机器学习常见知识汇总之决策树

田晓彬
xiaobin9652@163.com

决策树

1.决策树的基本思想

决策树有若干个内部节点和叶子节点。叶子节点对应决策结果，内部节点对应一个特征或者属性。从决策树的根节点开始，样本被划分到不同的子节点中，再根据子节点的特征进行进一步划分，直至所有样本都被归到某一类别（叶子节点）中。决策树学习的目的是为了产生一颗泛化能力强，即处理未见数据能力强的决策树。决策树的生成包含特征选择，树的构造，树的剪枝三个过程。下面详细介绍。

2.决策树的三种构造方法

再具体情况中，我们即希望决策树能有效地拟合数据，达到良好的分类效果，同时又希望可以控制决策树的复杂度，使模型具有一定的泛化能力。而决策树的选择的可能性有很多种，从中选取最优的决策树是一个NP难问题。我们通常使用启发式学习的方法来构造一颗满足启发式条件的决策树。根据选取特征的不同量化评估指标，可以将常用的决策树构造算法分为三种：ID3、C4.5、CART。

1. **ID3-最大信息增益**

信息熵是度量样本集合纯度最常用的一种指标。样本集合 D 的信息熵定义为

$$Ent(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

其中样本集合 D 中第 k 类样本所占的比例为 p_k。 $Ent(D)$ 的值越小，则 D 的纯度越高。

假设离散属性 a 有V个可能的取值，D^v 为 D 中所有在属性 a 上取值为 a^v 的样本集。不同的 D^v 包含的样本数不同，给子集赋予权重 D^v/D，即样本数越多的子集影响越大，于是可以计算属性 a 对样本集 D 进行划分所获得的的信息增益

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

信息增益越大，则意味着使用属性 a 来进行划分所获得的的信息纯度提升越大。我们使用信息增益来确定决策树的属性划分选择，即找到所有为确定属性中，信息增益最大的属性进行划分决策树。

2. **C4.5-最大信息增益比**

C4.5算法不直接使用信息增益，而是使用增益率来选择最优划分属性。增益率定义为：

$$Gainratio(D,a)=\frac{Gain(D,a)}{H_a(D)}$$

其中

$$H_a(D)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$$

称为属性的固有值，属性 a 的可能取值的数目越多，则 $H_a(D)$ 的值越大。

增益率准则对可取值数目较少的属性有所偏好，C4.5算法并不直接使用增益率最大的候选划分属性，而先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的。

3. **CART-最小基尼指数**

CART决策树使用基尼指数来选择划分属性。数据集 D 的纯度可以用基尼值来度量：

$$Gini(D)=\sum_{k=1}^{K}\sum_{k'\neq k}p_kp_{k'}=1-\sum_{k=1}^{K}p_k^2$$

$Gini(D)$ 反映了从数据集 D 中随机抽取两个样本，其标记不一致的概率。 因此 $Gini(D)$ 越小，则数据集 D 的纯度越高。由此可知属性 a 的基尼指数定义为：

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

在划分候选属性集合的时候，选择使划分后基尼指数最小的属性作为最优划分属性。

注：三种特征选择方法中，只有基尼指数为选择最小的，信息增益和信息增益比都为选择最大的。

3.三种决策树构造准则的异同

首先，ID3采用信息增益作为评价指标，但是信息增益反映的是给定条件后不确定性减少的程度，这个值会倾向于特征取值多的特征。特征取值越多意味着确定性越高，也就是条件熵越小，即信息增益越大。但通常特征数过多的特征的泛化能力是非常弱的。C4.5通过引入信息增益比，在一定程度上对取值比较多的特征进行惩罚，避免出现ID3中的过拟合的特征，提高决策树的泛化能力。

其次，从样本的角度，ID3只能处理离散型变量，而C4.5和CART都可以处理连续型变量。

从应用角度，ID3和C4.5只能用于分类任务，CART(Classification and Regression Tree)还可以用于回归任务（使用最小平方误差准侧）。

此外，ID3对样本特征缺失值比较敏感，而C4.5和CART可以对缺失值进行不同方式的处理。ID3和C4.5可以在每个节点上产生出多叉分支，且每个特征在层级之间不会复用，而CART每个节点只会产生两个分支，而且每个特征可以被重复使用。ID3和C4.5通过剪枝来权衡决策树的准确性和泛化能力，而CART直接利用全部的数据发现所有可能的树结构进行对比。

4.决策树的剪枝策略

一颗完全生长的决策树会面临过拟合这个严重的问题。这是因为在决策树的学习过程中，为了尽可能正确的确定划分训练样本，把训练样本的一些特点当作所有数据具有的一般特性。剪枝是决策树学习算法中应对过拟合的主要手段。决策树的剪枝通常有两种手段，预剪枝和后剪枝。

1. 预剪枝

   预剪枝是在对决策树中节点进行扩展之前进行评估，若当前节点的划分不能带来决策树泛化性能的提升，则停止划分并将当前节点标记为叶节点。标记的类别为当前节点集合中数量最多的类别。预剪枝对何时停止决策树的生长有以下几种方法：

   1）当树达到一定深度的时候，停止树的生长；

   2）当到达当前节点的样本数量小于某个阈值的时候，停止树的生长；

   3）计算每次分裂对测试集的准确度会有提升，当这个提升小于某个阈值的时候，不在继续扩展。

   预剪枝使得决策树的很多分支都没有展开，这不仅降低了过拟合的风险，还显著减少了决策树的时间开销。但决策树的预剪枝有一定的局限性，可能有欠拟合的风险，虽然当前的划分会导致测试集的准确率降低，但是后续的划分可能会使准确率有显著的上升。

2. 后剪枝

   后剪枝先从训练集中生成一颗完整的决策树，然后自底向上地对非叶子节点进行考察。首先将该节点对应的子树替换成叶子节点，叶子节点的类别为当前节点集合中数量较多的类别。然后计算替换后的决策树在测试集上的泛化能力，若泛化能力提升，那么将该节点剪枝。相对于预剪枝，后剪枝方法通常可以得到泛化能力强的决策树，但时间开销会更大。

   常见的后剪枝方法包括：错误率降低剪枝（REP）、悲观剪枝（PEP）、代价复杂度剪枝（CCP）、最小误差剪枝（MEP）、CVP、OPP等方法。

5.决策树的连续值处理

C4.5和CART都可以对连续值进行处理。这里介绍C4.5使用的简单二分法对属性进行划分。对于样本集 D 和连续属性 a，假定 a 在 D 上出现了 n 个不同的取值，首先将这 n 个不同的取值进行排序。假定划分点为 t，基于划分点 t 我们可以将数据集划分成两部分 D_t^- 和 D_t^+，其中 D_t^- 是在属性 a 上取值不大于 t 的样本，D_t^+ 是在属性 a 上取值大于 t 的样本。因此对于属性 a，我们可以得到 $n-1$ 个候选划分点集合

$$T_a = \left\{ \frac{a^i + a^{i+1}}{2} \mid 1 \leq i \leq n-1 \right\}$$

即把区间 $\left[a^i, a^{i+1}\right)$ 的中位点作为候选划分点。那么属性 a 在数据集 D 上的信息增益为

$$Gain(D,a)=\max_{t\in T_a} Gain(D,a,t)=\max_{t\in T_a} Ent(D)-\sum_{\lambda\in\{-,+\}}\frac{|D_t^\lambda|}{|D|}Ent(D_t^\lambda)$$

基于上述公式，我们可以得到在所有划分点集合中，使信息增益最大的划分点。根据此划分点便可以将属性 a 分为两个离散值。

与离散属性不同，若当前节点划分属性为连续属性，该属性还可以可以作为其后代节点的划分属性，即划分点不同。

6.决策树的缺失值处理

当数据集中有缺失值时，需要解决两个问题：如何在属性值确实的情况下进行划分属性选择？给定划分属性，若样本在该属性上的值缺失，如何对样本进行划分。C4.5和CART都可以对缺失值进行处理。这里介绍C4.5使用的处理方法。

给定样本集 D 和连续属性 a，令 $\widetilde{D}$ 表示属性 a 在样本集 D 上没有确实的样本子集。令 $\widetilde{D}^v$ 表示 $\widetilde{D}$ 在属性 a 上取值为 a^v 的样本子集， $\widetilde{D}_k$ 表示 $\widetilde{D}$ 中属于第 k 类的样本子集。假定给每一个样本都赋予一个权重 w_x ,在决策树学习的开始阶段，根节点各个样本的权重都初始化为1。

定义

$$\rho=\frac{\sum_{x\in\widetilde{D}}w_x}{\sum_{x\in D}w_x}$$

$$\tilde{p}_k=\frac{\sum_{x\in\widetilde{D}_k}w_x}{\sum_{x\in\widetilde{D}}w_x}$$

$$\tilde{r}_v=\frac{\sum_{x\in\widetilde{D}^v}w_x}{\sum_{x\in\widetilde{D}}w_x}$$

有公式可以看出，对于属性 a ， ρ 表示无缺失值样本所占的比例， $\tilde{p}_k$ 表示无缺失值样本中第 k 类样本所占的比例， $\tilde{r}_v$ 表示无缺失值样本中在属性 a 上取值 a^v 的样本所占的比例。

基于上述定义，信息增益的计算公式为：

$$Gain(D,a)=\rho\times Gain(\widetilde{D},a)=\rho\times\left(Ent(\widetilde{D})-\sum_{v=1}^{V}Ent(\widetilde{D}^v)\right)$$

其中

$$Ent(\widetilde{D})=-\sum_{k=1}^{K}\tilde{p}_k\log_2\tilde{p}_k$$

对于第二个问题，若样本 x 在划分属性 a 上的取值已知，则将 x 划分到对应的子节点，且样本权重在子节点中保持为 w_x 。若样本 x 在划分属性 a 上的取值未知，则将 x 同时划分到所有子节点中，且样本权值在子节点中调整为 $\tilde{r}_v w_x$ 。直观来说，就是让节点以不同概率划分到子节点中去。