**RandomForest特征重要性**

RF可以输出两种 feature\_importance，分别是Variable importance和Gini importance。

<https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>

**Variable importance**

选定一个特征M，在所有OOB样本的特征 M上人为添加噪声，再测试模型在OOB上的判断精确率，精确率相比没有噪声时下降了多少，就表示该特征有多重要。

假如一个特征对数据分类很重要，那么一旦这个特征的数据不再准确，对测试结果会造成较大的影响，而那些不重要的特征，即使受到噪声干扰，对测试结果也没什么影响。这就是 Variable importance 方法的朴素思想。

**Gini importance**

选定一个特征M，统计RF的每一棵树中，由M形成的分支节点的Gini指数下降程度（或不纯度下降程度）之和，这就是M的importance。

从sklearn的官方文档对feature\_importances\_参数的描述来看，sklearn应当是使用了Gini importance对特征进行排序，同时sklearn把所有的Gini importance以sum的方式做了归一化，得到了最终的feature\_importances\_输出参数

**Adaboost特征重要性**

feature\_importances\_:特征重要性，这个参数使用前提是基分类器也支持这个属性。

基分类器为决策树，计算方式与RF相同。

**GBDT特征重要性**

Friedman在GBM的论文中提出的方法：

特征*j*的全局重要度通过特征*j*在单颗树中的重要度的平均值来衡量：

其中，M是树的数量。特征*j*在单颗树中的重要度的如下：

其中，为树的叶子节点数量，即为树的非叶子节点数量（构建的树都是具有左右孩子的二叉树），是和节点相关联的特征，是节点分裂之后平方损失的减少值。

**XGBoost特征重要性**

xgb.plot\_importance有一个参数为importance\_type ，表示特征重要性的计算方式

计算方式有“weight”,“gain” or “cover”三种，默认为“weight”

”weight” is the number of times a feature appears in a tree

# “weight”为一个特征在一棵树中出现的次数

”gain” is the average gain of splits which use the feature

“gain”为使用某个特征分裂的平均增益

”cover” is the average coverage of splits which use the feature where coverage is defined as the number of samples affected by the split

“cover”为使用某个特征分裂的平均覆盖范围，覆盖范围定义为这次分裂影响的样本数量