C4.5算法

C4.5 Algorithm

deepshare.net 深度之眼

C4.5是对ID3的改进，在树生成过程中，使用信息增益比来选择特征

ID3计算信息增益公式：

$$g(D,A)=H(D)-H(D\mid A)$$

C4.5 计算信息增益公式：

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

其中 $H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$

n是特征A取值的个数

ID	年龄	有工作	有房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

在计算特征的每个取值下的熵的时候，需要估计每个类别k的概率！显然我们是直接用频率去近似的概率。然而根据大数定律，只有当样本数足够多的时候，频率才可以准确的近似概率。样本数越少，对概率的估计结果的方差就会越大（想象一下做抛硬币实验来近似正面向上的概率，如果只抛两次，那么得到的正面向上的概率可能会非常离谱。而如果抛1万次，不论何时何地几乎总能得到近似0.5的概率）而方差大会导致什么结果呢？显然就会导致该取值下的类别错估计为非均匀分布呀！而非均匀分布，不就是说导致该取值下的熵更小了。所以说，ID3决策树，或者说信息增益，或者说条件熵，并不是说一定会偏向于取值多的特征，而是数据集的不充足以及客观存在的大数定律导致取值多的特征在计算条件熵时容易估计出偏小的条件熵。

C4.5算法

C4.5 Algorithm

deepshare.net 深度之眼

C4.5 计算信息增益公式：

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

其中 $H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$

n是特征A取值的个数

ID	年龄	有工作	有房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

关注公众号深度之眼，后台回复资料，获取AI必学书籍及完整实战学习资料

计算 $H_A(D)$，以"信贷情况"为例：

$H_A(D)=-(\frac{5}{15}\log_2\frac{5}{15}+\frac{6}{15}\log_2\frac{6}{15}+\frac{4}{15}\log_2\frac{4}{15})$

一般　好　非常好

为什么信息增益除以 $H_A(D)$ 可以起到惩罚的作用？

基尼指数

Gini Index

deepshare.net 深度之眼

CART算法：在树生成过程中，使用基尼指数选择最优特征。生成的是二叉树。

1、设样本点属于第k类的概率为p_k，则：$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$

2、给定样本集合D，其基尼指数为：$Gini(D)=1-\sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2$

3、在特征A的条件下，集合D的基尼指数为：

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D1)+\frac{|D_2|}{|D|}Gini(D2)$$

基尼指数越大，样本的不确定性越大，这点与熵相似！

基尼指数

$Gini(D)=1-\sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2$　$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D1)+\frac{|D_2|}{|D|}Gini(D2)$

Gini Index

deepshare.net 深度之眼

用所给数据集D，根据基尼指数求最优特征：

$$Gini(D,A_1=1)=\frac{5}{15}(1-((\frac{2}{5})^2+(\frac{3}{5})^2))+\frac{10}{15}(1-((\frac{3}{10})^2+(\frac{7}{10})^2))=0.44$$

$Gini(D,A_1=2)=0.48$

$Gini(D,A_1=3)=0.44$

$Gini(D,A_2=1)=0.32$

$Gini(D,A_3=1)=0.27$

$Gini(D,A_4=1)=0.36$

$Gini(D,A_4=2)=0.47$

$Gini(D,A_4=3)=0.32$

ID	年龄	有工作	有房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

因此选A_3为最优特征，$A_3=1$为最优切分点。

为什么基尼指数越大，混乱程度越大？