4.树模型对特征重要性进行评估

1. 随机森林（RF）简介

随机森林的算法可以用如下几个步骤概括：

1）用有抽样放回的方法（bootstrap）从样本集中选取n个样本作为一个训练集

2）用抽样得到的样本集生成一棵决策树。在生成的每一个结点：

2.1 随机不重复地选择d个特征

2.2 利用这d个特征分别对样本集进行划分，找到最佳的划分特征（可用基尼系数、增益率或者信息增益判别）

3）重复步骤1到步骤2共k次，**k即为随机森林中决策树的个数。**

4）用训练得到的随机森林对测试样本进行预测，并用票选法决定预测的结果。

2.特征重要性评估

计算每个特征在随机森林中的每颗树上做了多大的贡献，然后取个平均值，最后比一比特征之间的贡献大小。

好了，那么这个贡献是怎么一个说法呢？通常可以用基尼指数（Gini index）或者袋外数据（OOB）错误率作为评价指标来衡量。

2.1 基于基尼系数

我们将变量重要性评分（variable importance measures）用VIM来表示，将Gini指数用GI来表示，假设有J个特征X_1，X_2，X_3，…，X_J，I棵决策树，C个类别，现在要计算出每个特征X_j的Gini指数评分$VIM_j^{(Gini)}$，亦即第j个特征在RF所有决策树中节点分裂不纯度的平均改变量。

第i棵树节点q的Gini指数的计算公式为

$$GI_q^{(i)} = \sum_{c=1}^{|C|}\sum_{c'\neq c} p_{qc}^{(i)} p_{qc'}^{(i)} = 1 - \sum_{c=1}^{|C|}(p_{qc}^{(i)})^2 \tag{3-1}$$

Pqc的含义：节点q 中类别 c 所占的比例。

特征X_j在第i棵树节点q的重要性，即节点q分枝前后的Gini指数变化量为

$$VIM_{jq}^{(Gini)(i)} = GI_q^{(i)} - GI_l^{(i)} - GI_r^{(i)} \tag{3-2}$$

其中，$GI_l^{(i)}$和$GI_r^{(i)}$分别表示分枝后两个新节点的Gini指数。

如果，特征X_j在决策树i中出现的节点为集合Q，那么X_j在第i颗树的重要性为

$$VIM_j^{(Gini)(i)} = \sum_{q \in Q} VIM_{jq}^{(Gini)(i)} \tag{3-3}$$

假设RF 中共有I颗树，那么

$$VIM_j^{(Gini)} = \sum_{i=1}^{I} VIM_j^{(Gini)(i)} \tag{3-4}$$

最后，把所有求得的重要性评分做一个归一化处理即可。

$$VIM_j^{(Gini)} = \frac{VIM_j^{(Gini)}}{\sum_{j'=1}^{J} VIM_{j'}^{(Gini)}} \tag{3-5}$$