SI 671/721 (Fall 2020) Data Mining: Methods and Applications

Instructor: Paramveer Dhillon

GSIs: Lingyun Guo, Kwame Porter Robinson

Homework 1: How Far Would You Go for Italian Mozzarella? Exploring the impact of product cost and purchase frequency on distance traveled by Italian consumers

Due: 10.5.2020

Total Points: 100

1 Summary

We will use the supermarket dataset described in Pennacchioli et al. (2013) for this homework [1]. The dataset was obtained from one of the largest Italian retail distribution companies, *Coop*. It includes three space delimited files:

- `supermarket_distances` A matrix of distances for customers and visited stores.
- `supermarket_prices` A matrix of prices and products.
- `supermarket_purchases` A transaction matrix of customers, products, visited stores, as well as the quantity of purchased products.

We will use concepts from data mining to explore interesting associations within the dataset. This exploration will support our investigation into the effects of product cost, purchase frequency on the distance traveled by Italian consumers.

We recommend that you use Jupyter notebooks and Python libraries `Pandas` and `Scipy` for this homework. We also recommend that you use `Numpy` , `Matplotlib` , `Seaborn` , and `apyori` . However, feel free to use whatever language and library that you would like.

Your Jupyter notebook should contain for each answer a few line description of your solution or approach along with any code. Do not only submit code.

Submit your homework through Canvas.

References:

[1] Pennacchioli et al. "Explaining the product range effect in purchase data" (2013) BigData. DOI: doi.org/10.1109/BigData.2013.6691634 (https://doi.org/10.1109/BigData.2013.6691634)

2 Details

This homework is divided into two parts.

Part 1. Data Exploration

Part 2. Exploring product frequency, price and distance traveled

2.1 Part 1: Data Exploration [50 Points]

This part of the homework is designed to help you familiarize yourself with the datasets and basic concepts of itemset mining. The insights from this part of the homework will help you in Part 2 of the homework.

2.1.1: Better understand the *Coop* supermarket dataset

a) [5 points] To gain more insight into the dataset and problem motivation, please read the following:

- The `Introduction` section of Pennacchioli et al. "Explaining the product range effect in purchase data" (2013) BigData. (http://www.michelecoscia.com/wp-content/uploads/2013/09/geocoop.pdf)
- A deeper explanation of the *Supermarket Data* here (http://www.michelecoscia.com/?page_id=379)

b) [5 points] Download the supermarket dataset here (http://michelecoscia.com/wp-content/uploads/2013/02/supermarket_data.zip). The dataset is ~65 megabytes big. We refer to the entire dataset as *S* or *Supermarket*. You may do this in code or manually.

c). [13 points] Briefly describe how the files and fields in the supermarket dataset are related. For each file be sure to detail what the file generally covers and what each field represents.

Also, the `supermarket_purchases` has a large number of rows. To reduce memory usage and future algorithmic runtime in Part 2 use a method that randomly reads in ~50% of the total rows (e.g. does not read in the entire file into memory but samples from the file as it is read); in `pandas` this would be the `skip_rows` argument to `read_csv` . Before reading in the file seed the random generator with a value of 671 (e.g. `random.seed(671)`) so that your sample is reproducible.

2.1.2: Exploratory Data Analysis (EDA)

(Part I, continued)

We want to better understand the data driven relationships within the dataset. To do so we will plot the distributions of key variables of interest and describe them. Some of the following questions can be answered through plots, histograms and/or pairwise plots. You may use other EDA plots and tools as desired as long as you describe what is observed and explain possible causes for what you observe.

d) [5 points] For `supermarket_distances` , please describe the shape of the `distance` distribution; is it normal? is it bi-modal? is it skewed? Briefly speculate why or why not your description is justified.

e) [5 points] For `supermarket_prices`, please describe the shape of the `price` distribution; is it normal? is it bi-modal? is it skewed? Briefly speculate why or why not your description is justified.

f) [17 points] In (e) you investigated the shape of the `supermarket_prices` distribution. For efficiency purposes, first take a random sample of 10,000 customers.

Then create a subset, C_{lower}, of customers for those customers that spent less than average on products. Create a subset, C_{higher}, of customers that spent more than average on products. Does the average `distance` traveled differ across these subsets? Is this difference *significant* (show your work using a 95% confidence interval)? In your answer include why or why not your explanation is justified.

The main deliverable for this part of the homework is 1) a step-by-step exploration and answers within a Jupyter Notebook, 2) a PDF document containing the answers to each of the questions above (this can be a PDF version of your Jupyter notebook).

2.2 Part 2: Applying Itemset Mining [50 Points]

For this part of the homework you will use itemset mining techniques to explore various relationships to the distance traveled by customers. Using the "language" of these data mining techniques to describe and justify patterns observed in a dataset is one of the primary contributions a data scientist can make in industry.

a) [6 Points] For all products sold what are the top 20 most frequent products?

b) [12 Points] Each top 20 product was brought by a set of customers. What is the average distance traveled to the top 20 products by customers (that is, people traveled to get the top 20 products, how much did they travel on average)? What is the average distance traveled to all other products not in the top 20? Is there a difference? Is it a signicant difference?

c) [6 Points] In itemset mining why would we ignore very high selling products that are readily available, like `toothbrush`? Are high selling, easily available products likely to yield high confidence and high interest association rules?

d) [6 points] Referring to Part 1, for C_{lower} customers what are the top 20 most frequent products? For C_{higher} customers what are the top 20 most frequent products?

e) [20 points total] Imagine that *Coop* wants to negotiate a better deal from its suppliers. To argue for a better deal, Coop wants to ask for bulk discounts on *groups* of products. Coop only cares about the top 5 products sold by quantity found in 2(d) above.

To help justify what groups of products should be considered Coop has asked you:

- [worth 7 points] For C_{lower} customers, using the Apriori algorithm, what are the most frequent itemsets that are purchased from stores? For computational efficiency only run the algorithm on data containing the products in C_{lower}'s top 5 most frequently sold products.

- [worth 7 points] For C_{higher} customers, using and Apriori algorithm, what are the most frequent itemsets that are purchased from stores? For computational efficiency only run the algorithm on data containing the products in C_{higher}'s top 5 most frequently sold products.
- [worth 6 points] Computer similarity between the C_{lower} and C_{higher} itemsets using Jaccard similarity. Compare only itemsets in the top 20 ranked by confidence. Are there any itemsets have a Jaccard similarity above 0?

The apriori algorithm parameters should be set to: `min_confidence=0.50` , `min_lift=1.0` and `max_length=5`

Apriori, C_{lower}

Apriori, C_{higher}

Jaccard Similarity

3 Submission

All submissions should be made electronically by 11:59 PM EST on October 5, 2020.

Here are the main deliverables:

- A PDF version of your executed Jupyter notebook
- The actual Jupyter notebook, so that we can check your results

4 Academic Honesty

Unless otherwise specified in the homework, all submitted work must be your own original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the University's policies on Academic and Professional Integrity may result in serious penalties, which might range from failing a homework, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to the concerned authorities. Consequences of academic misconduct are determined by the faculty instructor; additional sanctions may be imposed.

5 References

Pennacchioli et al. "Explaining the product range effect in purchase data" (2013) BigData. DOI: doi.org/10.1109/BigData.2013.6691634 (https://doi.org/10.1109/BigData.2013.6691634)