

贝叶斯分类器

贝叶斯分类器

1. 贝叶斯定理

2. 朴素贝叶斯分类器

3. 半朴素贝叶斯

3.1 独立估计 OED

3.1.1 SPODE

3.1.2 TAN

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 $X = x_1, x_2, \dots, x_m$ 包含 m 个属性。

朴素贝叶斯分类器的工作流程如下：

1. 设 $X = x_1, x_2, \dots, x_m$ 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 $C = y_1, y_2, \dots, y_n$ ；
3. 计算 $P(y_1|X), P(y_2|X), \dots, P(y_n|X)$ ；
4. 如果 $P(y_k|X) = \max P(y_1|X), P(y_2|X), \dots, P(y_n|X)$ ，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

$$\begin{aligned}
&P(x_1|y_1), P(x_2|y_1), \dots, P(x_m|y_1) \\
&P(x_1|y_2), P(x_2|y_2), \dots, P(x_m|y_2) \\
&\vdots \\
&P(x_1|y_n), P(x_2|y_n), \dots, P(x_m|y_n)
\end{aligned}$$

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

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：

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

其中 x_j^p 为特征 x_i 所依赖的特征，称作的 x_j 父特征。

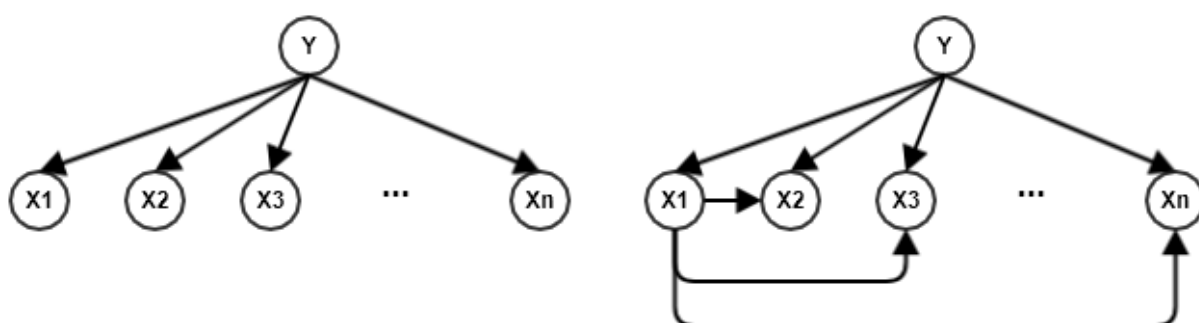
2. 如果父属性已知，那么可以用贝叶斯估计来估计概率值 $p(x_j|y, x_j^p)$ 。现在的问题是：如何确定每个特征的父特征？

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 x_j 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：

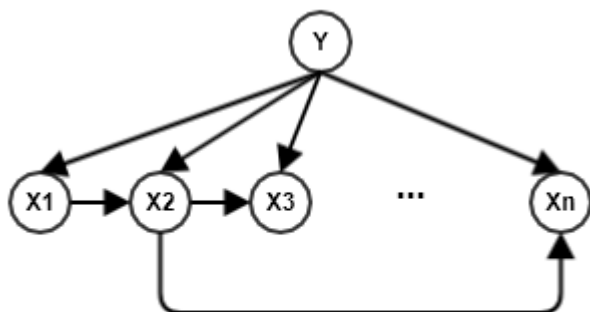
- 计算任意两个特征之间的条件互信息。记第 i 个特征 x_i 代表的结点为 X_i ，标记代表的节点为 Y 则有：

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

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

- 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。