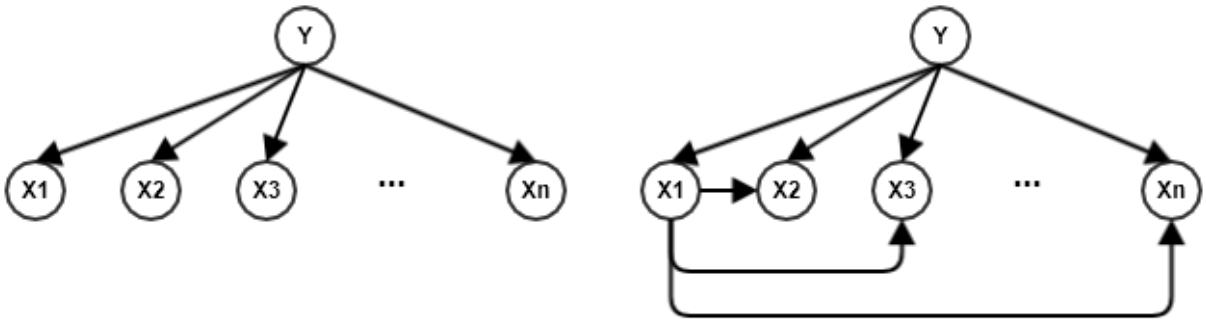
2. 如果父属性已知，那么可以用贝叶斯估计来估计概率值 $p(x_j|y,x_j^p)$。现在的问题是：如何确定每个特征的父特征？

   不同的做法产生不同的独依赖分类器。

3.1.1 SPODE

1. 最简单的做法是：假设所有的特征都依赖于同一个特征，该特征称作超父。然后通过交叉验证等模型选择方法来确定超父特征。这就是`SPODE:Super-Parent ODE`方法。

   假设节点`Y`代表输出变量，节点`Xj`代表属性x_j。下图给出了超父特征为 x_1时的`SPODE`。



朴素贝叶斯　　半朴素贝叶斯：SPODE

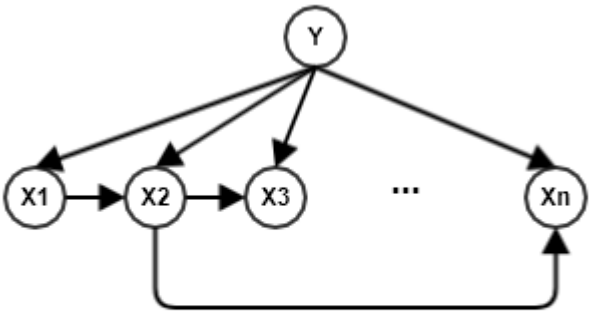
3.1.2 TAN

1. `TAN:Tree Augmented naive Bayes`是在最大带权生成树算法基础上，通过下列步骤将特征之间依赖关系简化为如下图所示的树型结构：
   - 计算任意两个特征之间的条件互信息。记第 i 个特征 x_i代表的结点为X_i，标记代表的节点为 Y 则有:

     $$I(X_i,X_j|Y)=\sum_{y}\sum_{x_i}\sum_{x_j}p(x_i,x_j|y)\log\frac{(p(x_i,x_j|y)}{p(x_j|y)p(x_j|y)}$$

     如果两个特征 x_i,x_j相互条件独立，则 $p(x_i,x_j|y)=p(x_i|y)p(x_j|y)$。则有条件互信息$I(X_i,X_j|Y)=0$，则在图中这两个特征代表的结点没有边相连。
   - 以特征为结点构建完全图，任意两个结点之间边的权重设为条件互信息$i(X_i,X_j|Y)$。
   - 构建此完全图的最大带权生成树，挑选根结点（下图中根节点为节点X_1），将边置为有向边。

- 加入类别结点Y ，增加 Y 到每个特征的有向边。因为所有的条件概率都是以 y 为条件的。



半朴素贝叶斯：TAN

4.贝叶斯网络

贝叶斯网络($Bayesian\ Network$)，又称信念网，借助有向无环图($Directed\ Acyclic\ Graph$)来刻画属性之间的依赖关系。使用条件概率表($Condition\ Probability\ Table$)来描述属性之间的联合概率。

具体来说,一个贝叶斯网 B 由结构 G 和参数 Θ 两部分构成，即$B = (G, \Theta\rangle$。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

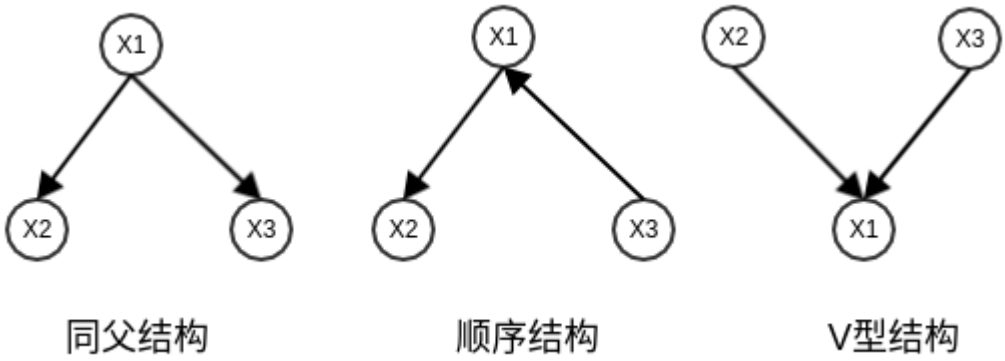
$$\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$$

4.1结构

贝叶斯的联合概率分布定义为

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

贝叶斯的三种变量之间的典型依赖关系分别是



4.1.1同父结构

给定父节点x_1 的取值，则 x_2 和 x_3 条件独立，即

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

4.1.2顺序结构

给定中间节点 x_1 的值，则 x_2 和 x_3 条件独立，即

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

当给定 x_1 时，x_2 和 x_3 之间的关系被阻断，因此他们关于 x_1 条件独立。

4.1.3 V型结构

给定子节点 x_1 的值，则x_2 与 x_3 必定不是条件独立的。即

$$P(x_2, x_3|x_1) \neq P(x_2|x_1)P(x_3|x_1)$$

事实上x_2 与 x_3 时独立的，但不是条件独立。即$P(x_2, x_3) = P(x_2)P(x_3)$

> 为了分析有向图中节点之间的条件独立性，可以使用有向分离技术：
>
> - 找出有向图中的所有`V`型结构，在`V`型结构的两个父节点之间加上一条无向边。
> - 将所有的有向边改成无向边。
>
> 这样产生的无向图称作道德图`moral graph`。父节点相连的过程称作道德化`moralization`。基于道德图能直观、迅速的找到结点之间的条件独立性。

4.2 学习

1. 贝叶斯网络的学习可以分为参数学习和结构学习两部分
   - 参数学习比较简单。只需要通过对训练样本“计数”，估计出每个结点的条件概率表即可。

     但是前提是必须知道网络结构。
   - 结构学习比较复杂，结构学习被证明是`NP`难问题。
2. 贝叶斯网络的结构学习通常采用`评分搜索`来求解。
   - 先定义一个评分函数，以此评估贝叶斯网络与训练数据的契合程度。然后基于这个评分函数寻找结构最优的贝叶斯网。
   - 最常用的评分函数基于信息论准则：将结构学习问题视作一个数据压缩任务。
     - 学习的目标是找到一个能以最短编码长度描述训练集数据集的模型。这就是`最小描述长度 Minimal Description Length:MDL`准则。
     - 此时的编码长度包括了：描述模型自身所需要的字节长度，和使用该模型描述数据所需要的字节长度。
3. 给定训练集 $D = \{x_1, x_2, \ldots, x_m\}$,贝叶斯网络$B = (G, \Theta)$ 再D上的评分函数可写为

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

   其中$f(\theta)$ 表示描述每个参数 θ 所需的字节数，$|B|$ 是贝叶斯网络参数的个数。

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

   是贝叶斯网B 的对数似然，因此：
   - 第一项 $f(\theta)|B|$是计算编码贝叶斯网络B 所需要的字节数。
   - 第二项 $\sum_{i=1}^{m} \log P_B(x_i)$是计算 B 所对应的概率分布 P 需要多少字节来描述 。
4. 现在结构学习任务转换为一个优化任务，即寻找一个贝叶斯网络 使评分函数$s(B|D)$ 最小。

问题是，从所有可能的网络结构空间中搜索最优贝叶斯网络结构是个`NP`难问题，难以快速求解。

有两种方法可以在有限时间内求得近似解：

- 贪心算法。如从某个网络结构出发，每次调整一条边，直到评分函数不再降低为止。
- 增加约束。通过给网络结构增加约束来缩小搜索空间，如将网络结构限定为树形结构等。

5. 贝叶斯网络训练好之后就能够用来进行未知样本的预测。

   最理想的是直接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，但问题是这样的"精确推断"已经被证明是`NP`难的。

   此时需要借助"近似推断"，通过降低精度要求从而在有限时间内求得近似解，常用的近似推断为吉布斯采样(`Gibbs sampling`)。

1.朴素贝叶斯分类器的优点： 性能相当好，它速度快，可以避免维度灾难。 支持大规模数据的并行学习，且天然的支持增量学习。

2.朴素贝叶斯分类器的缺点： 无法给出分类概率，因此难以应用于需要分类概率的场景。