决策树

主要有ID3, C4.5, CART算法.

分类决策树表示基于特征对实例进行分类的过程. 主要包括3个步骤:特征选择, 决策树的生成 和 决策树的修剪.

决策树学习的算法, 通常是一个递归地选择最优特征, 并根据该特征对训练数据进行分割, 使得对各个子数据集有一个最好的分类的过程.

主要优点:

模型具有可解释性, 容易向业务部门人员描述.

分类速度快.

基础知识

熵\条件熵\经验熵\经验条件熵\信息增益\信息增益比\基尼系数

熵

定义:表示随机变量X不确定性的度量. 设X是一个取有限个值的离散随机变量, 其概率分布为

$$P(X = x_i) = p_i, i = 1, 2, \ldots, n$$

则随机变量X的**熵定义**为

$$H(X) = -\sum_{i=1}^{n} p_i log(p_i)$$

条件熵

定义:

条件熵$H(Y|X)$表示在已知随机变量X的条件下随机变量Y的不确定性. 随机变量X给定的条件下随机变量Y的条件熵$H(Y|X)$, 定义为X给定条件下Y的条件概率分布的熵对X的数学期望:

$$H(Y|X)=\sum_{i=1}^{n}p_iH(Y|X=x_i)$$

这里$p_i=P(X=x_i),i=1,2,\ldots,n$.

经验熵\经验条件熵

当熵和条件熵中的概率由数据估计(特别是极大似然估计)得到时, 所对应的熵和条件熵分别称为经验熵和条件经验熵. 此时如果有0概率, 令$0log0=0$.

条件熵是由真实的概率分布求得的, 经验熵是由一个数据集估计出来的概率分布求得的.

信息增益

$$g(D,A)=H(D)-H(D|A)$$

表示得知特征X的信息而使得类Y的信息的不确定性减少的程度.

不同的特征往往具有不同的信息增益, 信息增益大的特征具有更强的分类能力.

设数据集为D, $|D|$表示其样本容量, 即样本个数. 设有K个类$C_k,k=1,2,\ldots,K,|C_k|$为属于类C_k的样本数, $\sum_{k=1}^{K}|C_k|=|D|$. 设特征A有n个不同的取值$a_1,a_2,\ldots,a_n$, 根据特征A的取值将D划分为n个子集$D_1,D_2,\ldots,D_n,|D_i|$为D_i的样本个数, $\sum_{i=1}^{n}|D_i|=D$. 记子集D_i中属于类C_k的样本的集合为D_{ik}, 即$D_{ik}=D_i\bigcap C_k,|D_{ik}|$为$D_{ik}$的样本个数. 于是信息增益的算法如下:

(1) 计算数据集D的经验熵H(D)

$$H(D)=\sum_{i=1}^{K}\frac{|C_k|}{|D|}log_2\frac{|C_k|}{|D|}$$

(2)计算特征A对数据集D的条件经验熵H(D|A)

$$H(D|A)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{i=1}^{K}\frac{|D_{ik}|}{|D_i|}log_2\frac{|D_{ik}|}{|D_i|}$$

(3)计算信息增益

$$g(D,A)=H(D)-H(D|A)$$

信息增益比

特征A对训练数据集D的信息增益比定义$g_R(D,A)$定义为其信息增益$g(D,A)$与训练数据集D关于特征A的值的熵$H_A(D)$之比:

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

其中, $H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}log2\frac{|D_i|}{D}$, n 是特征A取值的个数.

基尼指数

1.分类问题中, 假设有K个类, 样本点属于第k个类的概率为p_k, 则**概率分布的基尼指数**定义为

$$Gani(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}{p_k}^2$$

2.**对于给定样本集合D, 其基尼指数为**

$$Gani(D)=1-\sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2$$

这里C_k是D中属于第k类的样本子集, K是类的个数.

3.如果样本集合D根据特征A是否取某一可能值a被分隔成D_1和D_2两部分, 即

$$D_1=(x,y)\in D|A(x)=a,D_2=D-D_1$$

则**在特征A的条件下, 集合D的基尼指数**定义为

$$Gani(D,A=a)=\frac{|D_1|}{|D|}Gani(D_1)+\frac{|D_2|}{|D|}Gani(D_2)$$

基尼指数Gani(D)表示集合D的不确定性, 基尼指数Gani(D, A=a)表示经A=a分隔后集合D的不确定性.基尼指数值越大, 样本集合的不确定性也就越大, 这一点与熵相似.