# Ensemble Learning 实验报告

2014011326 魏星 计 43

## 一、实验设计

本次实验的代码基于 python 库 scikit-learn，需要 python 包 :sklearn，numpy 和 scipy。

大致实现思路如下：

1. 读入数据
   方式和上一次实验相同（按行读取，strip 去掉行尾换行符\n，split 进行分割）。
2. 数据处理
   先将数据分为特征和分类两个 list，分别对应 X 和 y，对 X 做标准化处理，然后用 train_test_split()函数分割出训练集和测试集
3. 进行学习
   根据参数，分别执行四个算法，即 bagging+svm，bagging+DTree，Adaboost+svm 和 Adaboost+DTree。SVM 和 DTree 由 sklearn 包中的函数实现，只需传入相应的参数即可（例如 boost 时，传入 weight 数组）
4. 打印结果

## 二、实验结果

实验得到的准确率如下

| | SVM | DTree |
|---|---|---|
| Bagging | 0.735 | 0.880 |
| Adaboost | 0.736 | 0.846 |

由此可见，DTree 比 SVM 的准确率更高。
而在集成学习对不同分类器的影响上，发现 bagging 和 boost 对 svm 的影响不大，而 bagging 对于 DTree 的提升要比 Adaboost 对于 DTree 的提升要大。

## 三、实验分析与思考

在集成学习对于分类器的影响上，我请教了其他同学，并进行了一些讨论，也比较了他人的学习结果。发现若是对数据集进行平衡，能够较好的提升 DTree 的准确率。

若是不使用 oversample，bagging 和 boost 对 DTree 的影响相差较多，而若是使用了 oversample，则相差无几。我认为可能是由于 bagging 对数据集的随机采样能够在一定程度上减轻数据集不平衡这一问题对于 DTree 学习的影响。

SVM 的准确率较低，应该是由于没有进行调参导致的。若是更加细致地调节参数，有望提升其准确率。