**人工智能实验--第一节特征工程**

**本实验采用python来写代码，主要引用 xlrd，xlwt，numpy，skleaarn，**

**通过 xlrd，xlet进行excl文件的读取和写入 (读取samplesubmission.csv文件)**

**split\_age\_range(user\_id)将特征值 用户id 进行离散化为8个特征值（以id为例）**

**if user\_id >0 and user\_id<=50000:**

**return [1,0,0,0,0,0,0,0,0]**

**elif user\_id >50000 and user\_id<=100000:**

**return [0,1,0,0,0,0,0,0,0]**

**elif user\_id >100000 and user\_id<=150000:**

**return [0,0,1,0,0,0,0,0,0]**

**elif user\_id >150000 and user\_id<=200000:**

**return [0,0,0,1,0,0,0,0,0]**

**elif user\_id >200000 and user\_id<=250000:**

**return [0,0,0,0,1,0,0,0,0]**

**elif user\_id >250000 and user\_id<=300000:**

**return [0,0,0,0,0,1,0,0,0]**

**elif user\_id >300000 and user\_id<=350000:**

**return [0,0,0,0,0,0,1,0,0]**

**elif user\_id >350000 :**

**return [0,0,0,0,0,0,0,1,0]**

**split\_gender(merchant\_id)将特征值merchant\_id进行离散化**

**if merchant\_id >0 and merchant\_id<=3000:**

**return [1,0,0]**

**elif merchant\_id >3000:**

**return [0,1,0]**

**def split\_log():分割数据文件中数据**

import numpy as np

import xlrd

import xlwt

# from sklearn import preprocessing # 进行标准化数据时，需要引入这个包

from sklearn.model\_selection import train\_test\_split

#

def open\_excel(file):

try:

data = xlrd.open\_workbook(file)

return data

except Exception as e:

print(str(e))

#

def split\_age\_range(user\_id):

"""

将特征值年龄进行离散化为8个特征值

:param age: 年龄区间值

:return: 离散化后的特征

"""

if user\_id >0 and user\_id<=50000:

return [1,0,0,0,0,0,0,0,0]

elif user\_id >50000 and user\_id<=100000:

return [0,1,0,0,0,0,0,0,0]

elif user\_id >100000 and user\_id<=150000:

return [0,0,1,0,0,0,0,0,0]

elif user\_id >150000 and user\_id<=200000:

return [0,0,0,1,0,0,0,0,0]

elif user\_id >200000 and user\_id<=250000:

return [0,0,0,0,1,0,0,0,0]

elif user\_id >250000 and user\_id<=300000:

return [0,0,0,0,0,1,0,0,0]

elif user\_id >300000 and user\_id<=350000:

return [0,0,0,0,0,0,1,0,0]

elif user\_id >350000 :

return [0,0,0,0,0,0,0,1,0]

# elif age == 8:

# return [0,0,0,0,0,0,0,0,1]

def split\_gender(merchant\_id):

"""

将特征值性别进行离散化

:param gender:

:return: 返回离散化的特征

"""

if merchant\_id >0 and merchant\_id<=3000:

return [1,0,0]

elif merchant\_id >3000:

return [0,1,0]

# elif gender == 2:

# return [0,0,1]

def split\_log(prob):

# 分割数据文件中的Log数据

# :param Log: Log数据

# :return: 处理后的特征值

items = Log.strip().split('#')

purchase = 0;total = 0

click = 0;add\_to\_card = 0;add\_to\_favourite = 0

for i in range(len(items)):

total += 1

item = items[i].strip().split(':')

if item[4] == '2':

purchase += 1

if item[4] == '1':

add\_to\_card += 1

if item[4] == '3':

add\_to\_favourite += 1

return [float(total),float(round(purchase/total,3)),float(add\_to\_card),float(add\_to\_favourite)]

def loadDataSet(path, training\_sample,colnameindex=0,by\_name=u'Sheet1'):

"""

加载数据

:param path: 数据文件存放路径

:param training\_sample: 数据文件名

:param colnameindex: 文件列名下标

:param by\_name: 表名

:return: 数据集和类别标签

"""

dataMat = [];

labelMat = [] # 定义列表

filename = path + training\_sample

print(filename)

data = open\_excel(filename)

table = data.sheet\_by\_name(by\_name) # 获得表格

nrows = table.nrows # 拿到总共行数

colnames = table.row\_values(colnameindex) # 某一行数据 ['user\_id', 'age\_range', 'gender', 'merchant\_id','label']

for rownum in range(1, nrows): # 也就是从Excel第二行开始，第一行表头不算

row = table.row\_values(rownum)

if row[1] == '' or row[2] == '' or row[5] == '':

continue

if row:

app = []

app = split\_age\_range(row[1])+split\_gender(row[2]) + split\_log(row[5]) # 将Log转化为特征值

dataMat.append(app)

labelMat.append(float(row[4])) # 获取类别标签

return dataMat, labelMat

#

#

#

def main():

"""

主函数

:return: null

"""

wb = xlwt.Workbook()

ws = wb.add\_sheet('Sheet1',cell\_overwrite\_ok=True)

path = "D:\\"

training\_sample = 'qq.xls' # 训练数据文件

#training\_sample = 'samplesubmission.cdv'

trainingSet, trainingLabels = loadDataSet(path, training\_sample) # 取训练数据

# print(len(trainingSet))

num = len(trainingSet)

for i in range(num):

for j in range(16):

ws.write(i,j,trainingSet[i][j])

ws.write(i,j+1,trainingLabels[i])

wb.save('D:\\featuredata.xls')

print("处理完成")

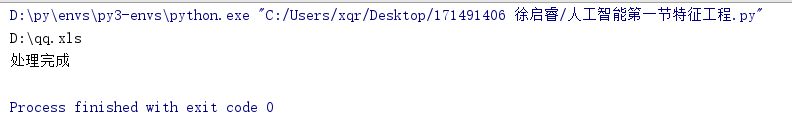
if \_\_name\_\_ == '\_\_main\_\_':

"""

程序入口

"""

main()



171491406 徐启睿