实验七：Ensemble Learning

一、 说明

- 实验采用 jupyter notebook, 请填写完代码后提交完整的 ipynb 文件
- 文件命名规则：班级_姓名_ML2018_HW7.ipynb，如计科 1701_张三_ML2018_HW7.ipynb
- 提交方式：采用在线提交至：

  http://pan.csu.edu.cn:80/invitation/bd1fea33-958f-4364-9ade-400e37f31bb3
- 实验提交截止日期：2018.12.26 24：00

二、 实验内容

集成学习（ensemble learning）通过构建并结合多个学习器来完成学习任务，常可获得比单一学习器显著优越的泛化性能。集成学习根据个体学习器间的关系分成两类，一类个体学习器间存在强依赖关系，必须串行生成学习模型，例如 AdaBoost、GBDT 等；另一类个体学习器间不存在强依赖关系，可同时生成学习模型，例如 Bagging、随机森林等。

AdaBoost 通过改变训练数据的权值来训练一组弱分类器，把这些弱分类器线性组合成为一个强分类器。GBDT 结合提升树模型和梯度提升的优点，使用新弱分类器拟合前一次迭代模型的样本余量，逐渐降低训练误差。Bagging 和随机森林利用自助采样采集 T 组训练样本集，分别训练 T 个分类器，对 T 个分类器的预测结果进行投票决定模型的最终预测结果。

本实验指导用户实现 AdaBoost 算法、GBDT 算法、Bagging 和随机森林算法。

三、 实验目标

- 熟悉并实现 AdaBoost 算法。
- 熟悉并实现 GBDT 算法。

- 熟悉并实现 Bagging 和随机森林算法。

四、 实验操作步骤

本实验需要用到的 python 环境包括

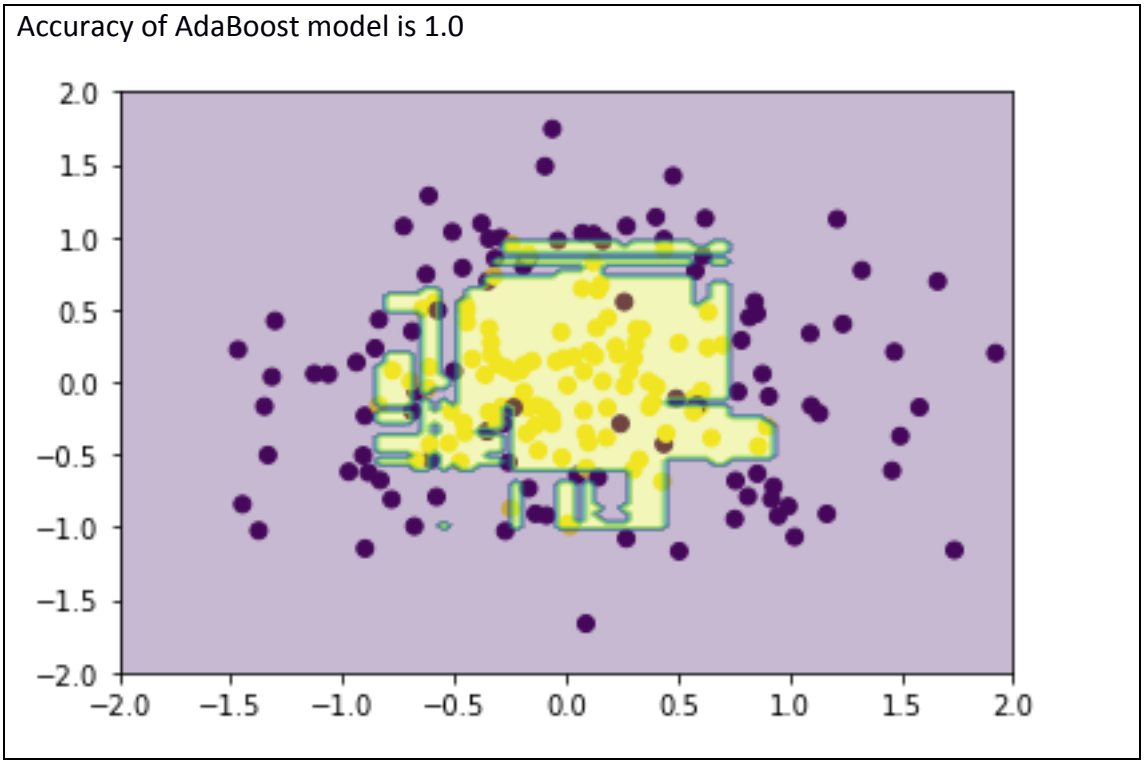
名称	版本
Python	3.6.5
Numpy	1.14.3
matplotlib	2.2.2
jupyter	1.0.0
Scikit-learn	0.19.1

1. 启动 jupyter notebook 使用

参照实验一的任务指导书，使用 jupyter notebook 打开本实验的 EnsembleLearning.ipynb 文件。

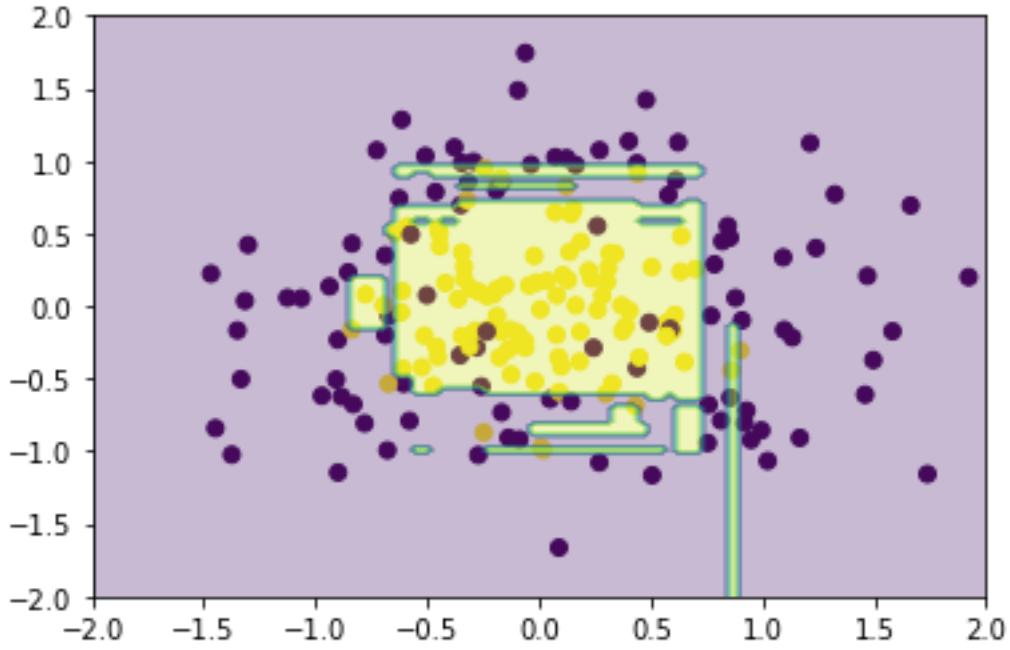
2. 完成实验任务

任务 1 实现 adaboost 函数。



任务 2 实现 gbdt_classifier 函数。

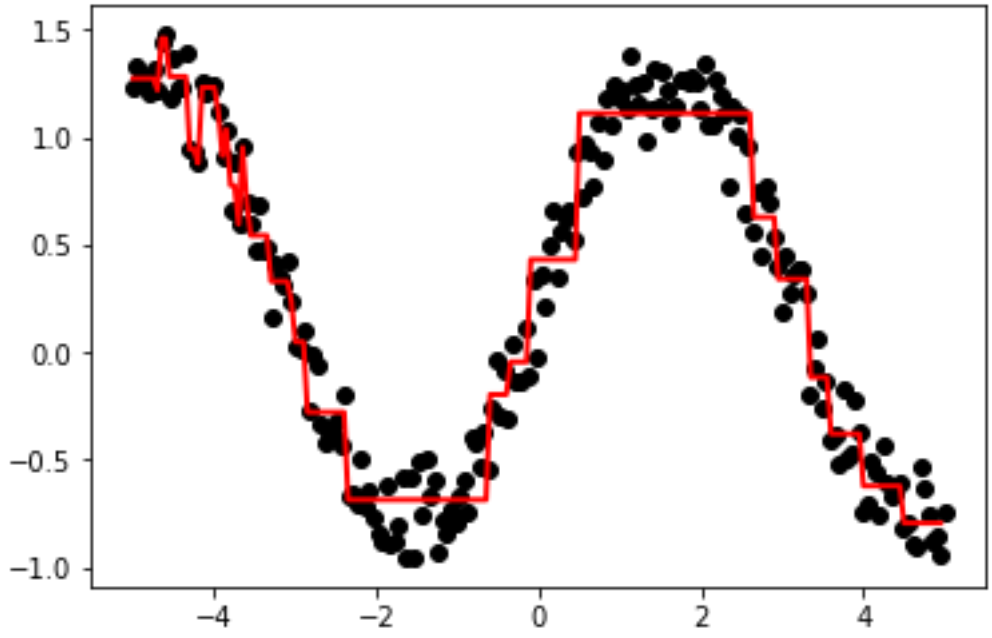
Accuracy of GBDT model is 0.955

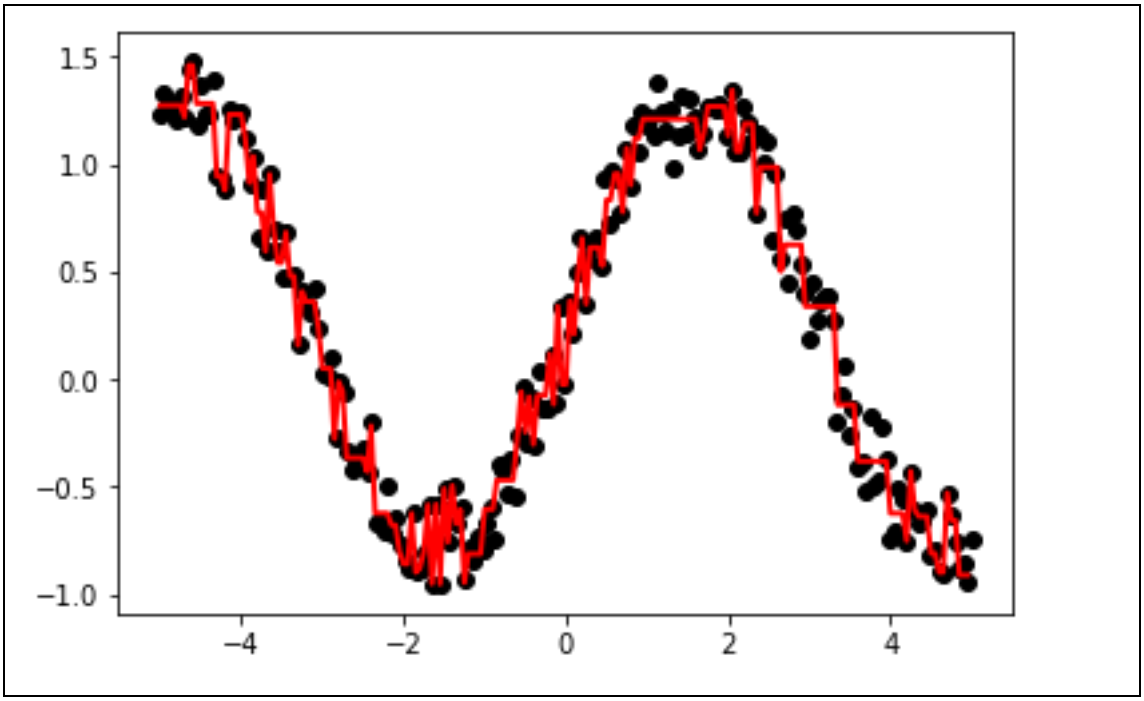


任务 3 实现 gbdt_regressor 函数。

loss of CART model is 0.008974256548067702

loss of GBDT model is 0.0022895551966146386





任务 4 实现 bagging 函数。

具体内容见 DecisionTree.jpynb 文件

输出的结果是：

```
Accuracy of bagging model in trainset is 0.7589220684632192
Accuracy of bagging model in validation set is 0.6650717703349283
```