计算机与信息工程学院实验报告（六）

姓名：王赫 学号：1828070097 专业：数据科学与大数据技术(明德计划) 年级：2018级

课程：机器学习与数据挖掘 主讲教师：罗慧敏 辅导教师：\_\_\_\_\_\_\_

实验时间：2020年12月 9日 下午17时至18时，实验地点：606

实验题目： 实现拉普拉斯修正的朴素贝叶斯分类器

实验目的： 掌握朴素贝叶斯分类器的原理及应用

实验环境（硬件和软件） Anaconda/Jupyter notebook/Pycharm

实验内容：

编码实现拉普拉斯修正的朴素贝叶斯分类器，基于给定的训练数据，对测试样本进行判别。

实验步骤：

1. 数学公式推导
2. 编程实现
3. 数据记录
4. 实验总结

实验数据记录：

主要分为四大部分：

（1）获取数据：



import numpy as np

# 1、获取数据

def loaddata():

X = np.array([[1,'S'],[1,'M'],[1,'M'],[1,'S'],

[1,'S'],[2,'S'],[2,'M'],[2,'M'],

[2,'L'],[2,'L'],[3,'L'],[3,'M'],

[3,'M'],[3,'L'],[3,'L']])

y = np.array([-1,-1,1,1,-1,-1,-1,1,1,1,1,1,1,1,-1])

return X, y

（2）拉普拉斯修正算法实现



# 2、训练数据

def trainData(dataSet, labels):

m = dataSet.shape[0] # 获取数据集实体数

n = dataSet.shape[1] # 获取数据集类型数

preProbability = {} # 先验概率字典保存

valueSample = {} # 存储值类型数

conditProbability = {} # 条件概率字典保存

trainLabels = set(labels) # 类型去重，用于计算概率值

# 计算先验概率------>labels出现的次数

for tl in trainLabels:

preProbability[tl] = float(len(labels[labels == tl]))

for i in range(m):

for j in range(n):

if dataSet[i,j] in valueSample.keys():

continue

else:

valueSample[dataSet[i,j]] = len(set(dataSet[:,j]))

# 计算条件概率----->每个属性出现的次数

for i in trainLabels:

dicValue = {}

for j in range(m):

for k in range(n):

if i == labels[j]:

if dataSet[j,k] not in dicValue.keys():

dicValue[dataSet[j,k]] = 1

else:

dicValue[dataSet[j,k]] += 1

conditProbability[i] = dicValue

# 计算拉普拉斯修正概率

for i in conditProbability.keys():

dicValue = conditProbability[i]

for j in dicValue.keys():

p = float(dicValue[j]+1)/float(preProbability[i]+valueSample[j])

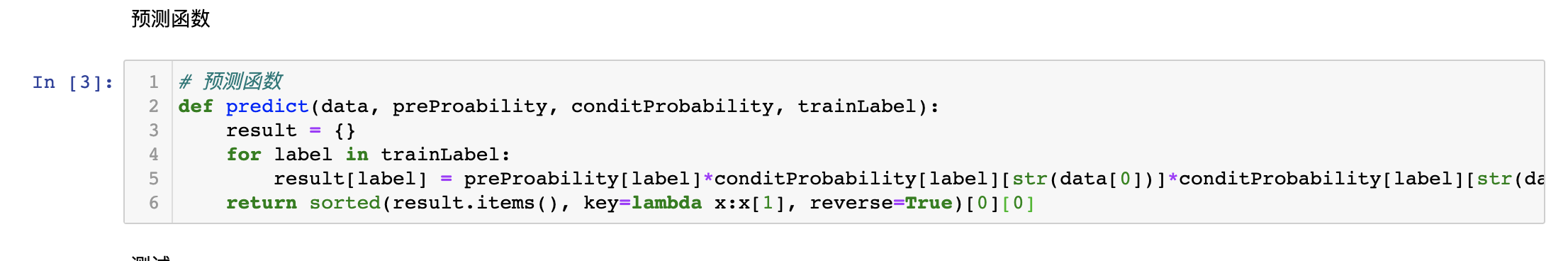
conditProbability[i][j] = p

for i in preProbability.keys():

preProbability[i] = float(preProbability[i])/float(m+len(preProbability.keys()))

return preProbability, conditProbability, trainLabels

（3）预测函数



# 预测函数

def predict(data, preProability, conditProbability, trainLabel):

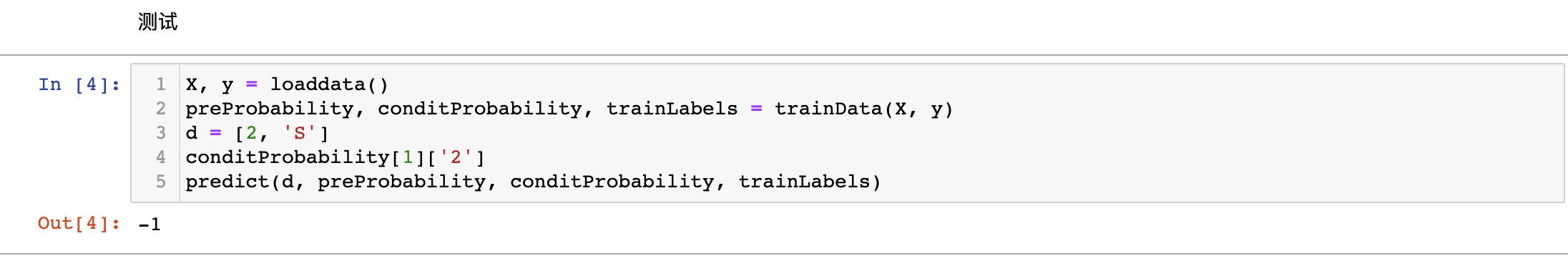
result = {}

for label in trainLabel:

result[label] = preProability[label]\*conditProbability[label][str(data[0])]\*conditProbability[label][str(data[1])]

return sorted(result.items(), key=lambda x:x[1], reverse=True)[0][0]

（4）测试



X, y = loaddata()

preProbability, conditProbability, trainLabels = trainData(X, y)

d = [2, 'S']

conditProbability[1]['2']

predict(d, preProbability, conditProbability, trainLabels)

问题讨论：

本次试验主要进行了朴素贝叶斯分类在拉普拉斯修正下的编程实现。整体实现没有太大问题，但是朴素贝斯分类只适用于小样本，对于大样本的分类学习，该方法迭代复杂程度就会提升，不能应对。