CS 484: Introduction to Machine Learning

Spring Semester 2023 Assignment 2

# Question 1 (20 points)

I invited six friends to watch a basketball game at home. They brought the following items along.

|  |  |
| --- | --- |
| Friend | Items |
| Andrew | Cheese, Cracker, Soda, Wing |
| Betty | Cheese, Soda, Tortilla, Wing |
| Carl | Cheese, Ice Cream, Wing |
| Danny | Cheese, Ice Cream, Salsa, Soda, Tortilla |
| Emily | Salsa, Soda, Tortilla, Wing |
| Frank | Cheese, Cracker, Ice Cream, Wing |

I noticed that my friends often brought Cheese, Soda, and Wing together. Since I prefer to spend on food instead of Soda, I study how likely my friends would bring Soda if they already bought Cheese and Wing. Therefore, please calculate the Lift of this association rule {Cheese, Wing} ==> {Soda} for me.

# Question 2 (40 points)

This question walks you through the typical process of discovering association rules. We will use the market basket data in the **Groceries.csv** file to discover association rules. Here are the data contents.

1. Customer: Customer Identifier
2. Item: Name of Product Purchased

For your information, we have sorted the observations in ascending order first by Customer and then by Item. Also, we have removed duplicated items for each customer.

1. (10 points) What is the number of items in the Universal Set? What is the maximum number of itemsets that we can find in theory from the data? What is the maximum number of association rules that we can generate in theory from the data?
2. (10 points) We are interested in the itemsets that can be found in the market baskets of at least seventy-five (75) customers. How many itemsets did we find? Also, what is the largest number of items, i.e., , among these itemsets?
3. (10 points) We will use up to the largest value we found in Part (b) and then generate the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Next, we plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for these association rules. We will use the Lift metrics to indicate the size of the marker. We will add a color gradient legend to the chart for the Lift metrics.
4. (10 points) Among the rules that you found in Part (c), list the rules whose Confidence metrics are greater than or equal to 60%. Please show the rules in a table that shows the Antecedent, the Consequent, the Support, the Confidence, the Expected Confidence, and the Lift.

# Question 3 (40 points)

This question demonstrates the effect of rescaling input variables on the cluster results. We will discover clusters using all the observations in the **TwoFeatures.csv** file with the following specifications.

* The input interval variables are x1 and x2
* The metric is the Manhattan distance
* The minimum number of clusters is 1
* The maximum number of clusters is 8
* Use the Elbow value for choosing the optimal number of clusters

Since the sklearn.cluster.KMeans class works only with the Euclidean distance, you will need to develop custom Python codes to implement the K-Means algorithm with the Manhattan distance.

1. (10 points) Plot x2 (vertical axis) versus x1 (horizontal axis). Add gridlines to both axes. Let the graph engine chooses the tick marks. How many clusters do you see in the graph?
2. (10 points) Discover the optimal number of clusters without any transformations. List the number of clusters, the Total Within-Cluster Sum of Squares (TWCSS), and the Elbow values in a table. Plot the Elbow Values versus the number of clusters. How many clusters do you find? What are the centroids of your optimal clusters?
3. (10 points) Linearly rescale x1 such that the resulting variable has a minimum of zero and a maximum of ten. Likewise, rescale x2. Discover the optimal number of clusters from the transformed observations. List the number of clusters, the Total Within-Cluster Sum of Squares (TWCSS), and the Elbow values in a table. Plot the Elbow Values versus the number of clusters. How many clusters do you find? What are the centroids of your optimal clusters in the original scale of x1 and x2?
4. (10 points) If you are doing everything correctly, you should discover two different optimal cluster solutions. In your words, how do you explain the difference?