**2017级机器学习方向第二次竞赛说明文档**

报告人：安博洋

**第一次尝试：**

当我拿到数据集时，发现这是一个多分类的问题，所以首先想到的算法模型便是softmax软最大化逻辑回归。因为之前用softmax算法模型写过一个手写数字识别的实例，那个实例每个样本有1024个属性，正确率还可以，能够达到84%多，而这个只有13个属性，所以我感觉应该没有问题，于是便开始用代码实现。在这之前，有一个问题需要解决，就是这个数据集中有很多？，我的想法是把？单独当为一个属性值，便对数据集进行了处理，把所有的？用数字9来代替。当把整个模型实现后，用训练集中留下来的一小部分进行争取率的预测，结果正确率只有20%。于是，我决定把数据进行标准化预处理，结果效果还是不理想，只有27%。于是便开始考虑其他的模型。

**第一次尝试：**

后来，我运用sklearn库，把以前学过的算法模型都试过，结果最高的也只有30%多。于是，我想到了最简单的也是最基本的算法模型，KNN模型，用它来预测，结果正确率也只有33%，我开始决定就这样了，就这样交上去吧。

**第一次尝试：**

后来，上课时，颜老师讲到了一个卡方检验的方法，它可以通过一些统计和分析，把最优的K个特征提取出来。于是，我下来自己实现了一下，结果也不尽人意，只提高到了35%多，于是，我便就以这种方法作为最终的算法模型提交。

代码中各个函数的功能简单介绍如下：（核心代码解释在代码.py 文件中以注释方式写出）

**def** str\_to\_float(line):

*# 将字符串的数据类型转换为浮点类型*

**def** load\_data(filename):

*# 读取数据，并对数据中的'?'进行处理*

**def** selectKBest(data):

*# 利用卡方检验方法进行最有特征的提取*

**def** dive\_data(model,data):

*# 将训练集中的数据分为训练集和测试集（用于前期自己验证正确率，调试算法模型和参数）*