Project 2 - Association Rule Mining & Clustering

CSCI 5523 - Introduction to Data Mining
University of Minnesota

Due - March 31, 2021

Instructions and Experiments

Note: Please read the entire project description before you begin. The goal of this project is to analyze the performance of association rule mining algorithms on several synthetic and real-world data sets. This will be done in the following steps:

- First, you will explore the data sets.
- Next, you will perform a series of experiments on which you will be asked to answer a series of questions. For these experiments, you will be running a python Jupyter notebook.
- Compile your answers in the form of a report where answer should be clearly labeled.

Python Jupyter Notebooks

We recommend installing Jupyter using Anaconda as it will also install other regularly used packages for scientific computing and data science. Some pointers to setup Jupyter notebooks on your system:

- Video link - *https://www.youtube.com/watch?v=MvN7Wdh0Juk*
- Medium Link - *https://medium.com/@neuralnets/beginners-quick-guide-for-handling-issues-launching-jupyter-notebook-for-python-using-anaconda-8be3d57a209b*
- Tutorials link - *https://www.dataquest.io/blog/jupyter-notebook-tutorial/*, *https://www.youtube.com/watch?v=3C9E2yPBw7s*

Before you Begin

- Visually explore the data sets in the experiments below, and consider the following:

  types of attributes

  class distribution

  which attributes appear to be good predictors, if any

  possible correlation between attributes

  any special structure that you might observe

  Note: The discussion of this exploration is not required in the report, but this step will help you get ready to answer the questions that follow

- Your goal is to learn everything that you can about the dataset. Answer the questions below as a starting point, but you should dig further. What more can you discover? The goal of this assignment is to give a helping hand for you to discover the most interesting and surprising things.

Report and Submission

- Collect output from your experiments. Submit all Jupyter notebook (cell displaying output) electronically as a single zipped file using the Project 2 **Canvas** submit tool. A submission not adhering to this policy will not be graded and you will get zero.

- Write a report addressing the experiment questions. **Clearly label responses to each problem / sub-problem. The report has to be submitted in PDF format electronically using Project 2 Gradescope submit tool**. Your project will be evaluated based only on what you write on the report.

- If you are a UNITE student, you should upload your Jupyter notebook (cell displaying output) and report (PDF) on canvas like other students.

- Your Jupyter notebook should be submitted electronically - we will look at your output if something is ambiguous in your report. Copy and paste the output from the Jupyter notebook into your report only to the limited extent needed to support your answers.

Package Requirements

- mlxtend (pip install mlxtend)

- squarify (pip install squrify)

- wordcloud (pip install wordcloud)

1 Problem 1 [25 points]

The files for this problem is under Experiment 1 folder. Datasets to be used for experimentation: **store_transaction.csv**. Jupyter notebook: **apriori_analysis.ipynb**. In this experiment, we give a dataset of a store with thousands of transactions of customers buying several items from the store. We will use the apriori algorithm to find correlations between various items in the store. Answer the following question :

1. How many records are there in the dataset?
2. In a single transaction, what is the maximum number of items a customer has bought? We assume that each record is a separate transaction
3. Write any five transactions a customer has done.
4. We use the wordcloud to generate a stunning visualization format to highlight crucial textual data points and convey essential information. Generate and paste the wordcloud with max_words set to 25 and 50. Briefly describe your understanding of the plot.
5. What are the top 5 most frequent items in the dataset?
6. Suppose we have the following transaction data: [['Apple', 'Beer', 'Rice', 'Chicken'], ['Apple', 'Beer', 'Rice'], ['Apple', 'Beer'], ['Apple', 'Bananas'], ['Milk', 'Beer', 'Rice', 'Chicken'], ['Milk', 'Beer', 'Rice'], ['Milk', 'Beer'], ['Apple', 'Bananas']]. Transform this input dataset into a one-hot encoded Boolean array. Hint: In the Jupyter notebook, we use TransactionEncoder to do the same.
7. In the input dataset, how many unique items are present?
8. Run Apriori to generate frequent itemsets at support thresholds of 1%, 2%, 3%, 4%, 5%, 6%, 8%,10%,12%,16% and 20%. In a single figure, for each threshold (X-axis), plot the number of itemsets (Y-axis). Comment on the general trends illustrated by the plots and the reason for the trend.
9. At support threshold 1%, we see frequent itemset of size three along with size 2 and 1. However, at the support threshold of 2%, we observe itemsets of size 1 and 2 only. Why do you think this is so?
10. Run Apriori to generate frequent itemsets of length 2 at support thresholds of 1%, 2%, 3%, 4% and 5%. In a single figure, for each threshold (X-axis), plot the number of itemsets of length 2 (Y-axis). Comment on the general trends illustrated by the plots and the reason for the trend.
11. For the following itemset, write down its corresponding support value:
    - Mineral Water
    - Chocolate
    - Eggs

- Eggs, Mineral Water
- Chocolate, Mineral Water

2 Problem 2 [25 points]

The files for this problem is under Experiment 2 folder. Datasets to be used for experimentation: **instacart_transaction.csv**. Jupyter notebook: **Instacart_association.ipynb**. Instacart, an online grocer, has graciously made some of their datasets accessible to the public (*https://www.instacart.com/datasets/grocery-shopping-2017*). In this experiment, we will use apriori algorithm to find correlations between the different items in the store. Answer the following question:

1. Given following transactions:

   - order 1: apple, egg, milk
   - order 2: carrot, milk
   - order 3: apple, egg, carrot
   - order 4: apple, egg
   - order 5: apple, carrot

   Using the apriori algorithm, write down the pair of items having a minimum threshold of 3. Briefly describe your steps.

2. How many unique orders and unique items are there in the dataset? Are unique order same as number of records in the dataset? On average, per order, how many items does a customer order?

3. Run apriori to generate pairs of itemset at support thresholds of 1%, 2%, 4%, 6%, 8%, and 10%. In a single figure, for each threshold (X-axis), plot the number of association rules (Y-axis). Comment on the trend of algorithm runtime at different thresholds.

4. Run apriori at support thresholds of 1%, 2%, 4%, 6%, 8% and 10%. For each threshold, write a pair of association rules (you can choose any) along with its key metrics (i.e. freqAB, supportAB, freqA,supportA, freqB, support, confidenceAtoB, confidenceBtoA, lift). As a data scientist for the retailer giant, after observing the association rules, what would you do to increase the sales.

3 Problem 3 [20 points]

The files for this problem is under Experiment 3 folder. Datasets to be used for experimentation: **2d_data, chameleon, elliptical, and vertebrate**. Jupyter notebook: **cluster_analysis.ipynb**. Cluster analysis seeks to partition the input data into groups of closely related instances so that instances that belong to the same cluster are more similar to each other than to instances that belong to other clusters. In this experiment, we provide

examples of using different clustering techniques provided by the scikit-learn library package. Answer the following question :

1. In the notebook, k-mean clustering assign users to two clusters i.e., cluster one has a higher rating for action movies, and cluster two has higher ratings for horror movies. Given the cluster centroid, assign the following users to their respective cluster assignment:

User	Exorcist	Omen	Star Wars	Jaws
Paul	4	5	2	4
Adel	1	2	3	4
Kevin	2	3	5	5
Jessi	1	1	3	2

2. A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data can be clustered. The Elbow Method is one of the most popular methods to find the optimal value of k. For the toy example of the movie rating dataset, what is the optimal value of K? Briefly explain your reasoning.

3. On the Vertebrate dataset, we illustrate the results of using three hierarchical clustering algorithms (1) single link (MIN), (2) complete link (MAX), and (3) group average. As a data scientist, given the class of the original dataset, which clustering algorithm makes more sense. Explain your reasoning.

4. For DBSCAN, how many clusters are formed when the minimum number of points (min_samples) to set to 1, 2, 3, 4, and 5, respectively. For each instance, copy and paste the plot of the clusters.

5. For elliptical and 2D data, we apply k-means and spectral clustering with the number of clusters(k) = 2. Repeat the same set of clustering for k = 4 and copy and paste the clusters formed. Which cluster method performs well when k =2 and k = 4.

4 Problem 4 [30 points]

The files for this problem is under Experiment 4 folder. Jupyter notebook: **covid-19-research-challenge.ipynb**. In this experiment, given the large amount of academic literature surrounding COVID-19, you will help overloaded scientists to keep up with the research happening all around the globe for faster development of the vaccine. Given that we have recently studied clustering, can we cluster similar research articles together to make it easier for health professionals to find relevant research articles? Clustering can be used to create a tool to identify related articles, given a target article. **Dataset Description**: In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 29,000 scholarly articles, including over 13,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global

research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in modern coronavirus literature, making it difficult for the medical research community to keep up.

1. After we handle duplicates, what is the count, mean, standard deviation minimum, and maximum values for abstract word count and body word count?

2. Briefly describe the data pre-processing steps done in the notebook for cleaning up the text.

3. For clustering, to create a feature vector, on what part of the article did we focus?

4. What is N-gram in machine learning? Given the following word list: ['the', '2019', 'novel', 'coronavirus', 'sarscov2', 'identified', 'as', 'the', 'cause', 'of'], what is its 2-gram ?

5. What does HashingVectorizer do? What is the feature size of HashingVector that we used in our analysis?

6. We have randomly chosen 10 clusters using k-means clustering, vectorized using hashingVector, which makes some sense if we plot the t-SNE plot as articles from the same cluster are near each other, forming groups. However, there are still overlaps. Can you improve this by changing the cluster size or choosing a different feature size? Give the size of the cluster and the feature size that makes more sense for you. Copy and paste the corresponding t-SNE plot.

7. We have randomly chosen 10 clusters using k-means clustering, vectorized using tf-idf, and we can see clusters more clearly. Can you improve this by changing the cluster size or changing the max_features value of TfidfVectorizer? Give the size of the cluster and the max_features value that makes more sense for you. Copy and paste the corresponding t-SNE plot.

8. In the interactive t-SNE with 20 clusters, can you do a manual analysis of each cluster to see what articles cluster together? Choose any 5 clusters and write 4-5 keywords that describe it. Hover your mouse over the cluster point, and you can see the article that it refers. You can moreover choose to display points of one cluster only in the plot. Also, name the clusters that include articles involving the social and economic impacts of the coronavirus?