PCA主成分分析：

主成分分析是一种将样本进行降维的算法，形象的说，就是用一个超平面，对样本点集合进行详细的表达，以求尽量在保持样本最大信息的情况下，将样本的维度降低，具体地，可以用矩阵可以表达为： （在这里，每个样本通过矩阵变换由p维降为m维），找到了变换的矩阵，也就找到了主成分分析的解

在这个过程中，这个超平面具有如下的两条性质：

1. 最近重构性:即样本点到这个超平面的距离应该足够近
2. 最大可分性：即样本点在这个超平面上的投影能尽可能的分开

我们从最大可分性出发进行主成分分析的原理推导，其数学推导如下：

由最大可分性知，样本点经过投影后方差应达到最大，设,使得,投影后样本点的方差可以表示为,等于,假设样本点是中心化的，则样本点的协方差矩阵,要使方差最大，只需：

用拉格朗日乘数法可得：

只需对协方差矩阵**V**进行特征值分解，将求得的特征值排序**：**,取前m个特征值构成),即为主成分分析的解。

总结来说，PCA算法的过程为以下几步：

1. 对所有样本进行中心化
2. 计算样本的协方差矩阵
3. 对协方差矩阵进行特征值分解
4. 取最大的m个特征值所对应的特征向量组成投影矩阵，即为算法的解

随机森林回归：

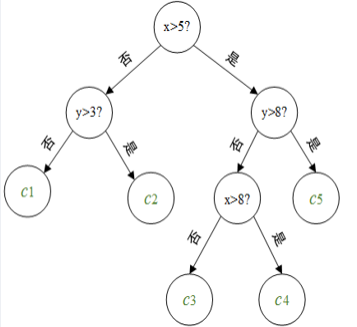
随机森林回归是利用随机森林算法解决回归问题的一种方法，而随机森林是Bagging算法的一种变体，属于机器学习中的集成学习方法，对于该学习方法，因此在介绍随机森林之前，首先来介绍一下集成学习方法和Bagging算法，从而引出随机森林算法

集成学习：通过构建并结合多个学习器来完成学习任务，它的基本思想是先产生一组“个体学习器”，再用某种策略将它们结合起来，每个个体学习器用一种学习算法从训练数据中产生。通过对多个学习器进行结合，常可获得比单一学习器显著优越的泛化性能。

Bagging算法：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，再把该样本放回初始数据集，使得下一次采样时该样本仍有可能被选中，这样经过m次随机采样操作，就可以得到含m个样本的采样集，照这样可以采样出T个含m个训练样本的采样集，基于每个采样集对这些基学习器进行结合，就是Bagging的基本流程。

而随机森林(RF)方法是Bagging算法的一个变体，以CART决策树为基学习器构建Bagging进行集成，并进一步在决策树的训练过程中引入了随机属性的选择。对基决策树的每个结点，先从该结点的属性集合中随机选择一个包含k个属性的子集，然后再从这个子集中选择一个最优属性用于划分。

在随机森林回归算法中，基学习器CART树要完成回归的任务，这就涉及到了决策树实现回归的算法，下面来介绍下决策树如何完成回归问题：

我们知道在决策树中，通过结点判断条件来选择不同分支，而回归任务中，判断条件就是具体的数值，而对于测试数据，我们只要按照特征将其归到某个单元，便得到对应的输出值。

而算法的关键便是如何找切分点以及如何确定最终的输出值，最常用的方法是利用最小二乘法寻找切分点，并通过计算单元内均值的方法确定输出，具体原理如下：

假设X和Y分别为输入和输出变量，并且Y是连续变量，给定训练数据集为

，其中为输入实例(特征向量)，n为特征个数,.

每次划分逐一考察当前集合中所有特征的所有取值，根据平方误差最小化准则选择其中最优的一个作为切分点。如对训练集中第j个特征变量\small x^{(j)}和它的取值s，作为切分变量和切分点，并定义两个区域和，为找出最优的j和s,对下式求解:

其中，为划分后两个区域内固定的输出值，方括号内的两个min意为使用的是最优的，也就是使各自区域内平方误差最小的，易知这两个最优的输出值就是各自对应区域内Y的均值，所以上式可写为

其中,, 一维空间中样本均值就是最优的输出值，至此，对CART决策树回归方法的原理介绍完毕。

在最后，我们总结一下随机森林回归的算法流程：

1. 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
2. 用采样集训练第t个决策树模型，在训练决策树模型的节点的时候，在节点上所有的样本特征中选择一部分样本特征，在这些随机选择的部分样本特征中选择一个最优的特征来做决策树的左右子树划分
3. 对于回归算法，T个弱学习器得到的回归结果进行算术平均得到的值为最终的模型输出。

该部分参考文献

[1]周志华.机器学习[M]. 北京:清华大学出版社, 2016. 171-181.

[2] https://blog.csdn.net/Albert201605/article/details/81865261