XGBoost:

可以在分类和回归问题使用

构建二叉树，通过梯度提升

0. 给出initial prediction（分类问题一般给0.5）

1. 选择特征作为分裂标准，计算similarity和gain值，选出gain值最大的特征

2. 回归问题中，每个叶节点的residual的平均值作为输出值。分类问题中，

3. 循环1,2,3步直到树的数量达到预设值

剪枝：从最深层开始计算gain - gamma，如果结果为正数，则不剪枝，反之进行剪枝（视具体情况，如子树gain-gamma>0,当根节点gain-gamma<0,就不剪枝）

计算similarity公式中的λ实现正则化操作，>0时可以解决过拟合

XGB regression和classification中的不同：

计算similarity的公式不同

classification中有Cover值，即每个节点中样本的最小个数

预测 initial guess + learning rate * 样本在每个树中的输出值

AdaBoost:

可以在分类和回归问题使用，只要用于分类

构建树桩（stump）,通过提高误分类样本的权重来进行拟合数据

0. 给每个样本赋予初始权重1/总样本数（每次更新权重后要实现标准化， 即所有样本权重之和为1）

1. 计算每个特征作为分裂准则的Gini值，选择最小的作为分裂特征

2. 计算每个树桩的权重：Amount of say = (1/2)*log((1-total erro)/total error)， total error为误分类样本的权重之和

Amount of say 越大，说明该Stump分类效果越好

3. 更新样本的权重

（1）提高误分类样本的权重： new weight = sample weight*e^(该树桩的amount of say)

（2）降低正确分类样本的权重： new weight = sample weight*e^(-该树桩的amount of say)

（3）归一化权重
4. 在0到1中随机选取一个数a，看a落在哪里，选取对应的样本加入新的数据集new_D中，重复，选取N个数据（跟原训练集大小一样）
5. 回到0，初始权重仍然是1/总样本数

预测时，看样本在每个stump中的分类情况，结合当前stump的权重，选择总权重大的类别

GBDT for regression：
0.用所有样本的均值作为initial guess（回归问题）
1. 算每个样本和guess的误差，构造决策树（涉及到特征分裂选择问题等），对样本对应的误差进行分类（即叶节点中是每个样本的误差）
2. 每个叶节点的值用其中样本误差的均值代替，作为输出值
3. 更新样本的目标值=initial guess + learning rate * 样本在每个树中的输出值
4. 继续构造tree，但是计算误差时使用新的目标值

回归问题预测： y = initial guess + learningrate * 样本在每个树中的输出

GBDT for classification（二分类问题）：
0. 用log(odds)计算initial prediction，(odds = 事件A发生频率/1-事件A发生频率)，，则假设所有样本都分类到A，
再把log(odds)放入到logistic函数中：P(A) = e^(log(odds))/(1+e^(log(odds)))，求得initial prediction，若P(A)>0.5，则P(A)作为initial guess,反之求另一分类的概率作为initial guess
1. 计算所有样本的residual。Residual = Observed - Predicted = **(1 - Predicted)** OR **(0 - Predicted)** 。详细例子看笔记
2. 构建决策树，跟regression问题一样，把每个样本的residual分到各个叶节点中，计算每个叶节点的输出（详细公式看笔记）
3. 更新每个样本的新概率 initial guess log(odds) + learning rate*样本在每个树中的输出值，结果为log(odds)，再把log(odds)转换为概率
4. 回到第一步重新构造新的tree

预测：
initial guess log(odds) + learning rate*样本在每个节点的输出，计算结果再通过logistic function转换为概率

Decision Tree:

1. 计算每个特征作为分裂条件的gini值，以gini值最小的特征作为分裂条件
2. 重复第1不不断构造树，若分裂后的子树的gini值比原来的大，则不构造该子树。

Random Forest:

0. 有放回抽样构造boostrapped dataset
1. 使用bootstrapped dataset构造决策树，但只用其中的某几个特征（即每棵树只包含其中的一部分特征）
2. 重复0,1, 构建random forest

新样本预测：

使用投票表决方式，看样本在这么多个树中，分到哪类的频率最高