3.随机森林

3.1随机森林原理

1）从样本集中有放回随机采样选出m个样本；

2）从所有特征中随机选择k个特征，对选出的样本利用这些特征建立决策树

3）重复以上两步次，即生成棵决策树，形成随机森林；

对于回归任务，就把这t 棵树的结果做平均；要是分类任务就使用投票法。

3.2随机森林优缺点

优点：

1） 每棵树随机选择样本并随机选择特征，使得具有很好的抗噪能力，性能稳定；

2） 能处理很高维度的数据，并且不用做特征选择,适用于数据集中存在大量位置特征；

3） 适合并行计算；

4） 能够做回归与分类两种任务。

缺点：

1） 参数较复杂；（需要训练多棵子树）

2） 对小量数据集和低维数据集的分类不一定可以得到很好的效果

3.3 特征选择

1.袋外误差（比较哪一个更重要）

1）首先要了解什么是袋外误差？

袋外的概念就是我们一次对样本进行采样，假设总共有 N 个样本，一次采样只采集 M 个样本，那么就有N - M 个样本没有被采集到，**这些样本就是用来作为测试样本后期衡量决策树的好坏，当然也拿来衡量特征的好坏。**

2）OOB 误差究竟怎么用？

首先，计算每一棵树的袋外数据的预测值与真实值之间的误差和，记为ERROR1；

其次，对每一个特征随机添加一定的噪音，然后再次计算对应的误差，记为ERROR2；

最后，计算ERROR1 与 ERROR2 之间的差值，差值越大，说明当前特征越重要。

（误差越大，说明该特征对模型影响越大，故其越重要）

2.基尼系数（计算出具体数据）

在决策树中 cart 树就是使用基尼系数来进行节点划分，在每一个节点划分的时候，计算每一个特征的基尼系数，选择基尼系数较小的特征，**基尼系数越小，反应得到的结果集数据越纯，也就是划分的效果比较好。**

计算划分前集合的基尼系数 和 划分后每个子集的基尼系数和 的差值作为当前特征的重要性，如果差值越大说明当前的特征重要性越高。

计算每一个特征在每一棵树的重要性，然后取加权平均得到最终的特征重要性评估