第一章 - 绪论与概述

张建章

阿里巴巴商学院

杭州师范大学

2022-09-01

杭州師範大學
Hangzhou Normal University

新大陆
Newland

目录

目录

课程考核说明

根据教学大纲要求，本课程的考核办法为：

$$\begin{aligned} \text{总成绩} = \text{期末成绩} \times 50\% \ &+ \text{日常作业} \times 30\% \\ + \text{日常考勤} \times 10\% \ &+ \text{课堂表现} \times 10\% \end{aligned}$$

其中，期末考试采用**上机考试**形式。

课程简介

课程名称：《商务数据分析实战》

课程目标：

① 掌握经典的数据分析方法；

② 培养数据驱动的商务计算思维；

③ 通过编程高效解决商务分析问题。

授课方式： 课堂讲授 + 实践案例

课程主旨：

① 数据能力传递，数据发现价值，以数据分析推动科学决策；

② 联合新大陆科技集团有限公司开展产学研融合深度合作，推动高校商科人才培养改革。

实验环境

编程语言： Python 3.X

开发环境： Pycharm + Anaconda

交互环境： Jupyter-lab (Anaconda 已内置)

常用软件包： NLTK，scikit-learn，pandas，numpy，matplotlib，上述软件包 Anaconda 均已内置，MLxtend，huggingface，需要通过 pip 命令自行安装。

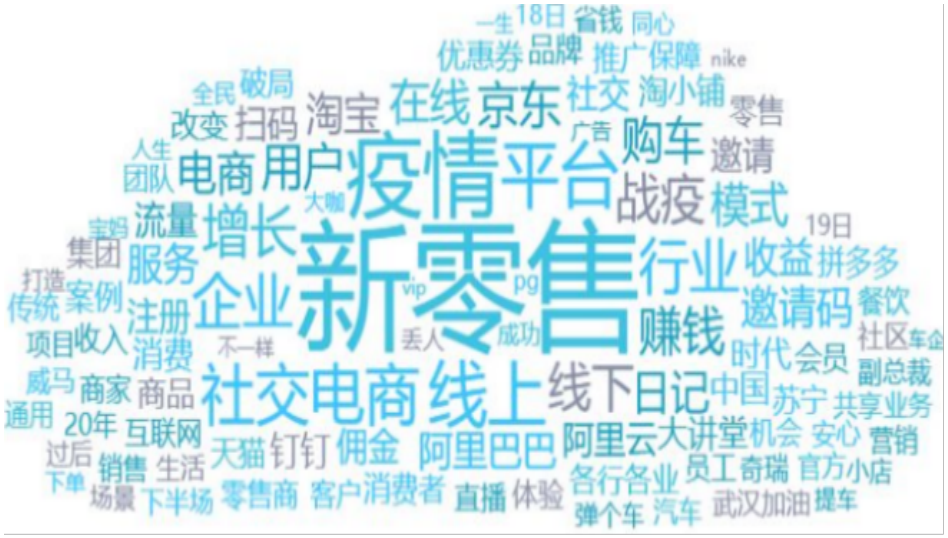
操作系统： Linux 桌面版 (推荐)，Windows，Mac OS (推荐)

在线环境： Kaggle (推荐)，Google colab

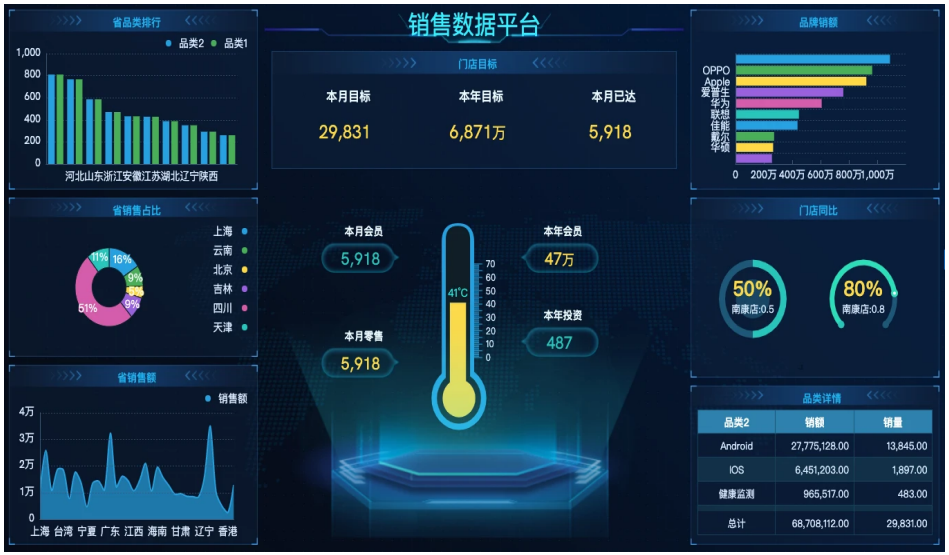
学习资源： Kaggle (推荐)，Towards Data Science (推荐)，Stack Overflow (推荐)，Github，CSDN，阿里云天池

1. 关于课程

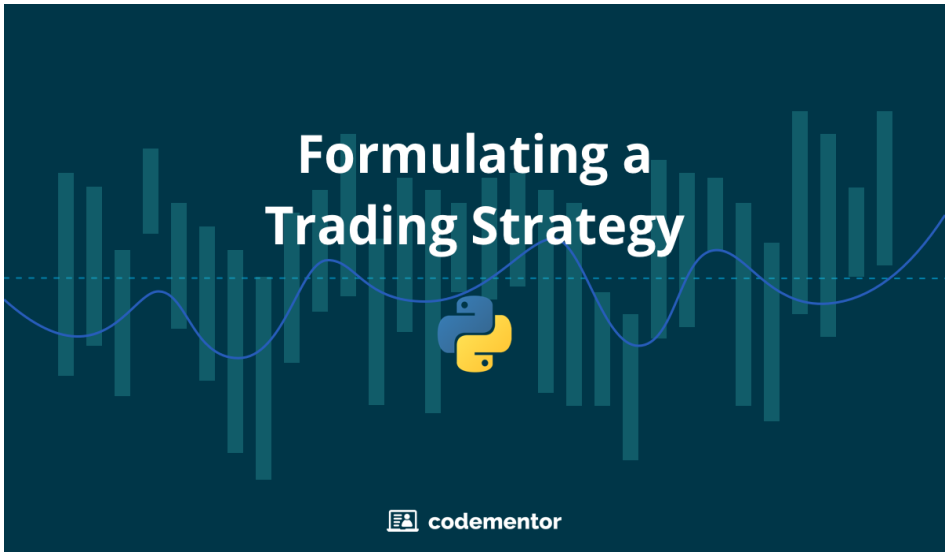
数据分析



文本分析



可视化



量化投资

目录

算法介绍

关联规则分析是零售行业一种常见的分析方法，也被称为**购物篮分析**。该方法也可用于个人商品推荐，如在 Spotify、Netflix 和 Youtube 等应用中推荐音乐、电影和视频。

最著名的案例是**啤酒和尿布之间的相关性**。沃尔玛在研究顾客的购物行为时发现，尿布和啤酒经常一起被购买，事实证明，母亲照顾婴儿期间，负责购物的宝爸会在买尿不湿时顺带买罐啤酒犒劳自己。

Apriori 算法

Apriori 是一种在关系数据库上进行**频繁项集挖掘**和**关联规则学习**的算法。它通过识别数据库中频繁出现的单个项目并将它们扩展到越来越大的项目集，由 Apriori 确定的频繁项集可用于确定表明数据库中总体趋势的关联规则。

令 X 和 Y 代表市场上的**产品集合**，N 代表产品的总数，一条关联规则 $X \Rightarrow Y$ 的强度通常**支持度** (support)、**置信度** (confidence) 和**提升度** (lift) 来确定，如下图

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

规则关联程度的度量

除置信度外，衡量规则的关联程度的其他度量有：

- $Lift = \frac{P(X,Y)}{P(X)P(Y)}$，$Lift = 1$ 时表示 X 和 Y 独立。$Lift$ 越大 (> 1)，越表明 X 和 Y 存在于一条记录中不是偶然现象, 有较强的关联度；
- $Leverage = P(X, Y) - P(X)P(Y)$，$Leverage = 0$ 时，X 和 Y 独立，$Leverage$ 越大 X 和 Y 的关系越密切；
- $Conviction = \frac{P(X)P(!Y)}{P(X,!Y)}$，$(!Y$ 表示 Y 没有发生$)$，用来衡量 X 和 Y 的独立性，这个值越大，X、Y 的关联越大。

目录

1-准备工作

安装机器学习包Mlxtend，`! pip install mlxtend`，导入所需软件包 pandas，mlxtend，加载并查看数据集。

```
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
# 使用pandas读取并查看数据集
df = pd.read_csv('./data/GroceryStoreDataSet.csv', names =
↪  ['products'], sep = ',')
df.head()
```

	products
0	MILK,BREAD,BISCUIT
1	BREAD,MILK,BISCUIT,CORNFLAKES
2	BREAD,TEA,BOURNVITA
3	JAM,MAGGI,BREAD,MILK
4	MAGGI,TEA,BISCUIT

图 1: 前 5 条商品交易记录

2-数据转换

将原始数据转换为 `mlxtend` 中 `apriori` 模型所需的**布尔矩阵**格式。

```
from mlxtend.preprocessing import TransactionEncoder
raw_data = list(df["products"].apply(lambda x:x.split(",") ))
bool_trans = TransactionEncoder()
bool_data = bool_trans.fit(raw_data).transform(raw_data)
```

	BISCUIT	BOURNVITA	BREAD	COCK	COFFEE	CORNFLAKES	JAM	MAGGI	MILK	SUGER	TEA
0	True	False	True	False	False	False	False	False	True	False	False
1	True	False	True	False	False	True	False	False	True	False	False
2	False	True	True	False	False	False	False	False	False	False	True
3	False	False	True	False	False	False	True	True	True	False	False

图 2: 商品交易记录布尔矩阵

每一行表示一条购物记录，每一列表示一种商品。

3-应用 Apriori 算法

使用 Apriori 算法寻找频繁项集，`apriori` 函数的详细用法可通过 `help` 函数查看。

```
# 将支持度设置为0.2，其他参数亦可根据需要调整
freq_df = apriori(df, min_support = 0.2, use_colnames = True,
↪  verbose = 1)
```

12	0.20	(BREAD, TEA)
13	0.20	(CORNFLAKES, COFFEE)
14	0.20	(COFFEE, SUGER)
15	0.20	(MAGGI, TEA)

图 3: 商品交易记录中的部分频繁项集

从上表中可以看到，(面包，茶)，(咖啡，糖) 等都是频繁出现 (频率大于 0.2) 的商品组合。

4-寻找关联规则

设置置信度阈值，使用 `association_rules` 函数寻找商品之间的关联规则。

```
# 置信度值设置为0.6，即，在购买商品X的条件下，
↪  至少有0.6的概率会购买商品Y
df_rules = association_rules(freq_df, metric = "confidence",
↪  min_threshold = 0.6)
df_rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(MILK)	(BREAD)	0.25	0.65	0.2	0.800000	1.230769	0.0375	1.75
1	(SUGER)	(BREAD)	0.30	0.65	0.2	0.666667	1.025641	0.0050	1.05
2	(CORNFLAKES)	(COFFEE)	0.30	0.40	0.2	0.666667	1.666667	0.0800	1.80
3	(SUGER)	(COFFEE)	0.30	0.40	0.2	0.666667	1.666667	0.0800	1.80
4	(MAGGI)	(TEA)	0.25	0.35	0.2	0.800000	2.285714	0.1125	3.25

图 4: 商品交易记录中的关联规则

5-结果分析

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(MILK)	(BREAD)	0.25	0.65	0.2	0.800000	1.230769	0.0375	1.75
1	(SUGER)	(BREAD)	0.30	0.65	0.2	0.666667	1.025641	0.0050	1.05
2	(CORNFLAKES)	(COFFEE)	0.30	0.40	0.2	0.666667	1.666667	0.0800	1.80
3	(SUGER)	(COFFEE)	0.30	0.40	0.2	0.666667	1.666667	0.0800	1.80
4	(MAGGI)	(TEA)	0.25	0.35	0.2	0.800000	2.285714	0.1125	3.25

图 5: 商品交易记录中的关联规则

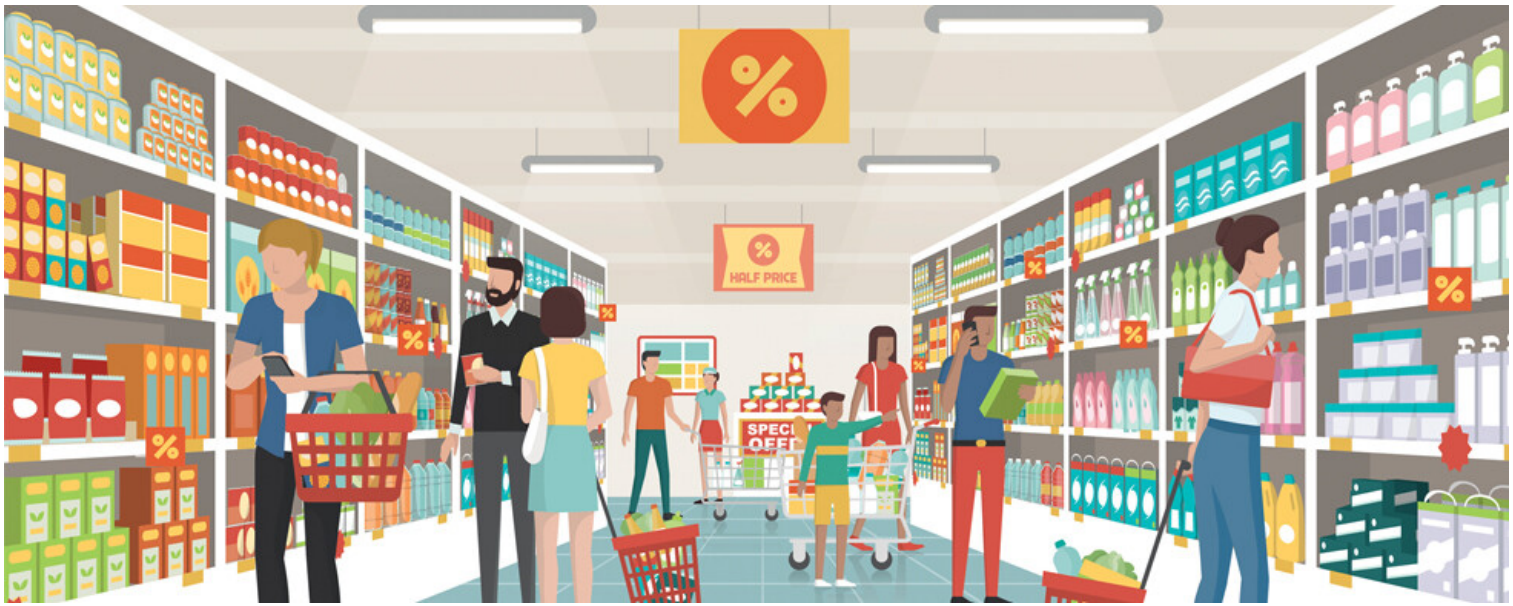
以上述结果中的第四（索引值为 3）条规则为例分析如下：

- 方糖在所有销售记录中的比率占 30%；
- 咖啡在所有销售记录中的比率占 40%；
- 方糖 + 咖啡组合在所有销售记录中的比率占 20%；
- 在购买方糖的顾客中，有约 67% 的顾客也会购买咖啡；
- 购买糖的用户可能会比不购买糖的用户多购买 8% 的咖啡；
- 方糖和咖啡的关联度可用 conviction 值衡量（1.8）。

6-商业决策支持

如果商品集合 X 和 Y 一起购买的频率高，关联性强，可采取以下几个步骤增加利润：

- 通过产品组合改善交叉销售；
- 更改商店布局，将高度关联商品放在一起以提高销售额；
- 为销量较低的产品设计促销活动，以提高销售额；
- 对关联产品，提供组合购买折扣。

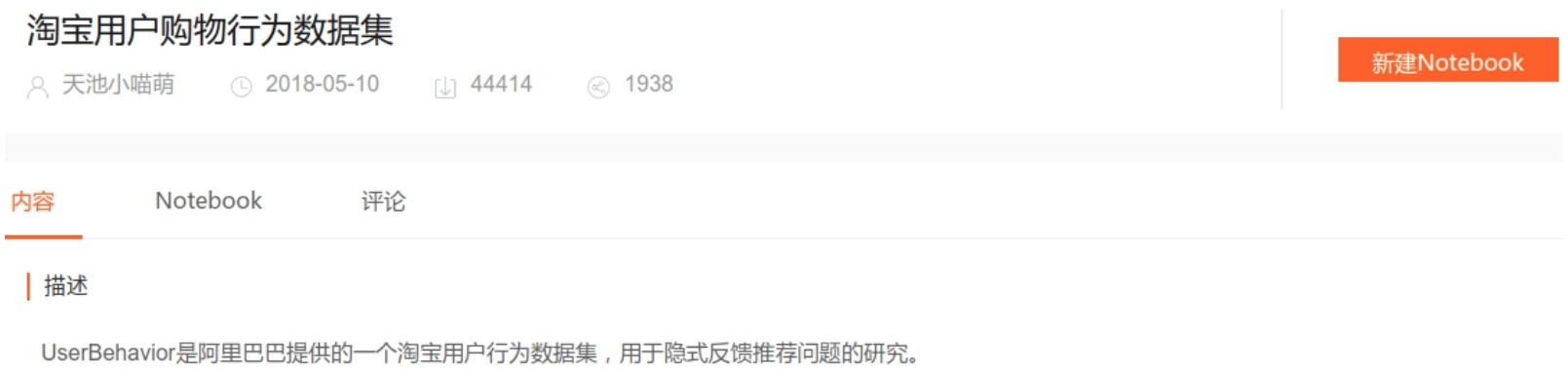


目录

6-商业决策支持

使用淘宝用户购物行为数据集 (数据详情和下载地址)，针对收藏、加入购物车、购买行为，分别挖掘商品之间以及商品类别之间的关联规则。建议流程如下：

- 理解数据各字段含义，理解每条记录含义；
- 数据预处理，按照三类用户行为划分数据；
- 分别挖掘商品和商品类别之间的关联规则；
- 参考课程示例代码应用关联规则挖掘；
- 分析挖掘结果，并为网店或淘宝平台提出有针对性的营销建议。



THE END