CS 484: Introduction to Machine Learning

Autumn 2020 Assignment 2

# Question 1 (35 points)

The file Groceries.csv contains market basket data. The variables are:

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

After you have imported the CSV file, please discover association rules using this dataset. For your information, the observations have been sorted in ascending order by Customer and then by Item. Also, duplicated items for each customer have been removed.

1. (5 points) Create a data frame that contains the number of unique items in each customer’s market basket. Draw a histogram of the number of unique items. What are the 25th, 50th, and the 75th percentiles of the histogram?
2. (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy five (75) customers. How many itemsets can we find? Also, what is the largest *k* value among our itemsets?
3. (10 points) Find out the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent. Please **do not** display those rules in your answer.
4. (5 points) Plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you have found in (c). Please use the Lift metrics to indicate the size of the marker.

1. (5 points) List the rules whose Confidence metrics are greater than or equal to 60%. Please include their Support and Lift metrics.

# Question 2 (10 points)

In the breakfast café example in Chapter 4 of *A Practitioner's Guide to Machine Learning*, you will take the three most frequent categories as the initial centroids. What is your final cluster solution? How does this final cluster solution compare with that of the example?

# Question 3 (10 points)

Let denote the matrix elements of the Adjacency matrix. Suppose is an eigenvalue and the corresponding eigenvector of the Laplacian matrix. Prove mathematically that where denote the elements of the eigenvector If possible, please type your answer using an Equation Editor (Press **Alt** and **=** together in Microsoft Word) or similar application.

# Question 4 (10 points)

Suppose Cluster 0 contains observations {-2, -1, 1, 2, 3} and Cluster 1 contains observations {4, 5, 7, 8}.

1. (4 points) Calculate the Silhouette Width of the observation 2 (i.e., the value -1) in Cluster 0.
2. (4 points) Calculate the cluster-wise Davies-Bouldin value of Cluster 0 (i.e., ) and Cluster 1 (i.e., ).
3. (2 points) What is the Davies-Bouldin Index of this two-cluster solution?

# Question 5 (35 points)

Apply the Spectral Clustering method to the FourCircle.csv. Your input fields are and . Wherever needed, specify random\_state = None in calling the KMeans function.

1. (5 points) Plot on the vertical axis versus on the horizontal axis. Based on your visual inspection, how many connected clusters are there?
2. (5 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifiers to control the color scheme. Please comment on this K-mean result.
3. (10 points) Apply the nearest neighbor algorithm using the Euclidean distance. We will consider the number of neighbors from 1 to 15. What is the smallest number of neighbors that we should use to discover the clusters correctly? Remember that you may need to iterate between parts c), d), and e) a couple of values and use the eigenvalue plot to validate your choice.
4. (5 points) Using your choice of the number of neighbors in (c), calculate the Adjacency matrix, the Degree matrix, and finally the Laplacian matrix. How many eigenvalues do you determine are practically zero? Please display values of the “zero” eigenvalues in scientific notation.
5. (10 points) Apply the K-mean algorithm on the eigenvectors that correspond to your “practically” zero eigenvalues. The number of clusters is the number of your “practically” zero eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme.