**机器学习框架 第二次竞赛 说明文档 蒋晓伟2017011850**

通过分析题目要求，本次竞赛的任务是实现对已给数据集的所分类任务。可以采取的方法有：决策树算法，随机森林算法，逻辑回归算法，KNN算法等。

在初期我尝试着第一次竞赛中的逻辑回归算法根据softmax回归求出各类别的概率，根据给定的weights得到digit的最大概率的类别，找到softmax回归代价函数最小的参数θ，即通过在θ减去θ处的代价函数的导数，让θ趋于稳定。但是预测所得的正确率较低，如果增加迭代次数计算量就大大增加了，代码跑不动。这时候必须换方法了。

于是我联想到使用K近邻算法在解决鸢尾花分类问题经典模型，所以在本次竞赛中我采用了引用 sklearn库中的K近邻算法实现收入水平的多分类问题：

观察所给数据train.csv和test.csv,其中第2列和第6列的数据类型为字符型，其中还穿插着不明数据用‘？’代替了，设计函数openTrainFile()和openTestFile()在加载训练集合测试集的同时进行数据的规范化，使用map()函数转换数据类型，使用reshape函数使得数据维度保持一致，得到可操作数据集和标签。

引入sklearn库中的KNeighborsClassifier算法，把整理好的数据集放入模型中进行训练，并通过pandas.DataFrame()方法把预测结果输出到csv文件中。

由此，就得到了预测结果jiangxiaowei.csv文件。通过整理相关格式赋予列名，提交到kaggle竞赛的平台上。