**学习小组2机器学习第二次作业**

王锦宏、吴泽辉、陶少聪、王海东、叶泽林、陈耀顺

（中山大学智能工程学院，广东 深圳，518107）

**摘要：**基于信用卡用户的年收入(income)、信用卡每月账单(balance)、是否是学生(student)，预测用户是否会拖欠信用卡还款(default)。对数据集使用10折交叉验证，训练一个线性分类模型，预测信用卡用户是否会拖欠还款。

**关键词：**线性分类 (logistic regression)

**1 描述10折交叉验证对数据集的处理**

所谓交叉验证法，就是将一个数据集分为K份，然后取其中一份作为测试集，剩余K-1份作为训练集。然后，取另一份作为测试集，其余K-1份作为训练集。如此循环，直到每一份都做过测试集为止。本题当中采取的是10折交叉验证，即取k=10时对数据进行处理。每次试验都会得出相应的正确率（或差错率）。10次的结果的正确率（或差错率）的平均值作为对算法精度的估计。一般还需要进行多次10折交叉验证（例如10次10折交叉验证），并且采取不同的划分数据集的方式，减小因样本划分不同而产生的误差。再求其均值，作为对算法准确性的估计。同时十折交叉验证法作为一种对数据处理结果的评估算法，也可以用于对不同模型拟合结果的一种选择。

**2 描述所使用的线性模型**

所用的模型：对数几率回归

对于输入的，计算与设定的阈值比较，大于阈值的为正类，反之为反类

**3 描述训练模型所使用的算法**

为了便于表示和计算，令，，m为样本数，利用最大化对数似然估计：

等价于最小化：

故所要求的参数向量为：

利用梯度下降算法得到最优解：

（1）取初值

=[0.1;0.1;0.1;0.1;0.1]，置k=0；

（2）计算

（3）计算梯度

并且一维搜索确定得到

（4）

并计算，

当时

令，停止迭代

（5）否则置k=k+1,返回（2）

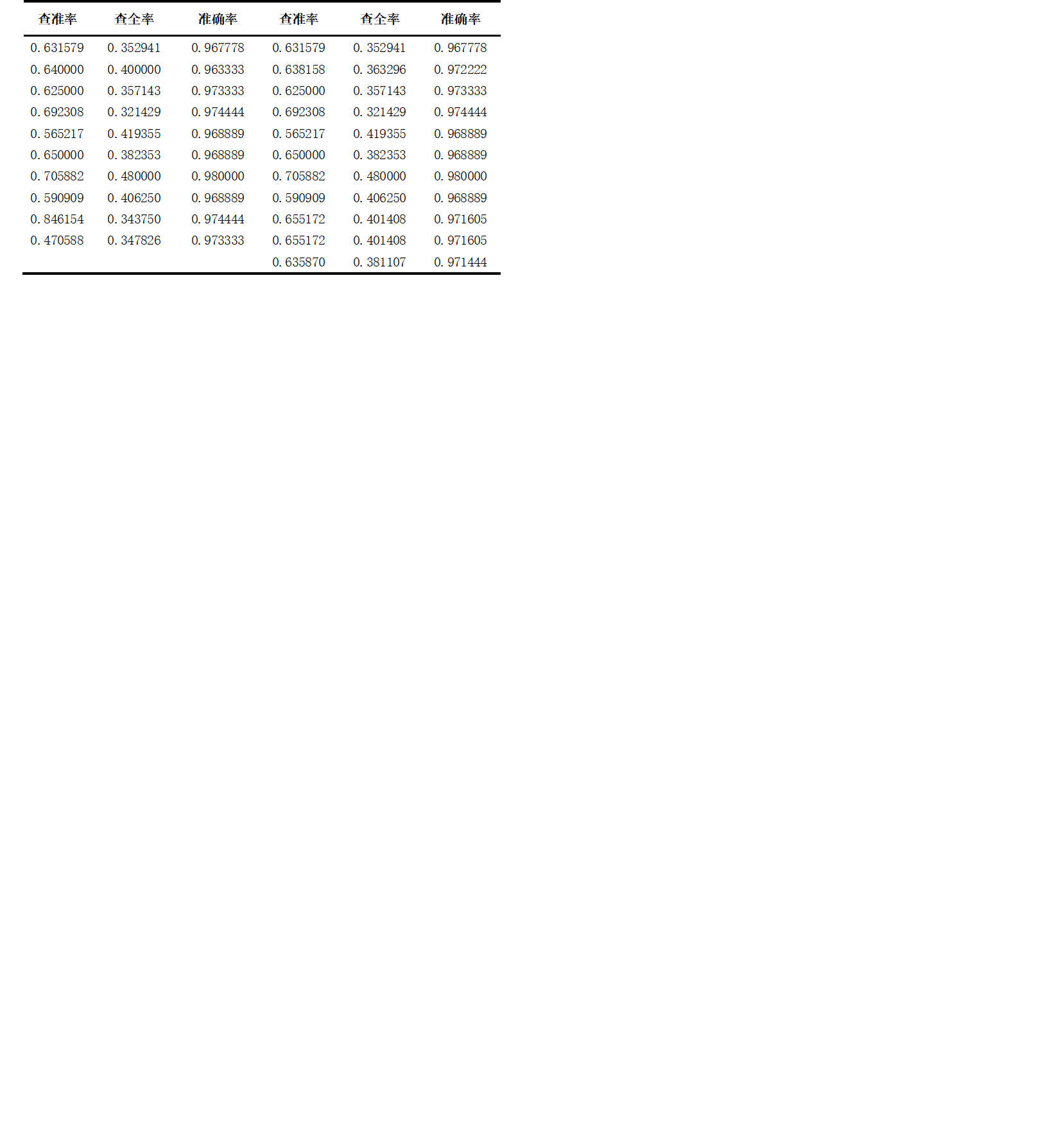


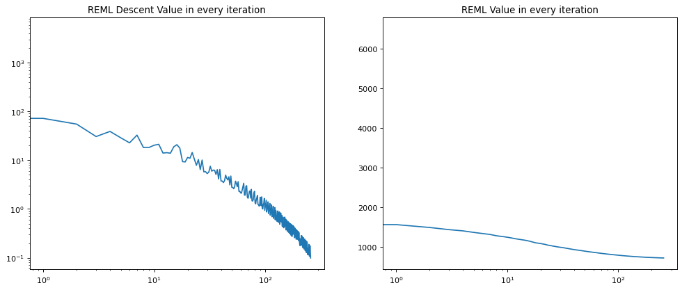
图 2 十折交叉验证对自身、对测试样本的拟合结果

最后一行为全数据样本训练后对自身的拟合结果

**4 模型训练结果，训练集错误率和测试集错误率分析**

**训练过程****的总结**

我组提交的T2.ipynb在进行拟合的时候，同样可以对模型的训练情况进行画图。由于篇幅关系，全部的训练过程图此处不附。取使用源数据训练模型，并用模型验证源数据的情况：

****

在T2模型的训练中，全部的训练组普遍出现了梯度下降值出现了周期性震荡的情况，且随着训练集的不同，测试集震荡的幅度均不同。

十折中所有训练组平均265步左右即可收敛，

由最终求出的ω算出的正例与实际上的正例比较得出真正例、假正例、真反例、假反例的值。查准率为，查全率为，准确率为。

我组的数据样本中，十折交叉验证的**平均查全率：0.381105，平均查准率：0.641764，平均准确率0.971333**

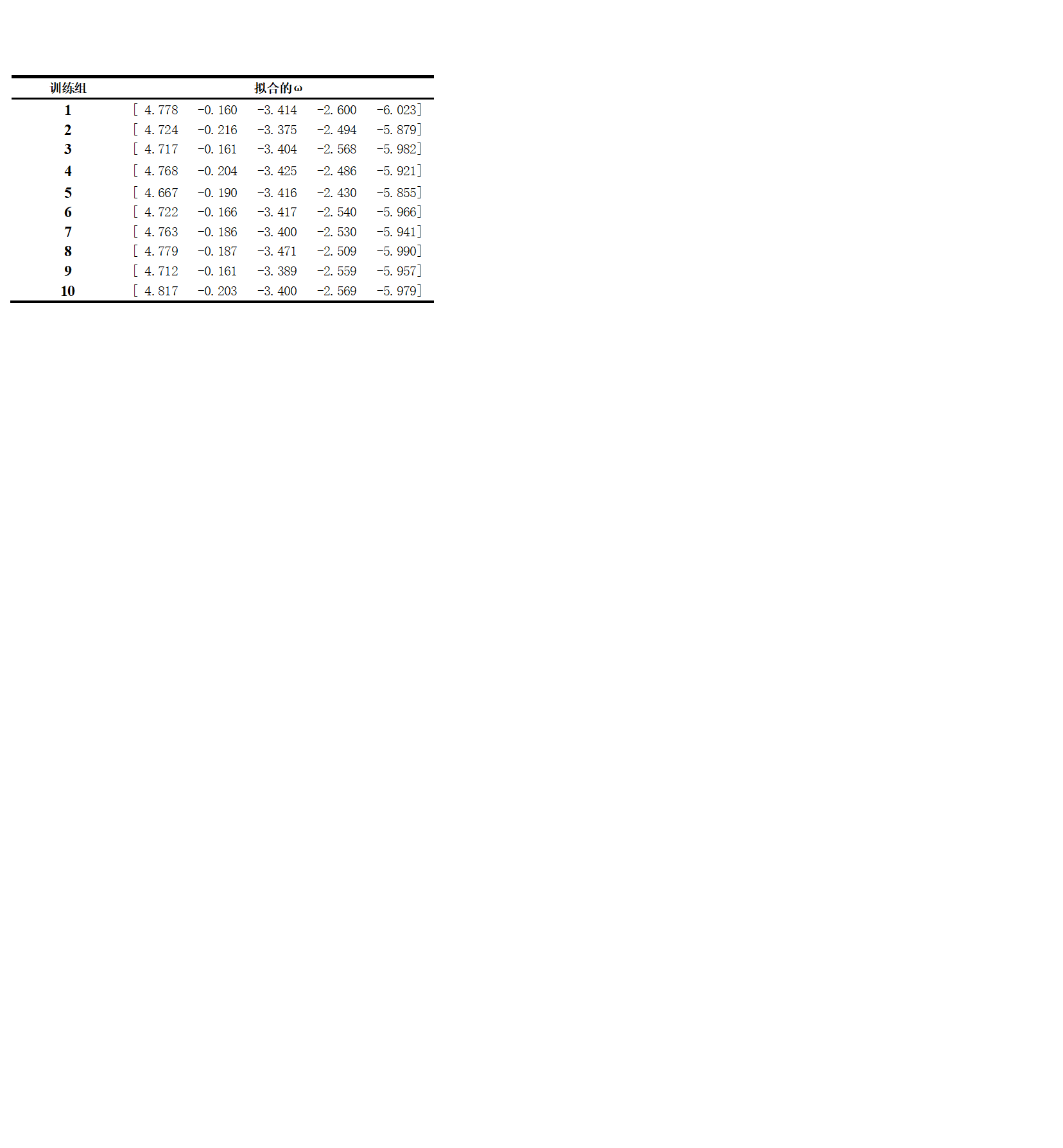
从下面的图2也可看出训练集和测试集之间这三种率很接近，因此泛化效果比较好。

图 1 十折验证训练模型分别拟合得到的ω

**5 模型训练过程中的收获**

**5.1 基础知识**

加深了对python的熟悉，如学习了dataframe等机器学习常用的数据储存方式，加强了pandas、numpy、matplotlib等库的使用；回顾了课本上的数学公式；将算法用代码实现，并用python完成矩阵运算，提升了编程能力；任务管理上，训练了JupyterNotebook集成环境的使用、利用Github完成了小组合作。

**5.2 数据处理**

学习了数据可视化时，对多变量情况需要进行归一化，所得图片才有对比价值；面对类别样本数不平衡问题，使用阈值移动处理，模型能有更好结果。

**5.3 模型调试**

学习步长的选取，太大会无法收敛，太小会使训练速度过慢；调试中建议选取较小的学习步长，虽然学习时间较长，但能避免不收敛的情况。

**5.4 代码改进**

使用矩阵运算，相比循环运算，能有效的提高效率；尝试了除具体迭代外，其他的判准，尝试了对模型效率的提高。