1. **Bagging和Boosting的概念与区别**

随机森林属于集成学习（Ensemble Learning）中的**bagging算法**。在集成学习中，主要分为bagging算法和boosting算法。我们先看看这两种方法的特点和区别。

* **Bagging（套袋法）**

bagging的算法过程如下：

1. 从原始样本集中**使用Bootstraping方法随机抽取n个训练样本，共进行k轮抽取，得到k个训练集。（**k个训练集之间相互独立，元素可以有重复）
2. 对于k个训练集，我们训练k个模型（这k个模型可以根据具体问题而定，比如决策树，knn等）
3. 对于分类问题：**由投票表决产生分类结果；**对于回归问题：**由k个模型预测结果的均值作为最后预测结果。**（所有模型的重要性相同）

* **Boosting（提升法）**

boosting的算法过程如下：

1. **对于训练集中的每个样本建立权值wi，表示对每个样本的关注度。当某个样本被误分类的概率很高时，需要加大对该样本的权值。**
2. **进行迭代的过程中，每一步迭代都是一个弱分类器。我们需要用某种策略将其组合，作为最终模型。**（例如AdaBoost给每个弱分类器一个权值，将其线性组合最为最终分类器。误差越小的弱分类器，权值越大）

* **Bagging，Boosting的主要区别**

1. 样本选择上：**Bagging采用的是Bootstrap随机有放回抽样；而Boosting每一轮的训练集是不变的，改变的只是每一个样本的权重。**
2. 样本权重：**Bagging使用的是均匀取样，每个样本权重相等；Boosting根据错误率调整样本权重，错误率越大的样本权重越大。**
3. 预测函数：**Bagging所有的预测函数的权重相等；Boosting中误差越小的预测函数其权重越大。**
4. 并行计算：**Bagging各个预测函数可以并行生成；Boosting各个预测函数必须按顺序迭代生成。**
5. 泛化能力：**bagging减少了variance，而boosting减少了bias？**

下面是将决策树与这些算法框架进行结合所得到的新的算法：

1）Bagging + 决策树 = 随机森林

2）AdaBoost + 决策树 = 提升树

3）Gradient Boosting + 决策树 = GBDT (梯度下降树)

1. **随机森林**

鉴于决策树容易过拟合的缺点，随机森林采用多个决策树的投票机制来改善决策树，我们假设**随机森林使用了m棵决策树，那么就需要产生m个一定数量的样本集来训练每一棵树，如果用全样本去训练m棵决策树显然是不可取的，全样本训练忽视了局部样本的规律，对于模型的泛化能力是有害的**

产生n个样本的方法采用Bootstraping法，这是一种有放回的**抽样方法**，产生n个样本。而最终结果采用Bagging的策略来获得，即**多数投票机制**

**随机森林的生成方法**：

1. 从样本集中**通过重采样的方式产生n个样本**
2. 假设样本特征数目为a，对n个样本**选择a中的k个特征**，用建立决策树的方式获得最佳分割点
3. 重复m次，**产生m棵CART决策树(cart树可以在一定程度上解决，cart树既可以做回归，也可以做分类。)**
4. **多数投票机制来进行预测**

**随机森林特征重要性估计：**

根据节点划分前后GINI系数的变化大小来决定。

**GBDT特征重要度估计：**

特征j在单颗树中的重要度的平均值来衡量。

**Xgboost特征评分原理：**

根据结构分数的增益情况计算出来选择哪个特征的哪个分割点，某个特征的重要性，就是它在所有树中出现的次数之和。

1. **AdaBoost**

略。统计学习方法~

1. **GBDT**

略~

1. **Xgboost**

对比GBDT