Random Forest

集成学习

集成学习(Ensemble learning)：训练一系列个体学习器，再通过结合策略将它们集成起来，形成更强的学习器。

个体学习器

如果集成中只包含**同种类型**的个体学习器，叫做同质集成，个体学习器称为**基学习器**

> e.g. 随机森林全是决策树集成

如果集成中包含**不同类型**的个体学习器，叫做异质集成，个体学习器称为**组件学习器**

> e.g. 同时包含决策树和神经网络进行集成

集成学习核心问题

1. 使用什么样的个体学习器
    - 个体学习器不能太弱，需要一定**准确性**
    - 个体学习器之间要有**多样性**
2. 如何选择合适的结合策略构建强学习器
    - 并行组合方式，e.g. 随机森林
    - 传统组合方式，e.g. boosting树模型

Bagging

Bagging是并行式集成学习的代表。

自助采样法(Bootstrap Sampling)：给定包含m个样本的数据集，先随机取出一个样本放入采样集中，再把样本放回初始数据集，使得下次采样时该样本仍有可能被选中。(也就是有放回的均匀抽样)上述过程重复m轮，可以得到m个样本的采样集。

Bagging(Bootstrap aggregating的缩写)：将自助采样法重复T次，采样出T个含m个训练样本的采样集，然后基于每个采样集训练出一个基学习器，将这些基学习器进行结合。

Random Forest

随机森林(RF)：基于树模型的Bagging的优化版本，使用了CART决策树作为基学习器。

算法过程

1. 输入为样本集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$
2. 对于$t=1,2,\cdots,T$
    - 对训练集进行第t次随机采样，共采集m次，得到包含m个样本的采样集D_T
    - 用采样集D_T训练第T个决策树模型$G_T(x)$
3. 对于分类场景，T个基模型投出最多票数的类别为最终类别

特点

1. 随机
    - **样本扰动**：直接基于自助采样法，使得初始训练集中约63.2%的样本出现在一个采样集中，带来数据集差异化
    - **属性扰动**：对基模型每个结点的特征属性集合中随机选择k个属性，然后从这k个属性中选择一个最优属性进行划分
2. 集成
    - 根据多个差异化采样集，训练得到多个差异化决策树，采用**简单投票/平均法**提高模型稳定性、泛化能力

优缺点

1. 优点
    - 适用于高维稠密型数据，不用降维，不用做特征选择
    - 借助模型构建组合特征
    - 并行集成，有效控制过拟合
2. 缺点
    - 噪声过大的分类和回归数据集上可能还是会过拟合
    - 模型解释复杂

影响参数与调优

核心参数

1. 生成单颗决策树的特征数 `max_feature`
    - 增加能提高单个决策树的性能，但是降低了树与树之间的差异性
    - 太小会影响单棵树的性能，进而影响整体集成效果
2. 决策树的棵树 `n_estimators`
    - 较多子树会让模型有很好的稳定性和泛化能力，同时也会让模型学习速度变慢
3. 树深 `max_depth`
    - 太大树深可能会过度学习，导致过拟合
    - 如果模型样本多特征多，则限制树深提高模型泛化能力

参数调优

1. RF划分考虑最大特征数 `max_feature` ，通常选择总数的$[0.5, 0.9]$
2. 决策树棵树 `n_estimators` ，通常设置为> 50的取值
3. 决策树最大深度 `max_depth` ，通常设置在$4 - 12$
4. 内部结点再划分所需最小样本数 `min_samples_split` ，样本不大不需要调整，样本太大可以设置为$16, 32, 64$等
5. 叶子结点最少样本数 `min_samples_leaf` ，通常设为> 1