决策树

决策树学习通常包括三个步骤：特征选择，决策树的生成，剪枝。

ID3，C4.5，CART的区别

在于树节点分裂准则的不同。

1. *熵(entropy)*

假设X是一个取有限个值的离散随机变量，其概率分布为

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

则该随机变量的熵的表示为

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

2. *条件熵(conditional entropy)*

设有随机变量(X,Y)，其联合概率分布为

$$P(X=x_i,Y=y_i)=p_{ij}, i=1,2,\dots,n; j=1,2...,m$$

给定X的条件下Y的条件熵为

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

条件熵是熵的加权平均。

3. *信息增益(information gain)*

特征X对训练目标Y的信息增益定义

$$g(Y,X)=H(Y)-H(Y|X)$$

使用信息增益存在偏向于选择去值较多的特征的问题，可以使用信息增益比较正。
思考：一般在做模型训练之前会把多分类变量转成二分类的，是不是就能解决这个问题了。

4. *信息增益比(information gain ratio)*

在信息增益的基础熵除数据集关于特征X的值的熵：

$$g_R(Y,X)=\frac{g(Y,X)}{H_X(Y)}$$

$H_X(Y)=-\sum_{i=1}^{n}\frac{D_i}{D}log_2\frac{D_i}{D}$，$n$为特征X的取值个数。

5. *ID3算法*

从根节点开始，对节点计算所有可能的特征的**信息增益**，选择最大的那个做分裂节点，对子节点递归调用以上方法，直到所有特征的信息增益均很小或者没有特征可以选择。

ID3树使用信息增益，且只有树的生成，容易过拟合。叶子节点的结果是选实例数最大的类做结果，所以只能用在分类上。

6. 剪枝

1. 梯度提升树（GBDT）

原理： 每次在前一棵决策树的残差上构建下一棵决策树，将所有树的结果加起来