决策树

1.决策树算法原理

2.CART回归树

3.决策树的剪枝

1.决策树算法原理

1.1 信息论基础

- 熵： 熵度量了事物的不确定性，越不确定的事物，熵就越大。随机变量X的熵的表达式如下：

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

举个例子，比如X有2个可能的取值，而这两个取值各为1/2时X的熵最大，此时X具有最大的不确定性。值为

$$H(X)=-(\frac{1}{2}log\frac{1}{2}+\frac{1}{2}log\frac{1}{2})=log2.$$

- 联合熵： 熟悉了一个变量X的熵，很容易推广到多个变量的联合熵，这里给出两个变量X和Y的联合熵表达式：

$$H(X,Y)=-\sum_{i=1}^{n}p(x_i,y_i)logp(x_i,y_i)$$

- 条件熵： 条件熵H(X|Y)度量了我们在知道Y以后X剩下的不确定性
- H(X)-H(X|Y)度量了X在知道Y以后不确定性减少程度。这个度量我们在信息论中称为**互信息**，记为$I(X,Y)$。在决策树ID3算法中叫做信息增益。信息增益大，则越适合用来分类。

1.2 决策树ID3算法

- 思想：用信息增益最大的特征来建立决策树的当前节点。
- 例子：15个样本D，其中9个输出为1,6个输出为0。样本中有特征A，取值为A1，A2，A3。在取值为A1的样本的输出中，有3个输出为1，2个输出为0，取值为A2的样本中，2个输出为1，3个输出为0，在取值为A3的样本中，4个输出为1，1个输出为0。
  - 样本D的熵为：$H(D)=-(\frac{9}{15}log_2\frac{9}{15}+\frac{6}{15}log_2\frac{6}{15})=0.971$
  - 样本D在特征下的条件熵为：$H(D|A)=\frac{5}{15}H(D1)+\frac{5}{15}H(D2)+\frac{5}{15}H(D3)$
    $=-\frac{5}{15}(\frac{3}{5}log_2\frac{3}{5}+\frac{2}{5}log_2\frac{2}{5})-\frac{5}{15}(\frac{2}{5}log_2\frac{2}{5}+\frac{3}{5}log_2\frac{3}{5})-\frac{5}{15}(\frac{4}{5}log_2\frac{4}{5}+\frac{1}{5}log_2\frac{1}{5})=0.888$
  - 对应的信息增益为：$I(D,A)=H(D)-H(D|A)=0.083$

    这里只有一个特征A，当有多个特征时，选择信息增益最大的特征作为节点特征，由该特征的不同取值最为子节点，以此类推，构建决策树。
- 具体算法过程：
  - 初始化信息增益的阈值ε
  - 判断样本是否为同一类输出Di，如果是则返回单节点树T。标记类别为Di

- 判断特征是否为空，如果是则返回单节点树T，标记类别为样本中输出类别D实例数最多的类别
- 计算A中的各个特征（一共n个）对输出D的信息增益，选择信息增益最大的特征Ag
- 如果Ag的信息增益小于阈值ε，则返回单节点树T，标记类别为样本中输出类别D实例数最多的类别。
- 否则，按特征Ag的不同值Agi将对应的样本输出D分成不同的类别Di。每个类别产生一个子节点。对应特征值为Agi。返回增加了节点的数T
- 对于所有子节点，令D = Di ，A = A- {Ag}递归调用2-6步，得到子树Ti并返回。

- **缺点与不足：**
  - ID3没考虑连续特征，比如长度，密度都是连续值，无法在ID3运用。（连续值处理）
  - ID3用信息增益作为标准容易偏向取值较多的特征。在相同条件下，取值多的特征信息增益大。（信息增益比）
  - ID3算法没考虑缺失值问题。（缺失值处理）
  - 没考虑过拟合问题。

1.3 决策树C4.5算法改进ID3

针对ID3算法4个主要的不足，一是不能处理连续特征，二是用信息增益比，最后是缺失值处理的问题和过拟合问题。

1.3.1 连续值处理

C4.5思路：将连续的特征离散化。

1. 将m个连续样本从小到大排列。(比如 m 个样本的连续特征A有 m 个，从小到大排列 a1，a2，……am)
2. 取相邻两样本值的平均数，会得m-1个划分点。(其中第i个划分点Ti表示为：$T_i = \frac{a_i + a_{i+1}}{2}$。)
3. 对于这m-1个点，分别计算以该点作为二元分类点时的信息增益。选择信息增益最大的点作为该连续特征的二元离散分类点。(比如取到的增益最大的点为at，则小于at的值为类别1，大于at的值为类别2，这样就做到了连续特征的离散化。注意的是，与离散属性不同，如果当前节点为连续属性，则该属性后面还可以参与子节点的产生选择过程。)
4. 用信息增益比选择最佳划分。

【注意】：选择连续特征的分类点采用信息增益这个指标，因为若采用增益比，影响分裂点信息度量准确性，若某分界点恰好将连续特征分成数目相等的两部分时其抑制作用最大，而选择属性的时候才使用增益比，这个指标能选择出最佳分类特征。

1.3.2 信息增益比

引入一个信息增益比 IR(Y, X)，它是信息增益与特征熵（也称分裂信息）的比。表达式：

$$I_R(D,A) = \frac{I(A,D)}{H_A(D)}$$

其中D为样本特征输出的集合，A为样本特征，对于特征熵 HA(D)，表达式：

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

其中n为特征A的类别数，|Di|为特征A的第i个取值对应的样本个数。|D|为样本个数。

特征数越多的特征对应的特征熵越大，它作为分母，可以校正信息增益容易偏向于取值较多的特征的问题。

- **1.3.3 缺失值问题**

  主要需要解决的是两个问题，一是在样本某些特征缺失的情况下选择划分的属性，二是选定了划分属性，对于在该属性上缺失特征的样本的处理。

  - 对于第一个子问题，对于某一个有缺失特征值的特征A。C4.5的思路是将数据分成两部分，对每个样本设置一个权重（初始可以都为1），然后划分数据，一部分是有特征值A的数据D1，另一部分是没有特征A的数据D2. 然后对于没有缺失特征A的数据集D1来和对应的A特征的各个特征值一起计算加权重后的信息增益比，最后乘上一个系数，这个系数是无特征A缺失的样本加权后所占加权总样本的比例。
  - 对于第二个子问题，可以将缺失特征的样本同时划分入所有的子节点，不过将该样本的权重按各个子节点样本的数量比例来分配。比如缺失特征A的样本a之前权重为1，特征A有3个特征值A1,A2,A3。 3个特征值对应的无缺失A特征的样本个数为2,3,4.则a同时划分入A1，A2，A3。对应权重调节为2/9,3/9, 4/9。

1.4决策树C4.5算法的不足与改进

- 决策树算法非常容易过拟合，因此对于生成的决策树要进行剪枝。C4.5的剪枝方法有优化的空间。思路主要是两种，一种是预剪枝，即在生成决策树的时候就决定是否剪枝。另一个是后剪枝，即先生成决策树，再通过交叉验证来剪枝。后面在下篇讲CART树的时候我们会专门讲决策树的减枝思路，主要采用的是后剪枝加上交叉验证选择最合适的决策树。
- C4.5生成的是多叉树，在计算机中二叉树模型会比多叉树运算效率高。多叉树改二叉树，可以提高效率。
- C4.5只能用于分类。
- C4.5由于使用了熵模型，里面有大量的耗时的对数运算,如果是连续值还有大量的排序运算。如果能够加以模型简化减少运算强度但又不牺牲太多准确性的话，因此用基尼系数代替熵模型。