链滴

机器学习方法——随机森林

作者：yudake

原文链接：https://ld246.com/article/1519913530581

来源网站：链滴

<h2 id="定义">定义</h2>
随机森林建立多个决策树并将它们合并在一起以获得更准确和稳定的预测。
<h2 id="实现方法">实现方法</h2>
随机森林有两种实现方法：

数据的随机化：就是每一个树构建只利用随机抽取的部分数据，使得随机森林的决策树更普遍化

有放回的采样可以保证不同子集的数量级一样（不同子集/同一子集之间的元素可以重复）。

待选特征的随机化：

在所有特征里随机选取部分特征；
在随机选取的特征里选取最优特征（由树的构建完成）。

<h2 id="特征的重要性">特征的重要性</h2>
随机森林可以衡量每个特征对预测的相对重要性。一个特征被越多树使用说明这个特征越重要，过查看特征的重要性，可以决定要放弃哪些特征（防止过拟合）。
<h2 id="随机森林和决策树的区别">随机森林和决策树的区别</h2>

决策树的特征选择使用信息增益或者基尼指数计算，随机森林是随机选择；
决策树可能会过深，导致过拟合，随机森林可以防止大部分过拟合，因为创建的都是小树。

<h2 id="随机森林的参数">随机森林的参数</h2>

n_estimators：树的数量，越多越稳定，但是速度回变慢；
max_features：每个树可以使用的最多特征，应该是为了防止过拟合；
max_depth：每个树的最大深度，应该也是为了防止过拟合；
min_samples_split：一个结点的分支想分割需要的最少样本，防止离群点对模型产生影响；

min_sample_leaf：一个叶子上需要的最小样本数，根上一个差不多。
其他的也都差不多...

<h2 id="优点">优点</h2>

既可以用于分类，也可以用于回归；
容易查看特征的重要性；
易于实现，超参数较少，超参数最佳取值范围比较固定；
不容易过拟合；

<h2 id="缺点">缺点</h2>

树的数量太多时，算法变慢；
无法描述数据中的关系。
