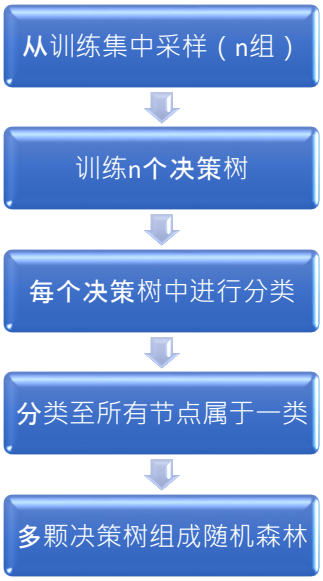
一、 随机森林算法简介

随机森林是一种比较新的机器学习模型。经典的机器学习模型是神经网络，有半个多世纪的历史了。神经网络预测精确，但是计算量很大。上世纪八十年代 Breiman 等人发明分类树的算法，通过反复二分数据进行分类或回归，计算量大大降低。2001 年 Breiman 把分类树组合成随机森林，即在变量（列）的使用和数据（行）的使用上进行随机化，生成很多分类树，再汇总分类树的结果。随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元公线性不敏感，结果对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用，被誉为当前最好的算法之一。

随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。因此随机森林算法实际上可以看做是组合在一起的多个决策树。关于决策树的原理、构造等参考决策树算法介绍的文章。

二、 随机森林的构造



1. 从原始训练集中使用 Bootstraping 方法随机有放回地采样选出 m 个样本，共进行 n_tree 次采样，生成 n_tree 个训练集
2. 对于 n_tree 个训练集，我们分别训练 n_tree 个决策树模型
3. 对于单个决策树模型，假设训练样本特征的个数为 n，那么每次分裂时根据信息增益/信息增益比/基尼指数选择最好的特征进行分裂
4. 每棵树都一直这样分裂下去，直到该节点的所有训练样例都属于同一类。在决策树的分裂过程中不需要剪枝
5. 将生成的多棵决策树组成随机森林。对于分类问题，按多棵树分类器投票决定最终分类结果；对于回归问题，由多棵树预测值的均值决定最终预测结果

三、 随机森林算法的优缺点

优点：

- 在数据集上表现良好，两个随机性的引入，使得随机森林不容易陷入过拟合

- 在当前的很多数据集上，相对其他算法有着很大的优势，两个随机性的引入，使得随机森林具有很好的抗噪声能力
- 它能够处理很高维度的数据，并且不用做特征选择，对数据集的适应能力强：既能处理离散型数据，也能处理连续型数据，数据集无需规范化
- 可生成一个 Proximities=P_{ij}矩阵，用于度量样本之间的相似性：P_{ij}=a_{ij}/N, a_{ij}表示样本 i 和 j 出现在随机森林中同一个叶子结点的次数，N 随机森林中树的颗数
- 在创建随机森林的时候，对 generlization error 使用的是无偏估计
- 训练速度快，可以得到变量重要性排序（两种：基于 OOB 误分率的增加量和基于分裂时的 GINI 下降量
- 在训练过程中，能够检测到 feature 间的互相影响
- 容易做成并行化方法
- 实现比较简单

缺点

- 当随机森林中的决策树个数很多时，训练时需要的空间和时间会较大
- 随机森林模型可解释性不够好