<https://tianchi.aliyun.com/getStart/information.htm?spm=5176.11165320.5678.2.47783c3ad1psT1&raceId=231576>

该网址是天池比赛的其中一个比赛项目，里面有详细说明可以帮助理解数据，方便特征处理。

import numpy as np  
import xlrd  
import xlwt  
  
def open\_excel(file):  
 try:  
 data = xlrd.open\_workbook(file) #打开excel文件  
 return data  
 except Exception as e:  
 print(str(e))  
  
def dateing\_age(path, training\_sample,colnameindex=0,by\_name=u'Sheet1'):  
 filename = path + training\_sample  
 data = open\_excel(filename)  
 table = data.sheet\_by\_name('Sheet1')  
 nrows = table.nrows #对excel的打开处理  
 colnames = table.row\_values(colnameindex) #某一行数据 ['user\_id', 'age\_range', 'gender', 'merchant\_id','label','acticity\_log]  
 zeromum = 0  
 zeroson = 0  
 onemum = 0  
 oneson = 0  
 twomum = 0  
 twoson = 0  
 thrmum = 0  
 thrson = 0  
 fourmum = 0  
 fourson = 0  
 fivemum = 0  
 fiveson = 0  
 sixmum = 0  
 sixson = 0  
 sevenmum = 0  
 sevenson = 0  
 eightmum = 0  
 eightson = 0  
 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)#先处理年龄(1-9)  
 if row[1] == 0:  
 zeromum = zeromum + 1  
 zeroson = zeroson + row[4]  
 elif row[1] == 1:  
 onemum = onemum + 1  
 oneson = oneson + row[4]  
 elif row[1] == 2:  
 twomum = twomum + 1  
 twoson = twoson + row[4]  
 elif row[1] == 3:  
 thrmum = thrmum + 1  
 thrson = thrson + row[4]  
 elif row[1] == 4:  
 fourmum = fourmum + 1  
 fourson = fourson + row[4]  
 elif row[1] == 5:  
 fivemum = fivemum + 1  
 fiveson = fiveson + row[4]  
 elif row[1] == 6:  
 sixmum = sixmum + 1  
 sixson = sixson + row[4]  
 elif row[1] == 7:  
 sevenmum = sevenmum + 1  
 sevenson = sevenson + row[4]  
 elif row[1] == 8:  
 eightmum = eightmum + 1  
 eightson = eightson + row[4]  
 zeroper = float(zeroson / zeromum)  
 oneper = float(oneson / onemum)  
 twoper = float(twoson / twomum)  
 thrper = float(thrson / thrmum)  
 fourper = float(fourson / fourmum)  
 fiveper = float(fiveson / fivemum)  
 sixper = float(sixson / sixmum)  
 sevenper = float(sevenson / sevenmum)  
 eightper = float(eightson / eightmum)  
 return zeroper,oneper,twoper,thrper,fourper,fiveper,sixper,sevenper,eightper  
  
 # app = [] #定义列表  
 # 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 dateing\_gender(path, training\_sample,colnameindex=0,by\_name=u'Sheet1'):#对性别处理  
 filename = path + training\_sample  
 data = open\_excel(filename)  
 table = data.sheet\_by\_name(by\_name)  
 nrows = table.nrows #对excel的打开处理  
 colnames = table.row\_values(colnameindex) # 某一行数据 ['user\_id', 'age\_range', 'gender', 'merchant\_id','label','acticity\_log]  
 zeromum1 = 0  
 zeroson1 = 0  
 onemum1 = 0  
 oneson1 = 0  
 twomum1 = 0  
 twoson1 = 0  
 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:  
 if row[2] == 0:  
 zeromum1 = zeromum1 + 1  
 zeroson1 = zeroson1 + row[4]  
 if row[2] == 1:  
 onemum1 = onemum1 + 1  
 oneson1 = oneson1 + row[4]  
 if row[2] == 2:  
 twomum1 = twomum1 + 1  
 twoson1 = twoson1 + row[4]  
 zeroper = float(zeroson1/zeromum1)  
 oneper = float(oneson1/onemum1)  
 twoper = float(twoson1/twomum1)  
 return zeroper,oneper,twoper  
  
def dateing\_log(path, training\_sample,colnameindex=0,by\_name=u'Sheet1'):#对log处理  
 filename = path + training\_sample  
 data = open\_excel(filename)  
 table = data.sheet\_by\_name(by\_name)  
 nrows = table.nrows #对excel的打开处理  
 colnames = table.row\_values(colnameindex) # 某一行数据 ['user\_id', 'age\_range', 'gender', 'merchant\_id','label','acticity\_log]  
 zerotofivemum = 0  
 zerotofiveson = 0  
 fivetotenmum = 0  
 fivetotenson = 0  
 morethantenmum = 0  
 morethantenson = 0#将log数据分为0-5,5-10,10-15三个层次来处理  
  
 one\_actmum = 0  
 one\_actson = 0  
 two\_actmum = 0  
 two\_actson = 0  
 thr\_actmum = 0  
 thr\_actson = 0#将action分为123类  
 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[5] != '':  
 log = row[5]  
 log2 = log.strip().split("#")  
 if len(log2) <= 5:  
 zerotofivemum = zerotofivemum + 1  
 zerotofiveson = zerotofiveson + row[4]  
 elif len(log2) > 5 and len(log2) <= 10:  
 fivetotenmum = fivetotenmum + 1  
 fivetotenson = fivetotenson + row[4]  
 elif len(log2) >10:  
 morethantenmum = morethantenmum + 1  
 morethantenson = morethantenson + row[4]  
  
  
  
 zerotofiveper = float(zerotofiveson/zerotofivemum)  
 fivetotenper = float(fivetotenson/fivetotenmum)  
 morethantenper = float(morethantenson/morethantenmum)  
  
 return zerotofiveper,fivetotenper,morethantenper  
  
def dateing\_test(path, training\_sample,a1,a2,a3,a4,a5,a6,a7,a8,a9,g1,g2,g3,l1,l2,l3,colnameindex=0,by\_name=u't2'):#对test数据预测概率  
 filename = path + training\_sample  
 data = open\_excel(filename)  
 table = data.sheet\_by\_name(by\_name)  
 nrows = table.nrows #对excel的打开处理  
 colnames = table.row\_values(colnameindex) # 某一行数据 ['user\_id', 'age\_range', 'gender', 'merchant\_id','label','acticity\_log]  
 workbook = xlwt.Workbook(encoding='ascii')  
 worksheet = workbook.add\_sheet('prediction')  
 worksheet.write(0,0, 'user\_id')#从空表格开始写  
 worksheet.write(0,1,'merchant')  
 worksheet.write(0,2,'prob')  
  
 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)  
 log = row[5]  
 log2 = log.strip().split("#")  
 if row[1] == 0: #以下通过age gender log三个特征判别与数据处理  
 if row[2] == 0 :  
 if len(log2) <= 5:  
 label = float(a1 + g1 +l1)/3 #label即为概率  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a1 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a1 + g1 + l3) / 3  
 elif row[2] == 1:  
 if len(log2) <= 5:  
 label = float(a1 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a1 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a1 + g2 + l3) / 3  
 elif row[2] == 2:  
  
 if len(log2) <= 5:  
 label = float(a1 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a1 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a1 + g3 + l3) / 3  
 elif row[1] == 1:  
 if row[2] == 0:  
  
 if len(log2) <= 5:  
 label = float(a2 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a2 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a2 + g1 + l3) / 3  
 elif row[2] == 1:  
 if len(log2) <= 5:  
 label = float(a2 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a2 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a2 + g2 + l3) / 3  
 elif row[2] == 2:  
 if len(log2) <= 5:  
 label = float(a2 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a2 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a2 + g3 + l3) / 3  
 elif row[1] == 2:  
 if row[2] == 0:  
 if len(log2) <= 5:  
 label = float(a3 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a3 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a3 + g1 + l3) / 3  
  
 elif row[2] == 1:  
 if len(log2) <= 5:  
 label = float(a3 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a3 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a3 + g2 + l3) / 3  
  
 elif row[2] == 2:  
 if len(log2) <= 5:  
 label = float(a3 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a3 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a3 + g3 + l3) / 3  
 elif row[1] == 3:  
 if row[2] == 0:  
 if len(log2) <= 5:  
 label = float(a4 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a4 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a4 + g1 + l3) / 3  
 elif row[2] == 1:  
 if len(log2) <= 5:  
 label = float(a4 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a4 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a4 + g2 + l3) / 3  
 elif row[2] == 2:  
 if len(log2) <= 5:  
 label = float(a4 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a4 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a4 + g3 + l3) / 3  
 elif row[1] == 4:  
 if row[2] == 0:  
  
 if len(log2) <= 5:  
 label = float(a5 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a5 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a5 + g1 + l3) / 3  
 elif row[2] == 1:  
  
 if len(log2) <= 5:  
 label = float(a5 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a5 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a5 + g2 + l3) / 3  
 elif row[2] == 2:  
  
 if len(log2) <= 5:  
 label = float(a5 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a5 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a5 + g3 + l3) / 3  
 elif row[1] == 5:  
 if row[2] == 0:  
  
 if len(log2) <= 5:  
 label = float(a6 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a6 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a6 + g1 + l3) / 3  
 elif row[2] == 1:  
  
 if len(log2) <= 5:  
 label = float(a6 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a6 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a6 + g2 + l3) / 3  
  
 elif row[2] == 2:  
  
 if len(log2) <= 5:  
 label = float(a6 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a6 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a6 + g3 + l3) / 3  
 elif row[1] == 6:  
 if row[2] == 0:  
  
 if len(log2) <= 5:  
 label = float(a7 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a7 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a7 + g1 + l3) / 3  
  
 elif row[2] == 1:  
  
 if len(log2) <= 5:  
 label = float(a7 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a7 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a7 + g2 + l3) / 3  
  
 elif row[2] == 2:  
  
 if len(log2) <= 5:  
 label = float(a7 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a7 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a7 + g3 + l3) / 3  
 elif row[1] == 7:  
 if row[2] == 0:  
  
 if len(log2) <= 5:  
 label = float(a8 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a8 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a8 + g1 + l3) / 3  
  
 elif row[2] == 1:  
  
 if len(log2) <= 5:  
 label = float(a8 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a8 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a8 + g2 + l3) / 3  
  
 elif row[2] == 2:  
  
 if len(log2) <= 5:  
 label = float(a8 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a8 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a8 + g3 + l3) / 3  
 elif row[1] == 8:  
 if row[2] == 0:  
  
 if len(log2) <= 5:  
 label = float(a9 + g1 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a9 + g1 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a9 + g1 + l3) / 3  
  
 elif row[2] == 1:  
  
 if len(log2) <= 5:  
 label = float(a9 + g2 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a9 + g2 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a9 + g2 + l3) / 3  
  
 elif row[2] == 2:  
  
 if len(log2) <= 5:  
 label = float(a9 + g3 + l1) / 3  
 elif len(log2) > 5 and len(log2) <= 10:  
 label = float(a9 + g3 + l2) / 3  
 elif len(log2) > 10:  
 label = float(a9 + g3 + l3) / 3  
  
 worksheet.write(rownum, 0,row[0]) # 写入数据！ #0-1为user与gender  
 worksheet.write(rownum, 1,row[3])  
 worksheet.write(rownum, 2,label) #2为概率数据  
  
 workbook.save('prediction.csv') # 保存文件为prediction.csv  
  
def main():  
  
  
 wb1 = xlwt.Workbook()  
 ws1 = wb1.add\_sheet('sheet1',cell\_overwrite\_ok=True)  
 path = 'D:\\'  
 training\_sample = 'lostt2.xlsx' # 打开已知数据进行数据处理  
  
 a1,a2,a3,a4,a5,a6,a7,a8,a9 = dateing\_age(path, training\_sample)#age的9个概率结果  
 print('由年龄(1-9)得出:')  
 print(a1)#年龄0到8的购买概率  
 print(a2)  
 print(a3)  
 print(a4)  
 print(a5)  
 print(a6)  
 print(a7)  
 print(a8)  
 print(a9)  
 print(a1+a2+a3+a4+a5+a6+a7+a8+a9)  
  
 print('分割线')  
 print('由性别(1-3)得出:') #gender的结果：  
 g1,g2,g3 = dateing\_gender(path,training\_sample)  
 print(g1)#女男未知的概率  
 print(g2)  
 print(g3)  
 print(g1+g2+g3)  
  
 print('分割线')  
  
 l1,l2,l3 = dateing\_log(path,training\_sample) #log的结果：  
 print('由操作(1-3)得出')  
 print(l1)#0-5 5-10 10以上操作次数的概率  
 print(l2)  
 print(l3)  
 print(l1+l2+l3)  
  
 wb2 = xlwt.Workbook()#打开test文件进行预测  
 ws2 = wb2.add\_sheet('sheet1',cell\_overwrite\_ok=True)  
 path2 = 'D:\\'  
 training\_sample2 = 't2backup.xlsx' # 训练数据文件  
 dateing\_test(path2, training\_sample2, a1, a2, a3, a4, a5, a6, a7, a8, a9, g1, g2, g3, l1, l2, l3)  
  
  
  
main()

实验结果：

生成文件：

