AI实验一 特征处理

171491328 罗晓璐

一、背景介绍：

为了吸引大量新买家，商家有时会在特定的节日（11月11日）进行大促销。不幸的是，被吸引的买家大多是一次性交易猎手，而这些促销活动可能对销售没什么持久影响。为缓解这个问题，商家要识别谁可以转化为重复的买家。瞄准这些潜在的忠诚客户，商家可以大大降低促销成本，提高投资回报。在网络广告领域，客户定位是极具挑战性的，尤其是对于新买家。随着积累的长期用户行为日志，也许能够解决这个问题。

二、问题：

我们提供了在双11促销期间一组商人他们相应增长的买家。你的任务是预测未来这些新买家是否会成为这些商家的忠诚客户。换句话说，您需要预测这些新购买者在6个月内再次从同一商家购买商品的可能性。

数据集包含匿名用户在“双11”之前和过去6个月的购物日志，以及他们是否是重复购买者的标签信息。由于隐私问题，对数据的采样存在偏差，因此该数据集的统计结果会偏离实际情况，但不会影响解的适用性。

在这个问题中为了找到一个人对应一个商家的忠诚客户概率值，每一个user\_merchant\_id应该是一个样例,而activity\_log是与user\_merchant\_id相关的购买记录。

三、实验处理：

1、预处理：

数据清洗：检测异常样本与缺省处理：

data=data[data.lable>=0]  
*#清洗数据，去掉数据中lable为null和-1的数据*

2、特征工程：

数据归一化：

data\_list=[]  
**for** i **in** range(1,5):  
 **try**:  
 data=pd.read\_csv(**'./formate/formate{}.csv'**.format(i),encording=**'gbk'**)  
 **except**:  
 pd.read\_csv(**'./formate/formate{}.csv'**.format(i), encording=**'gbk'**)  
 *#数据读取并统一格式*

data=pd.concat(data\_list)  
data.head()  
*#合并数据*

离散化

**def** split\_age\_range(age):  
 *"""  
 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** age == 0:  
 **return** [1,0,0,0,0,0,0,0,0]  
 **elif** age == 1:  
 **return** [0,1,0,0,0,0,0,0,0]  
 **elif** age == 2:  
 **return** [0,0,1,0,0,0,0,0,0]  
 **elif** age == 3:  
 **return** [0,0,0,1,0,0,0,0,0]  
 **elif** age == 4:  
 **return** [0,0,0,0,1,0,0,0,0]  
 **elif** age == 5:  
 **return** [0,0,0,0,0,1,0,0,0]  
 **elif** age == 6:  
 **return** [0,0,0,0,0,0,1,0,0]  
 **elif** age == 7 **or** age == 8:  
 **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(gender):  
 *"""  
 将特征值性别进行离散化* **:param** *gender:* **:return***: 返回离散化的特征  
 """* **if** gender == 0:  
 **return** [1,0,0]  
 **elif** gender == 1:  
 **return** [0,1,0]  
 **elif** gender == 2:  
 **return** [0,0,1]

抽取特征

**def** split\_log(Log):  
 *"""  
 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)]

特征选择：

data=data[[**'user\_id'**,**'merchant\_id'**,**'lable'**,**'activity\_log'**]]  
*#提取需要的数据用户ID，商家ID以及是否重复购买和购买日志*

加载数据并存储

**def** loadDataSet(path, training\_sample,colnameindex=0,by\_name=**u'Sheet1'**):  
 *"""  
 加载数据* **:param** *path: "C:\Users\apple\Documents\LXL\AI\data\_format1\format.csv"* **: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  
 data = open\_excel(filename) *#获取文件* table = data.sheet\_by\_name(by\_name) *# 获取Sheet1* 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 = [] *#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

实验结果：

四、实验总结：

1、学会了在项目开始前数据处理的具体步骤。

2、学会了如何使用python去处理数据

3、大概了解了项目的具体流程