机器学习大作业实验说明

1. 总体要求

大作业要求分别实现以对数几率回归和决策树桩为基分类器的 AdaBoost 算法，其中对数几率回归请参考课件第六讲，决策树桩指只有一层的决策树，决策树请参考课件第四讲，AdaBoost 算法请参考讲义。

实现以上算法时不允许调用已实现好的任何机器学习模型库。

2. 输入输出格式说明

你的代码需读取 data.csv 及 targets.csv 两个文件，并输出在不同数目基分类器条件下的 10 折交叉验证的预测结果至 experiments/base#_fold#.csv，以供评测。基分类器数目取 1，5，10，100 这四种数值。输入样例，输出样例及评测代码详见提供的压缩包。

对预测结果所在文件命名格式说明如下：基分类器数目为 x 对应的预测文件为 basex_fold1.csv~basex_fold10.csv，1~10 指的是用作测试集的子集编号。每个预测文件分成两列，第一列为样例的序号（序号从 1 开始），第二列为该样例的预测标记。评测时子文件夹 experiments 会建立好，请不要在你的代码中强行建立此文件夹以免出错。如果对文件名和格式还有疑问，可以参考压缩包中提供的输出样例和评测代码。

3. 提交说明

你只需要提交你的代码文件，不需要提交数据文件。我们会在新的测试数据上运行你的入口代码 main.py 以评测,main.py 应该能将基分类器编号以及新的测试数据作为输入，并输出你使用基分类器实现的 Adaboost 算法对新的测试数据的预测标记。我们在运行你的入口代码时，基分类器编号为 0 代表对数几率回归，基分类器编号为 1 代表决策树桩。

你可以本地运行评测代码 evaluate.py 来验证自己程序的有效性，如果评测代码正常运行，应当输出四个精度（accuracy）值，分别对应在不同基分类器数目下的评测结果。

4. 评分标准

30/100 运行代码后可以生成预测文件。

60/100 评测代码能够运行。

90/100 精度值达到要求。

100/100 完成实验后，试回答下面的问题：

(1) 你对 AdaBoost 算法有何新的认识？

(2) 关于基分类器类型、超参数设置对最终模型性能的影响，你有何发现？