**天猫购物行为分析实践**

吴捷成

2017213851

jasonwood2031@gmail.com

胡晗

2017213872

543160303@qq.com

胡昱坤

2017213814

437932979@qq.com

1. **商品购物篮分析**
   1. **问题研究意义**

本问题注重对天猫用户行为日志的分析，对该一数据集上进行商品的购物篮分析的实践研究，为后续的盈利模式和客户个性化服务提供有益的参考。

# 问题研究思路

### **数据预处理**

使用user\_log\_format1.csv构造交易列表，但由于user\_log\_format1.csv数据量太大，这里只取了交易列表中的16\*1024条进行下一步计算。

### **研究方法**

使用经典的Apriori算法进行商品购物篮分析，即回答哪些商品经常在一起购买。

### **算法说明**

Python中用专门的Apriori算法工具包可以使用，为了更加熟悉算法流程，对有关算法代码进行了重新编写。伪代码描述如下：

L1 = {frequent items};

for (k = 1; Lk !=∅; k++) do begin

Ck+1 = candidates generated from Lk;

for each transaction t in database do

increment the count of all candidates in Ck+1 that are contained in t

Lk+1 = candidates in Ck+1 with min\_support

end

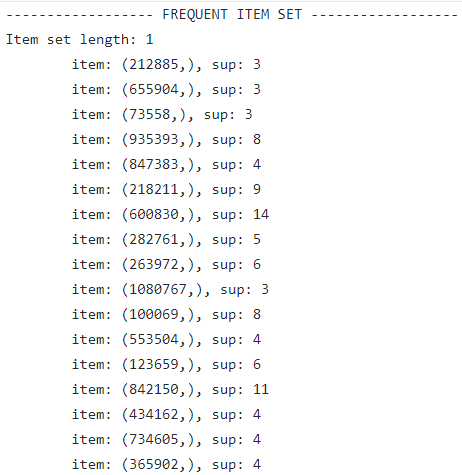
return ∪k Lk;

### **算法优化**

在计算频繁k项集时，可以将交易列表长度小于k的所有记录删除。因为长度小于k的记录一定不包含频繁k项集。

可以将商品id转换为整型值，然后将每条交易记录的商品id排序，在计算频繁k项集的时候可以直接比较整型值而非字符串。

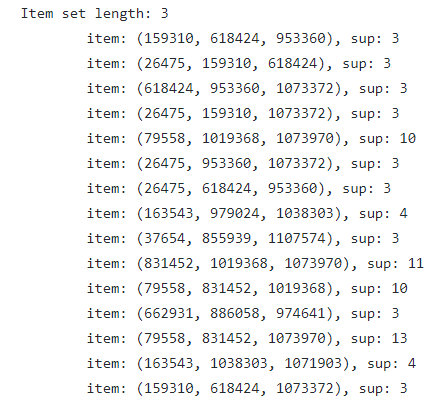
# 实验结果



#### 图 1-1 商品频繁1项集分析

#### 

#### 图 1-2 商品频繁2项集分析



#### 图 1-3 商品频繁3项集分析

#### 

#### 图 1-4 商品频繁4项集、5项集分析

#### 注：item括号中均为购物篮商品的id

# 同一商家用户特征分析

* 1. **研究方法**

使用SOM网络进行基于用户行为的聚类，对同一商家的用户可以分为哪几类 ，这几类有什么显著特征这一问题进行实践研究，从而可以为商家可以为商家可以为商家促销提供重要依据 。

# 算法说明

使用SOM网络进行基于同一商家的用户行为特征的聚类分析。SOM全称为自组织特征映射（Self Organizing feature Map），它可以保留在高维度空间下，簇之间的拓扑关系并将这种拓扑关系映射至低维度（通常是一维或者二维），它是一种非监督性的竞争型神经网络。比较详细系统的介绍<https://en.wikipedia.org/wiki/Self-organizing_map>，不再赘述。

# 算法说明

同样，Python中用专门的SOM算法工具包可以使用，为了更加熟悉算法流程，对有关算法代码进行了重新编写。基于SOM网络进行用户行为的聚类算法说明如下：

1. 首先根据user\_log\_format1.csv构造用户行为。这个表中的action\_type字段记录了用户对商品的行为：

0 - 用户点击该商品

1 - 用户将该商品加入了购物车

2 - 用户购买了该商品

3 - 用户将该商品加入喜欢列表

可以将这个字段下的记录当做用户对该商品的喜好程度的分值。在这里，将所有的用户行为自增1，并将用户对该商品无上述行为标记为0，即

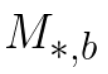
0 - 用户对该商品无操作

1 - 用户点击该商品

2 - 用户将该商品加入了购物车

3 - 用户购买了该商品

4 - 用户将该商品加入喜欢列表

以此构造一个|U|\*|I|稀疏矩阵M，其中U代表用户集合，I代表商品集合。代表第a个用户对所有商品的行为，代表所有用户对第b个商品的行为，代表第a个用户对第b个商品的行为，取值为0-4。

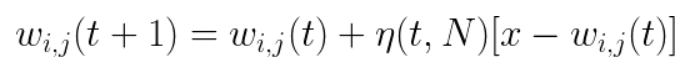
由于计算量大，这里只取512个用户进行聚类，即|U|=512。

2. 确定输出的维度、粒度、初始学习速率η和初始学习半径r，这里取输出维度为二维，粒度为50\*50。这里的50\*50指的是SOM输出层有2500个神经元，最后将输出一张50\*50的热度图，表示用户的聚类拓扑关系热度图。

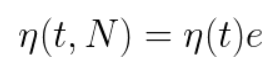
3. 随机生成一个2500\*|I|的权重矩阵，代表每个输出神经元对应的权重向量，类似于一个用户的行为向量。

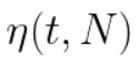
4. 当输入一条记录时，使用距离计算公式（欧氏距离或者余弦距离等）计算与该记录距离最“近”的获胜神经元，以获胜神经元为中心，学习半径r内的所有神经元作为优胜邻域。

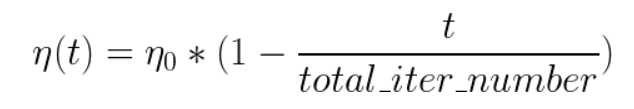
对优胜邻域内的所有神经元使用以下公式调整权重：



其中每次迭代需要使用以下公式计算当次学习速率



上式中，是训练时间t和优胜邻域内第j个神经元与获胜神经元之间的拓扑距离N的函数，其中可采用单调递减函数，即

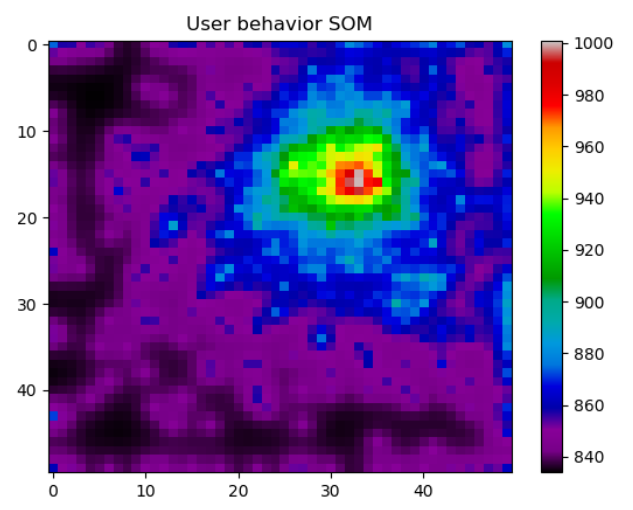


距离函数与学习半径有关，而学习半径也将采用单调递减函数。

5. 最终以热力图的方式展现聚类拓扑结构。

# 实验结果和分析

同一商家用户特征分析热力图，如图2-1所示：



#### 图 2-1 用户特征分析热力图

从图2-1可以比较容易得出用户的分类数目，距离比较集中的例如图中的红色、绿色部分，距离比较分散的可能是购买商品比较小众的用户，对于这些长尾用户的有关特征的研究是下阶段实践研究的一个方向。

1. **商家老用户特征分析**
   1. **数据预处理**

对于用户日志数据中性别不清楚或者缺失的值使用NULL或者2进行填充。

* 1. **研究方法**

从数据集中筛选出老客户数量最多的前10名商家，分析这些商家的老客户的特征，帮助商家更好的制定销售策略。

* 1. **问题建模**

### **特征工程**

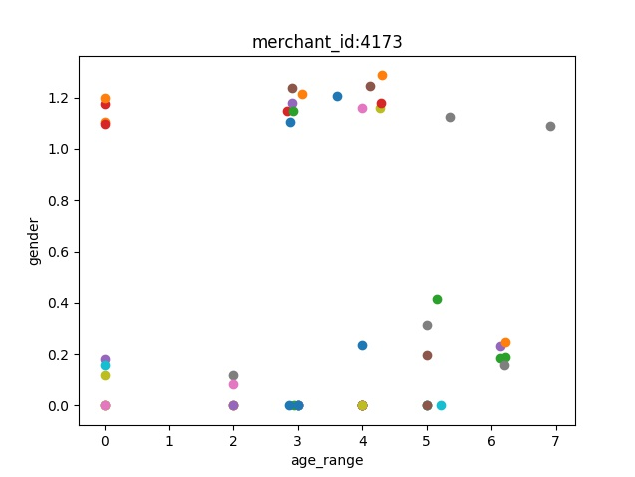
从train\_format1.csv中找出老客户数量最多的前10名商家，从user\_info\_format1.csv表中找出对应老客户的年龄和性别特征，user\_log\_format1.csv表中的特征比如点击、购买某件商品对用户画像意义不大， 所以暂时摒弃。

### **聚类**

对老客户进行kmeans聚类分析。

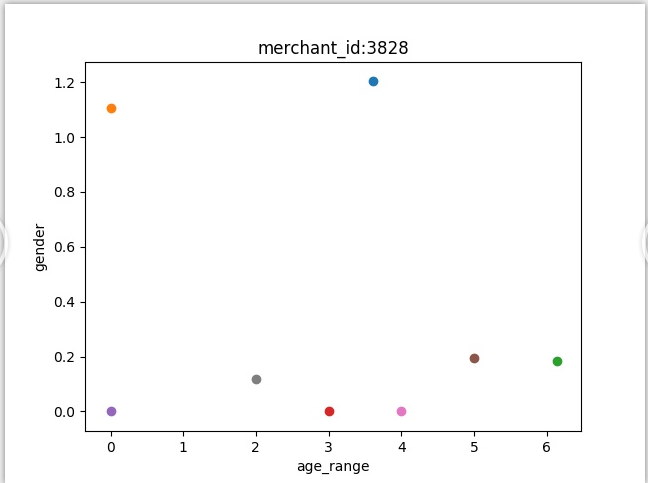
# 实验结果与分析

1.基于商家4173的老用户特征分析，如图3-1所示：



#### 图3-1 基于商家4173的老用户特征分析

从图3-1可以分析出来，商家4173的客户女性多于男性，用户年龄集中在25到34岁。

2. 基于商家3828的老用户特征分析，如图3-2所示： 

#### 图3-2 基于商家3828的老用户特征分析

从图3-2可以明显得知，客户基本是女性，并且年龄分布均匀，可能是买日常用品的等等。

1. **用户回头客预测**
   1. **问题研究意义**

企业的资源是有限的，把资源用在刀刃上才能产生更大的收益。预测用户是否会成为回头客无疑是重要的，可以帮助企业更好的制定战略，将主要的时间和精力放在可能会产生交易的客户身上，以避免弄错方向，做出无用功。

# 研究方法

通过特征工程筛选出有用的特征，融合三张表，使用随机森林、逻辑回归方法训练模型，进行天猫商城的用户回头客预测。最后比较两种模型的相对优劣。

# 问题建模

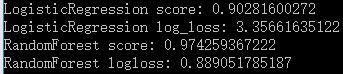
### **特征工程**

从user\_log\_format1.csv中选出favorite的用户，通过user\_d和merchan\_id把三张表的特征集合到一起。对数据进行处理，比如归一化，去掉空缺项，补充一些空缺项。

### **训练过程**

使用随机森林、逻辑回归训练模型。

# 实验结果与分析



通过交叉验证，逻辑回归预测精准度保持在90%左右，logloss在3.3左右，因为train\_format1中负样本比例就接近93%，所以逻辑回归模型在这里并不合格；随机森林预测准确度保持在97.5%左右，logloss在0.88左右，运行时间更长，但可接受。综合来看，随机森林模型在此处预测效果更好。

# 实验展望

在训练集train\_format1中，正负样本并不均衡，在完成作业的时候采取过增大正样本权值的方法平衡正负样本，但是效率提升并不太明显，希望可以找到更好的平衡正负样本的方式和更好的模型研究此问题。

# 文献引用

[1] SOM https://en.wikipedia.org/wiki/Self-organizing\_map