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 _ML2019_HW7 .ipynb,如计科 1701_张三 _ML2019_HW3.ipynb
- 提交方式：采用在线提交至：http://pan.csu.edu.cn:80/invitation/89868ee4-dd48-4df8-b805-5436831e9be1
- 实验提交截至日期：2019.12.16 23:59

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。

本实验指导用户实现 AdaBoost 算法、GBDT 算法、Bagging 和随机森林算法。有所集成的弱学习器均采用 scikit-learn 的 CART 树来快速实现。最后将所实现的集成学习算法应用到 Kaggle 入门经典竞赛 Titanic 存活预测的任务中。

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

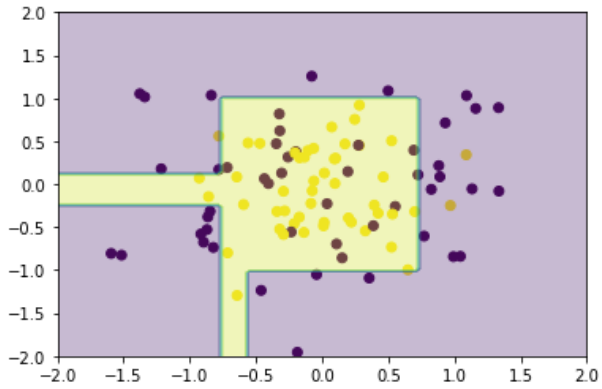
任务 1 采用 scikit-learn 实现 CART 决策树。

任务 2 实现 adaboost 函数。

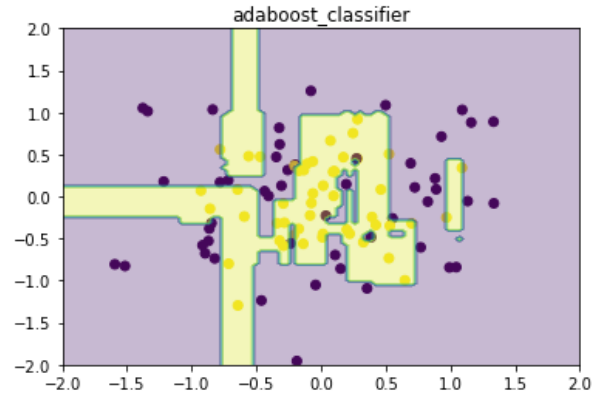
任务 3 实现 gbdt_classifier 函数。

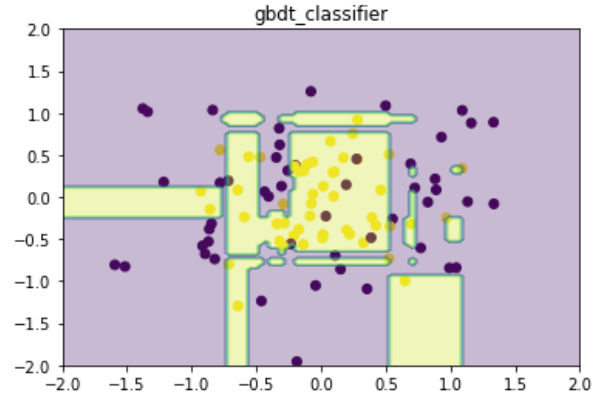
提示：对应简单二分类数据集，CART 树，adaboost 和 GBDT 实现效果类似以下输出

```
Accuracy of CART model is 0.81
<matplotlib.contour.QuadContourSet at 0x17d16b45898>
```



```
Accuracy of Adaboost model is 1.0
Accuracy of GBDT model is 0.97
```





任务 4 实现 bagging 函数。

任务 5 处理 Titanic 数据集，并采用集成学习算法训练模型。