Decision Tree

目录

1 、决策树 3
1.1 分裂特征选择 3
1.2 树生长 4

1、决策树

CART (Classification and Regression Trees) 可以应用于分类和回归的建模，属于非参数的有监督模型。

开始的时候，全部样本都在一个叶子节点上。然后叶子节点不断通过二分裂，逐渐生成一棵树。

优点：

简单易懂和解释，可视化,

可以处理二分类/多分类和回归问题.

缺点：

当创建出过于复杂的树时，容易过拟合，

模型的稳定性不好，

单一决策树的拟合效果一般。

1.1 分裂特征选择

节点上的数据表示为Q，分裂的备选集是$\theta = (j, t_m)$, j表示第 j个特征，t_m表示分裂阈值。基于θ把数据集划分为$Q_{left}(\theta)$和$Q_{right}(\theta)$：

$$Q_{left}(\theta) = (x, y)|x_j \leq t_m$$

$$Q_{right}(\theta) = Q \backslash Q_{left}(\theta)$$

函数H是计算节点纯净程度的指标，分类问题常用的指标有 gini 指数，交叉熵；回归问题常用的指标有 mse，mae。节点越纯净表示分裂效果越好。

$$G(Q, \theta) = \frac{n_{left}}{N_m} H\left(Q_{left}(\theta)\right) + \frac{n_{right}}{N_m} H\left(Q_{right}(\theta)\right)$$

$$\theta^* = argmin_\theta G(Q, \theta)$$

计算案例

buyer_user_id	diff_buy_count	avg_per_discount	all_per_discount	label
路人甲	0.33	0.048	0.06	0
路人乙	1	0.2	0.204	1
路人丙	0.71	0.23	0.31	0
路人丁	0.83	0.018	0.073	0

路人甲 2	0.5	0.43	0.704	1

基于上述样本，要构建一个树模型。我们需要计算用哪个特征的哪个阈值，解决办法是遍历所有的特征和阈值，分别计算得到指标。分类树的常用指标有 gini 指数和交叉熵，以 gini 指数为例子，

公式为$H(Q)=\sum_k p_k(1-p_k)$（那么原始样本的 gini 指数是 2/5*(1-2/5)+3/5*(1-3/5)=0.48）。

以第一个特征 diff_buy_count 为例，计算该特征的最佳阈值，计算过程如下：

第一步：

原始数据为[0.33,1,0.71,0.83,0.5],[0,1,0,0,1]

将数据从小到大排序为[0.33,0.5,0.71,0.83,1],

对应的 label [0,1,0,0,1]

第二步：

遍历所有可能的分裂方式：

阈值为 0.33，即小于等于 0.33 的为分裂后的左节点，大于 0.33 为分裂后的右节点。

1/5*(1*0+0*1)+4/5*(1/2*1/2+1/2+1/2)=0+2/5=2/5

阈值为 0.5

2/5*(1/2*1/2+1/2+1/2)+3/5*(1/3*2/3+2/3*1/3)=1/5+4/15=7/15

阈值为 0.71

3/5*(1/3*2/3+2/3*1/3)+2/5*(1/2*1/2+1/2+1/2)= 1/5+4/15=7/15

阈值为 0.83

4/5*(1/4*3/4+3/4*1/4)+1/5*0=4/5*3/8=3/10

[2/5,7/15,7/15,3/10]的最小值为 3/10，对应的阈值是 0.83。

一个特征的最佳分裂点计算完成。遍历所有的特征就能得到每个特征的最佳分裂点，特征间再比较 gini 指数的值，就能得到最佳分裂特征和分裂阈值。

1.2 树生长

通过前面我们已经知道树的节点是如何分裂的，那么我只要再知道树的生长是如何停止的，这样我输入一些参数，树模型就能够生长成我们预期的形态。

1.最大叶子节点数 max_leaf_nodes

叶子节点通俗来讲就是不再分裂的节点

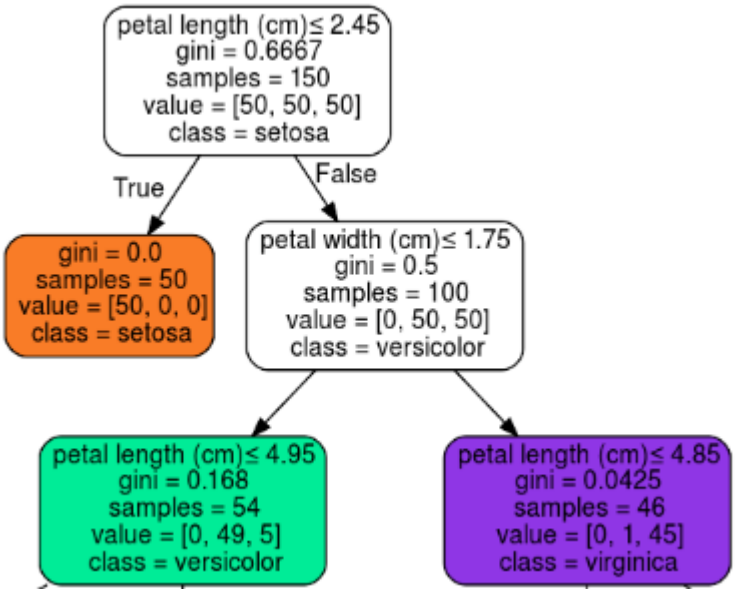
2.最小分裂节点样本数 min_samples_split

当节点的数小于最小分裂节点样本数就不继续分裂。

3.最小叶子样本数 min_samples_leaf

叶子所需要的最小样本数。

4.树的深度 max_depth



橙色，绿色和紫色的属于叶子节点，树的深度为 2。