

- Due Jul 12, 2023 by 11:59pm
- Points 100
- Submitting a file upload
- Available until Jul 15, 2023 at 11:59pm

This assignment was locked Jul 15, 2023 at 11:59pm.

A6 Clustering and Association Rule Mining

Instructions

Objectives: Find similar customer shopping visits and mine association rules in Walmart baskets.

Input files:

- **Walmart_visits_7trips.csv** (<https://utah.instructure.com/courses/882385/files/147060853/download?wrap=1>). [↓](#)
(https://utah.instructure.com/courses/882385/files/147060853/download?download_frd=1) for K-means clustering
- **Walmart_baskets_1week.csv** (<https://utah.instructure.com/courses/882385/files/147060849/download?wrap=1>). [↓](#)
(https://utah.instructure.com/courses/882385/files/147060849/download?download_frd=1) for Association rule mining

Data Source: Walmart and Kaggle

Walmart held its 3rd Kaggle “recruiting” competition (<https://www.kaggle.com/c/walmart-recruiting-trip-type-classification> (<https://www.kaggle.com/c/walmart-recruiting-trip-type-classification>)) in Fall 2015 to attract data scientists interested in getting jobs at Walmart.

The raw data is a long market basket format of items purchased during each visit and some other item-related information and trip information (see Appendix A).

For Walmart's competition in Kaggle, data scientists compete on classifying shopping trip types based on the items that customers purchased. To give a few hypothetical examples of trip types: a customer may make a small daily dinner trip, a weekly large grocery trip, a trip to buy gifts for an upcoming holiday, or a trip to buy seasonal items. For classifying a customer visit's trip_type, the raw data needs to be processed to aggregate items by VisitNumber and to extract meaningful and potentially effective predictors for TripType. Besides being regarded as data preparation, this is also about feature (predictor) engineering.

To explore trip data by finding similar customer shopping trips using clustering, **Walmart_visits_7trips.csv** (<https://utah.instructure.com/courses/882385/files/147060853/download?wrap=1>). [↓](#)
(https://utah.instructure.com/courses/882385/files/147060853/download?download_frd=1) was created via such a feature engineering process to contain the following features:

- TripType - a categorical id representing the type of shopping trip the customer made. Trip type, 999, is an "other" category.
Walmart_visits_7trips.csv (<https://utah.instructure.com/courses/882385/files/147060853/download?wrap=1>). [↓](#)
(https://utah.instructure.com/courses/882385/files/147060853/download?download_frd=1) only contains a few of the original 38 trip types.
- DOW – Day of Week of the trip
- UniqueItems – the number of unique UPC numbers of the products purchased in a visit
- TotalQty - the total number of the items that were purchased in a visit
- TotalRtrnQty - the total number of the items returned in a visit
- NetQty = total_purchase_quantity – total_return_quantity
- UniqueDepts – the number of unique departments representing the purchased items in a visit.
- OneItemDepts – the number of unique departments representing single-departmental-product purchases in a visit
- RtrnDepts – the number of unique departments representing the returned items in a visit.

To practice Association Rule Mining, **Walmart_baskets_1week.csv** (<https://utah.instructure.com/courses/882385/files/147060849/download?wrap=1>) [↓](#) (https://utah.instructure.com/courses/882385/files/147060849/download?download_frd=1) was derived from the original Kaggle data to represent shopping baskets also in the **long format** based on VisitNumber as transaction id and DepartmentDescription as a high-level indication of item type in a basket.

Packages required: Install C50, psych, RWeka, caret, rminer, matrixStats, knitr and arules packages.

Submit two files – A6_LastName_FirstName.Rmd and A6_LastName_FirstName.html.


(Please do not compress these two files into one zip file. Submit them as separate files. Be sure that the HTML output file contains assignment title, author name – you, the file creation date, and table of content including section numbers.)

Note: If there is any discrepancy in the task descriptions in the assignment page and the short rubric descriptions, please follow the task descriptions in the assignment page.

Task I (95%)

Create A6_LastName_FirstName.Rmd to meet the following requirements:

1. Code chunk 1 - Load packages, prepare and inspect the data (18 points)

- A. (7 points) Package loading, and [Walmart_visits_7trips.csv](https://utah.instructure.com/courses/882385/files/147060853/download?wrap=1) (<https://utah.instructure.com/courses/882385/files/147060853/download?wrap=1>)  (https://utah.instructure.com/courses/882385/files/147060853/download?download_frd=1) import and transformation. Show the overall structure of the input file. Transform factor variables, and show a summary of the input data file.
- B. (3 points) Understand this data set using correlation analysis (pairs.panels from psych)
- C. (8 points) Build a descriptive C5.0 decision tree using the entire data set (TripType is the target variable). Prune the tree so that the number of tree leaves is smaller than 15 (use CF value to prune the tree). Plot the tree and show summary of the model to view tree rules and confusion matrix.

2. Code chunk 2 - Use SimpleKMeans clustering to understand visits (42 points)

- A. Save the number of unique TripType in the imported data as *TripType.levels*. Remove TripType from input data.
- B. Generate clusters with the default (i.e. random) initial cluster assignment and the default distance function (Euclidean). The number of clusters equals to *TripType.levels*. Show the clustering information with the standard deviations and the centroids of the clusters.
- C. Keep the number of clusters at *TripType.levels* and the Euclidean distance function. Change the initial cluster assignment method to the Kmeans++ method. Cluster the visits again and show the standard deviations and the centroids of the clusters.
- D. Keep the number of clusters at *TripType.levels* and the initial cluster assignment method to be the Kmeans++ method. Change the distance function to "weka.core.ManhattanDistance". Cluster the visits again and show the standard deviations and the centroids of the clusters.
- E. Choose your own distance function and initial cluster assignment method, increase or decrease the number of clusters. Cluster the visits again and show the standard deviations and the centroids of the clusters.

3. Code Chunk 3 - Market Basket Analysis with the Walmart dept baskets (35 points).

- A. (7 points) Import [Walmart_baskets_1week.csv](https://utah.instructure.com/courses/882385/files/147060849/download?wrap=1) (<https://utah.instructure.com/courses/882385/files/147060849/download?wrap=1>)  (https://utah.instructure.com/courses/882385/files/147060849/download?download_frd=1) using the following read.transactions() with the "single" format (for long format) and save it in a sparse matrix called, e.g., Dept_baskets.

```
Dept_baskets <- read.transactions("Walmart_baskets_1week.csv", format="single", sep = ",", header = TRUE,
  cols=c("VisitNumber", "DepartmentDescription"))
```

- B. (3 points) Inspect the first 15 transactions.
- C. (5 points) Use the itemFrequencyPlot command to plot the most frequent 15 items in the descending order of transaction frequency in percentage.
- D. (20 points) Associate rule mining
 - i. Use the apriori command to generate about 50 to 100 association rules from the input data. Set your own minimum support and confidence threshold levels. Remember if the thresholds are too low, you will generate more rules than desired, or if you set them too high, you may not generate any or a sufficient number of rules. Show the rules in the descending order of their lift values.
 - ii. Similar to the last task, use the apriori command now to generate about 100 - 200 association rules from the input data. Set your own minimum support and confidence threshold levels. Show the rules in the descending order of their lift values.

For each chunk:

- Add some simple descriptive text in the text area before the code chunk.

- Add a name or description of each code chunk in {r}. Be sure that you allow code and output from executing code to be included in the file from rendering A6_LastName_FirstName.Rmd.
- Feel free to add comment lines with the requirement item numbers (e.g., # 3.A or # 3.B) to your code cell to help TAs and instructors easily identify your code that addresses a particular requirement.

Task II Reflections (5%):

What have you learned from building each of these models and the modeling impact of your adjustments to the hyperparameters or dataset? What can you say about the clusters that were formed? Is there anything interesting to point out? Recall clustering is often used to discover latent (hidden) information. What have you discovered? Make sure to discuss the association rule mining results as well.

If you were explaining the results of these models to a supervisor what would you say about them? Attempt to do more than just state facts here, interpret the results. Coding is great, interpretation of output is even more important. Discuss each model. Write at least 150 words.

Use the RMD header to include assignment title, author name – you and the file creation date. Also include header specifications to generate a table of contents and section numbers of your code chunks.

Render A6_LastName_FirstName.Rmd to a HTML file and submit.

(Please do not compress these two files into one zip file. Submit them as separate files. Be sure that the HTML output file contains assignment title, author name – you, the file creation date, and table of content including section numbers.)

Appendix A:

Data fields in the raw Walmart 2015 data:

- TripType - a categorical id representing the type of shopping trip the customer made. This is the ground truth that you are predicting. TripType_999 is an "other" category.
- VisitNumber - an id corresponding to a single trip by a single customer
- Weekday - the weekday of the trip
- Upc - the UPC number of the product purchased
- ScanCount - the number of the given item that was purchased. A negative value indicates a product return.
- DepartmentDescription - a high-level description of the item's department
- FinelineNumber - a more refined category for each of the products, created by Walmart

Criteria	Ratings			Pts
1.A. Package loading, and Walmart_visits_7trips.csv import and transformation. Show the overall structure of the input file. Transform factor variables, and show a summary of the input data file	7 to >6.0 pts Full Marks	6 to >0.0 pts Partially	0 pts No Marks	7 pts
1.B. Using pairs.panels for Walmart_visits_s2018.csv	3 to >2.0 pts Full Marks	2 to >0.0 pts Partially	0 pts No Marks	3 pts
1.C.i. Build a C5.0 decision tree using the entire data set (TripType is the target variable). Prune the tree so that the number of tree leaves is smaller than 15.	4 to >3.0 pts Full Marks	3 to >0.0 pts Partially	0 pts No Marks	4 pts
1.C.ii. Plot the tree and show summary of the model	4 to >3.0 pts Full Marks	3 to >0.0 pts Partially	0 pts No Marks	4 pts
2.A.i. Create variable TripType.levels and save the number of unique TripType on it.	3 to >2.0 pts Full Marks	2 to >0.0 pts Partially	0 pts No Marks	3 pts
2.A.ii. Remove TripType from input data.	3 to >2.0 pts Full Marks	2 to >0.0 pts Partially	0 pts No Marks	3 pts
2.B.i. Generate clusters with the default (i.e. random) initial cluster assignment and the default distance function (Euclidean). The number of clusters equals to TripType.levels	7 to >6.0 pts Full Marks	6 to >0.0 pts Partially	0 pts No Marks	7 pts
2.B.ii. Show the clustering information with the standard deviations and the centroids of the clusters.	2 to >1.0 pts Full Marks	1 to >0.0 pts Partially	0 pts No Marks	2 pts
2.C.i. Keep the number of clusters at TripType.levels and the Euclidean distance function. Change the initial cluster assignment method to the Kmeans++ method.	7 to >6.0 pts Full Marks	6 to >0.0 pts Partially	0 pts No Marks	7 pts
2.C.ii. Show the clustering information with the standard deviations and the centroids of the clusters.	2 to >1.0 pts Full Marks	1 to >0.0 pts Partially	0 pts No Marks	2 pts
2.D.i. Keep the number of clusters at TripType.levels and the initial cluster assignment method to be the Kmeans++ method. Change the distance function to "weka.core.ManhattanDistance".	7 to >6.0 pts Full Marks	6 to >0.0 pts Partially	0 pts No Marks	7 pts
2.D.ii. Show the clustering information with the standard deviations and the centroids of the clusters.	2 to >1.0 pts Full Marks	1 to >0.0 pts Partially	0 pts No Marks	2 pts
2.E.i. Choose your own distance function and initial cluster assignment method, increase or decrease the number of clusters.	7 to >6.0 pts Full Marks	6 to >0.0 pts Partially	0 pts No Marks	7 pts

Criteria	Ratings			Pts
2.E.ii. Show the clustering information with the standard deviations and the centroids of the clusters.	2 to >1.0 pts Full Marks	1 to >0.0 pts Partially	0 pts No Marks	2 pts
3.A. Import Walmart_baskets_1week.csv file using the read.transactions() with the "single" format (for long format) and save it in a sparse matrix.	7 to >6.0 pts Full Marks	6 to >0.0 pts Partially	0 pts No Marks	7 pts
3.B Inspect the first 15 transactions.	3 to >2.0 pts Full Marks	2 to >0.0 pts Partially	0 pts No Marks	3 pts
3.C. Use the itemFrequencyPlot command to plot the most frequent 15 items in the descending order of transaction frequency in percentage.	5 to >4.0 pts Full Marks	4 to >0.0 pts Partially	0 pts No Marks	5 pts
3.D.i. Use the apriori command to generate about 50 to 100 association rules from the input data. Show the rules in the descending order of their lift values.	10 to >9.0 pts Full Marks	9 to >0.0 pts Partially	0 pts No Marks	10 pts
3.D.ii. Use the apriori command to generate about 100 to 200 association rules from the input data. Show the rules in the descending order of their lift values.	10 to >9.0 pts Full Marks	9 to >0.0 pts Partially	0 pts No Marks	10 pts
What have you learned from building each of these models and the modeling impact of your adjustments to the hyperparameters or dataset? If you were explaining the results of these models to a supervisor what would you say about them? Attempt to do more than just state facts here, interpret the results. Coding is great, interpretation of output is even more important. Discuss each model. Write at least 150 words.	5 to >4.0 pts Full Marks	4 to >0.0 pts Partially	0 pts No Marks	5 pts
Total Points: 100				