决策树

Half

2022 年 1 月 28 日

1 基本概念

决策树的基本原理是我们不断地根据我们的属性去对元素进行筛选, 得到最终的结果, 但是根据在不同时间内对属性的选择, 我们算法的执行效率和最终的结果可能会有所不同, 所以我们需要一种选择划分属性的算法

2 划分选择

2.1 信息增益

信息熵是我们度量样本集合纯度, 定义信息熵的公式为

$$Ent(D) = \sum_{k=1}^{|y|} p_k \log_2 p_k \tag{1}$$

我们的信息熵越小, 越能说明样本的纯度越高

此外我们还有一个信息增益, 通过给不同的子节点赋予不同的权值, 来表示不同分支的影响

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{D} Ent(D^v) \tag{2}$$

我们的 ID3 决策树算法就是使用信息增益为准则来选择划分的属性

2.2 增益率

当我们的样本的种类过多的时候, 我们将发现此时的划分的泛化能力较弱, 比如我们用编号作为唯一的划分标准的时候, 我们将发现甚至每个节点都有不同的属性...

增益率的定义如下

$$\begin{aligned} Gain\_ratio(D,a) &= \frac{Gain(D,a)}{IV(a)} \\ IV(a) &= -\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|} \end{aligned} \tag{3}$$

我们会选择最大的增益率, 当我们的样本的种类过多的时候, 这个值会变小

2.3 基尼系数

下面我们给出我们的基尼值的计算方式

$$Gini(D) = \sum_{k=1}^{|y|}\sum_{k'!=k} p_k p_{k'} \tag{4}$$

我们可以将之前用在增益率上面的方法用在我们的基尼系数上面, 可以得到属性 a 的基尼质数

$$Gini\_index(D,a) = \sum_{v=1}^{V}\frac{|D^v|}{|D|}Gini(D^v) \tag{5}$$

我们通常选择基尼系数最小的那个作为我们的划分属性

3 剪枝处理

在进行拟合地时候, 可能会出现过拟合的现象, 即将训练集本身的一些特点当作是所有数据的一般性质而导致过度拟合, 为了解决这个问题, 我们会使用预减枝和后减枝条两种方式

预剪枝指的是在决策树生成过程中, 对每个即诶单在划分前先进行估计, 若当前的节点的划分不能带来决策树泛化性能的提升, 那么就停止划分并将当前的节点标记为叶节点, 后剪枝则是从头训练集中生成一颗完整的决策树, 然后自底向上地对叶节点进行考察, 若将该节点的子树替换为叶节点能带来决策树泛化性能的提升, 那么将该子树替换为叶节点

对于判断性能的提升, 我们可以使用留出法, 将一部分数据作为训练集, 一部分作为检测集