随机森林原理介绍

参考如下资料：

[随机森林概述]

周志华老师《机器学习》书中对应章节

随机森林（Random Forest，RF）是对基本的Bagging算法的一个扩展变体。简单来说，**RF在以决策树为基学习器构建的Bagging集成的基础上，进一步在决策树的训练过程中引入了随机属性选择。**

RF是多颗决策树分类器的集成。用Bootstrap抽样得到各训练子集后，对于每个训练子集，分别训练一颗决策树。同时与传统决策树不同的是，在RF中，对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含k个属性的子集，然后再从这些子集中选择一个最优属性用于划分。

这里的参数k控制了随机性的引入程度。假设当前结点的全部属性数目为d，当$k = d$时，基决策树的构建与传统决策树相同；若$k = 1$，则是随机选择一个属性进行划分；一般情况下，推荐值为$k = \log_2 d$。

- 计算变量的重要性：

随机森林有一个特点，可以在训练过程中输出变量的重要性，即哪个特征分量对分类更有用。实现的方法是置换法。它的原理是，如果某个特征分量对分类很重要，那么改变样本的该特征分量的值，样本的预测结果就容易出现错误。也就是说这个特征值对分类结果很敏感。反之，如果一个特征对分类不重要，随便改变它对分类结果没多大影响。

对于分类问题，训练某决策树时在包外样本集中随机挑选两个样本，如果要计算某一变量的重要性，则置换这两个样本的这个特征值。统计置换前和置换后的分类准确率。变量重要性的计算公式为：

$$\nu = \frac{\text{置换之前正确分类的样本数} - \text{置换之后正确分类的样本数}}{OOB\text{样本总数}}$$

这反应的是置换前后的分类准确率变化值。

上面定义的是单棵决策树的变量重要性，计算出每棵树的变量重要性之后，对该值取平均就得到随机森林的变量重要性。计算出每个变量的重要性之后，将该值归一化得到最终的重要性值。