贝叶斯分类器

1、贝叶斯定理:

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

含义是给定一个样本x,它的分类结果是c的概率为$P(c|x)$

$P(c|x)$也称为后验概率。$P(c)$称为先验概率

2. 朴素贝叶斯分类器

假设条件: ① 所有样本独立同分布(i.i.d.)

② 样本的所有属性相互独立

所以有:

总共有d个属性.

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

,其中$x^{(i)}$表示样本x的第$\boxed{i}$个属性.

我们的目标是 $\max P(c|x)$,这样可以使$R(c|x)=1-P(c|x)$最小。若保证对每个独立样本x_i均有$R(c|x_i)$最小,那么$E(R(c|x))$也可以最小化,整个分类器性能就是全局最优的

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

,对于所有的单个x,$P(x)$均相同,所以$P(x)$的取值不影响$P(c|x)$. $P(c|x)\sim P(c)\cdot\prod_{i=1}^{d}P(x^{(i)}|c)$

其中: $P(c)=\frac{|D_c|}{|D|}$ 表示标签为c的样本占总样本的比例

$$P(x^{(i)}|c)=\begin{cases}\frac{|D_{x^{(i)},c}|}{|D_c|} & \text{离散样本中,属性为}x^{(i)}\text{且标签为}c\text{的样本占}|D_c|\text{的比例}\\ p(x^{(i)}|\theta_c) & \text{连续样本中,使用密度函数表示取值.}\end{cases}$$

θ_c为密度函数的参数,使用最大似然方法去计算。对每个分类标签c,均对应一个一维的密度函数$p(x^{(i)}|c)$

3. 最大似然估计

假设有密度函数 $P(X|\theta)$，我们知道函数表达式，但是不知道具体的参数，我们可根据有限的样本X去估计θ的值。

定义 $L(\theta)=\prod_{i=1}^{N}P(x_i|\theta)$，$l(\theta)=\sum_{i=1}^{N}\ln P(x_i|\theta)=\ln L(\theta)$

对 $l(\theta)$ 中所有参数求偏导，有：

$$\begin{cases}\frac{\partial l(\theta)}{\partial\theta_1}=0\\ \vdots\\ \frac{\partial l(\theta)}{\partial\theta_n}=0\end{cases}\Rightarrow\begin{cases}\theta_1=a_1\\ \vdots\\ \theta_n=a_n\end{cases}$$

就可以得到基于最大似然估计的密度函数 $P(X|\hat{\theta})$

可用于估计连续属性的值

4. 拉普拉斯修正

1）对于 $P(c)=\frac{|D_c|+1}{|D|+N}$，其中N是所有标签的种类数。

2）对于 $P(x^{(i)}|c)=\frac{|D_{x^{(i)},c}|+1}{|D_c|+N^{(i)}}$，其中 $N^{(i)}$ 是属性 $x^{(i)}$ 的所有可取值个数

5. 半朴素贝叶斯分类器

$$P(c|x)\sim P(c)\prod_{i=1}^{d}P(x_i|c,pa_i)$$

朴素贝叶斯分类器基于的假设是所有属性都相互独立，但是这在现实中是很难实现的。半朴素贝叶斯分类器则假设所有的属性 x_i 均依赖一个父属性 pa_i

决定父属性的方法：

① 假设所有属性都依赖同一个父属性，这个父属性被称为超父（super parent）
使用SPODE算法去枚举验证超父

② TAN算法是构建最大生成树，对每个属性确定父属性

- TAN算法定义 $I(x_i, x_j | y) = \sum_{x_i, x_j, c \in y} P(x_i, x_j | c) \log \frac{P(x_i, x_j | c)}{P(x_i | c) P(x_j | c)}$

  是条件互信息，定义这个值是边的权重 $w = I(x_i, x_j | y)$
- 构建最大生成树，生成的 $(n-1)$ 条边具有最大权值和 $\sum w_i$ 可使用Kruskal实现
- 之后挑选根变量，把边置为有向，就构成了父结点的有向图.
- 最后增加 y 指向所有的属性 x_i

③ AODE算法也可以确定超父结点