

## 第十四周问答作业

朱士杭 231300027

2024 年 5 月 29 日

## 1 问题一：电商数据预处理

具体见“第十四周问答作业.ipynb”

```
[7]: ##使用数据源处理
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

[8]: ##数据清洗
#加载m10instdata1数据集
Out[8]: read_excel("Online Retail.xlsx")
print("数据维度:", "Data")
#删除缺失值的交易记录
Data = Data.dropna()
Data.reset_index(inplace=True)
Data.reset_index(inplace=True)#更新Data的索引
print("是否包含缺失值的交易记录之前数据:", "Data")
```

Quantity	Invoice	Stockcode	Description	Quantity
0	536765	8502326	WHITE HANGING HEART T-LIGHT HOLDER	6
0	536765	718953	WHITE METAL LANTERN	6
0	536765	844808	CREAM CUPPED METALS COAT HANGER	8
3	536765	8402926	KNITTED UNDER FLAG HOT WATER BOTTLE	6
0	536765	8402926	RED WOOLLY WHITE BOTTLE	6
...	...	...	...	...
514904	581587	22619	PACK OF 20 SPACERY NAPKINS	12
514905	581587	22833	CHILDREN'S APRON DOLLY GIRL	6
514906	581587	23124	CHILDRENS CULINARY DOLLY GIRL	6
514907	581587	23255	CHILDRENS CULINARY CROWN PARADE	4
514908	581587	23138	BAKING SET 9 PIECE RETROSPOF	3

	InvoiceDate	UnitPrice	CustomerID	Country
0	2010-12-01 00:20:00	2.55	17850.0	United Kingdom
1	2010-12-01 00:20:00	3.39	17850.0	United Kingdom
2	2010-12-01 00:20:00	2.75	17850.0	United Kingdom
3	2010-12-01 00:20:00	5.39	17850.0	United Kingdom
4	2010-12-01 00:20:00	3.39	17850.0	United Kingdom
***				
519404	2011-12-09 12:50:00	0.85	12600.0	France
519405	2011-12-09 12:50:00	2.10	12600.0	France
519406	2011-12-09 12:50:00	4.15	12600.0	France
519707	2011-12-09 12:50:00	4.15	12600.0	France
519408	2011-12-09 12:50:00	4.95	12600.0	France

```

[54140] row x 8 columns
主键为表头关键字的索引信息(表头索引)
0 0 536365 051234 INDEX HANGING HATE T-LIGHT HOLOID
1 1 536365 72803 WHITE HATE, LANTERN
2 2 536365 040408 CREAM CUP HEARTS COOL, HANGER
3 3 536365 040492 KNITED UNITED FLAG HOT WATER HEAT,
4 4 536365 040492 RED MOOLY MOOT WATER MOITTE,
...
040324 541504 501587 22613 PACK OF 20 SPOONCOP NAPKINS
040325 541505 501587 22809 CHILDREN'S APRON DOLLY GIRL
040326 541506 501587 23254 CHILDREN'S CUPPLY DOLLY GIRL
040327 541507 501587 23255 CHILDREN'S CUPPLY CIRCUS PARADE
040328 541508 501587 22138 BAKING SET 9 PIECE RETROSPOD

```

	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	6	2010-12-08 08:26:00	2.55	127850	United Kingdom
1	8	2010-12-08 08:26:00	5.79	127850	United Kingdom
2	8	2010-12-08 08:26:00	2.75	127850	United Kingdom
3	6	2010-12-08 08:26:00	3.39	127850	United Kingdom
4	6	2010-12-08 08:26:00	3.39	127850	United Kingdom
...	...	...	...	...	...
608324	12	2011-12-20 12:50:00	0.85	126800	France
608325	6	2011-12-20 12:50:00	2.18	126800	France
608326	4	2011-12-20 12:50:00	4.15	126800	France
608327	1	2011-12-20 12:50:00	4.15	126800	France
608328	1	2011-12-20 12:50:00	4.15	126800	France

[40822 rows x 9 columns]		Index: InvoiceNo: StockCode	Description
InvoiceNo	StockCode	InvoiceNo	StockCode
1	536345	015138	WHITE HANGING HAT T-LIGHT HOLDER
2	536345	718051	WHITE METAL LANTERN
3	536345	048400	CREAM CUPID HEARTS CONT. HANGER
4	536345	048029	KNOTTED UNUSED FLAG HOT WATER BOTTLE
5	536345	048029	RED MIDDLE WHITE METAL HAT
...	...	...	...
408062	541504	005187	PACK OF 20 SPACEDOUT NAPKINS
408065	541505	005187	CHILDREN'S APRON DOLLY GIRL
408066	541506	005187	CHILDREN'S APRON DOLLY GIRL
408067	541507	005187	CHILDREN'S CUTLERY CIRCUS FAIR

Q00020	5411008	5081587	22218	BAKING SET 9 PRICE RETROSPOT		
	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	
0	6	2010-12-04 08:26:00	2.55	278509	United Kingdom	
0	6	2010-12-04 08:26:00	3.39	278509	United Kingdom	
2	8	2010-12-04 08:26:00	2.75	278509	United Kingdom	
3	6	2010-12-04 08:26:00	3.39	278509	United Kingdom	
4	6	2010-12-04 08:26:00	3.39	278509	United Kingdom	
...	...	...	...	...	...	...
0000026	12	2011-12-09 12:50:00	0.85	326808	France	
0000026	6	2011-12-09 12:50:00	2.18	326808	France	
0000026	6	2011-12-09 12:50:00	4.13	326808	France	
0000027	4	2011-12-09 12:50:00	4.15	260809	France	

```
[397924 rows x 9 columns]
```

图 1: 电商数据清洗

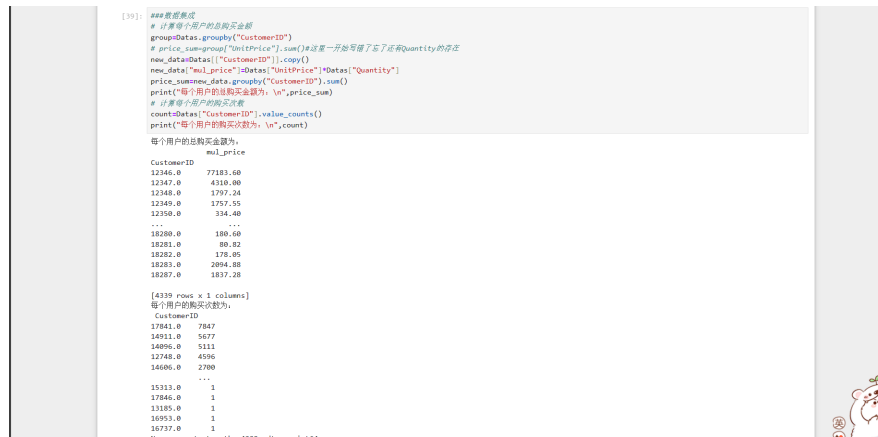


图 2: 电商数据集成

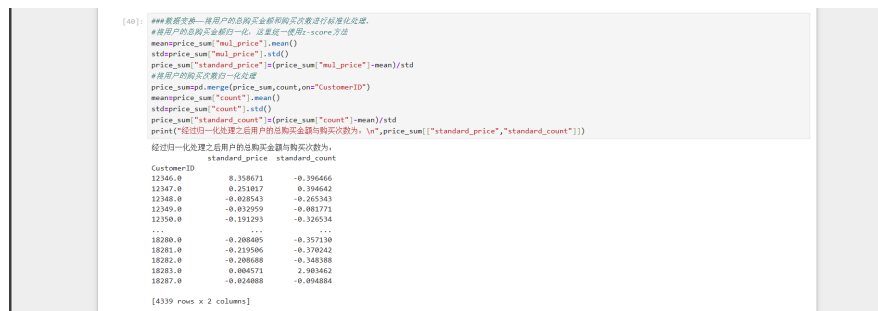


图 3: 电商数据变换

## 2 问题二：为什么在 PCA 之前需要进行标准化处理

如果不进行标准化处理的话很有可能会出现“大数吃小数”的现象，比如说

$$((100000, 0.5), (111000, 0.55), (99900, 0.32))$$

这三个数据点，其实沿着 y 轴方向相对变化也是很大的，但是根据 PCA 的公式，第二维 (y 轴) 的数据直接被“吃掉”了，而我们实际在比较的时候应当是相对分布而不是绝对分布，因此需要使用归一化的手段使得数据分布更加“公平”。



如果是回归任务，则比较和真正结果相差多少（方差）；如果是分类任务，则求出分类的正确率即可

## 4.2 常用评估指标

### 4.2.1 拟合优度（Goodness of Fit）

决定系数（ $R^2$ ）表示模型解释的变异性的比例，值越接近 1 说明模型解释能力越强

调整后的决定系数（Adjusted  $R^2$ ）考虑到自变量数量的影响，当自变量增多时， $R^2$  往往会上升，调整后的  $R^2$  对自变量的数量进行了惩罚，能更准确地反映模型的性能

### 4.2.2 F 检验

用于检验模型整体是否显著，即至少有一个自变量对因变量有显著影响。

### 4.2.3 t 检验

对每个回归系数进行检验，判断每个自变量是否对因变量有显著影响。

### 4.2.4 置信区间

每个回归系数的置信区间，如果置信区间不包括 0，通常认为该系数是显著的

### 4.2.5 均方误差（MSE）

预测值与实际值之间差异的平方的平均数，MSE 越小，说明模型的预测精度越高。

另一种为均方根误差（RMSE）为 MSE 的平方根，它和因变量的量纲一致，更直观地反映了预测误差的大小。

### 4.2.6 交叉验证

通过将数据集分成多个部分，在不同的子集上训练和测试模型，来评估模型的稳健性和预测能力。

#### 4.2.7 残差图

检查残差是否随机分布，如果残差图中存在明显模式（如曲线形状或扇形），则说明模型可能存在问题。

另一种是标准化残差，将残差除以其标准差，用于检测异常值和模型的不稳定性。

## 5 问题五：比较对数几率回归（逻辑斯蒂回归）和 KNN

### 5.1 对数几率回归（Logistic Regression）

对数几率回归是一种广义线性模型，用于二分类或多分类问题。它通过使用逻辑函数（Logistic Function）将线性回归的输出映射到 0 和 1 之间的概率

对数几率回归关注于找到特征和类别概率之间的最佳拟合，核心公式是使用一个连续可微的函数  $y=1/(1+\exp(-wx+b))$  使得一个二分类函数映射到 0-1 之间

训练阶段需要计算参数的梯度并更新，测试阶段对于新的数据点，计算其预测概率是快速的。

通常需要对特征进行缩放，并且可以加入正则化项（如 L1 或 L2 正则化）来防止过拟合

当特征数量较多时，对数几率回归可以有效地进行特征选择和权重分配。它适用于数据量较大，特征维度较高，且需要解释模型的情况

### 5.2 K-最近邻（KNN）

KNN 是一种基于实例的学习算法，它不需要显式地训练模型参数

预测过程中 KNN 算法会在训练集中找到与待预测数据点最接近的 K 个数据点，并根据这 K 个邻居的类别进行投票来决定待预测点的类别。说白了就是找到跟想要预测的点最相似的 K 个点，再根据他们的类别少数服从多数

KNN 适用于数据量较小，特征维度较低的情况。当数据分布不均匀或存在噪声时，KNN 通过平均多个近邻的响应来减少误差。

## 5.3 异同点

### 5.3.1 模型复杂度

对数几率回归在训练阶段有较高的计算复杂度，而 KNN 在测试阶段计算复杂度较高。

### 5.3.2 数据需求

对数几率回归需要较少的数据来训练模型参数，而 KNN 在训练阶段不需要数据来学习参数，但在预测时需要较多的数据来找到近邻。

### 5.3.3 泛化能力

对数几率回归通过模型参数来泛化，而 KNN 通过训练数据来泛化。

### 5.3.4 可解释性

对数几率回归的参数可以直接解释为特征对类别概率的影响，而 KNN 没有显式的参数来解释。

### 5.3.5 对异常值的敏感性

对数几率回归对异常值相对不敏感，因为它寻找的是全局最优解；而 KNN 对异常值非常敏感，因为异常值会影响近邻的选择。

## 5.4 适用场景选择

当数据量较大，特征维度较高，且需要模型具有较好的解释性时，优先选择对数几率回归。当数据量较小，特征维度较低，或者数据分布不均匀，噪声较多时，可以考虑使用 KNN

在实际应用中，通常需要根据具体问题和数据集的特点来选择合适的算法，并通过交叉验证等方法来评估和比较不同算法的性能。

## 6 分类模型对手写数字数据集分类

具体见“第十四周问答作业.ipynb”