CS 584-04: Machine Learning

Fall 2018 Assignment 2

# INSTRUCTIONS

1. Students should complete this assignment independently.
2. Students are encouraged to use Python to complete this assignment.
3. Students must submit their answers to Blackboard before 12:00 Noon on September 22, 2018.

# Question 1 (20 points)

Suppose a market basket can possibly contain these seven items: A, B, C, D, E, F, and G.

1. (1 point) What is the number of possible itemsets?
2. (3 points) List all the possible 1-itemsets.
3. (3 points) List all the possible 2-itemsets.
4. (3 points) List all the possible 3-itemsets.
5. (3 points) List all the possible 4-itemsets.
6. (3 points) List all the possible 5-itemsets.
7. (3 points) List all the possible 6-itemsets.
8. (1 point) List all the possible 7-itemsets.

# Question 2 (30 points)

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

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

The data is already sorted in ascending order by Customer and then by Item. Also, all the items bought by each customer are all distinct.

After you have imported the CSV file, please discover association rules using this dataset.

1. (2 points) How many customers in this market basket data?
2. (2 points) How many unique items in the market basket across all customers?
3. (5 points) Create a dataset which contains the number of distinct items in each customer’s market basket. Draw a histogram of the number of unique items. What are the median, the 25th percentile and the 75th percentile in this histogram?
4. (5 points) Find out the *k*-itemsets which appeared in the market baskets of at least seventy five (75) customers. How many itemsets have you found? Also, what is the highest *k* value in your itemsets?
5. (5 points) Find out the association rules whose Confidence metrics are at least 1%. How many association rules have you found? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent.
6. (5 points) Graph the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (e). Please use the Lift metrics to indicate the size of the marker.
7. (5 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.
8. (1 point) What similarities do you find among the consequents that appeared in (g)?

# Question 3 (20 points)

You are asked to write a Python program to calculate the Elbow value and the Silhouette value. For this question, you will use the CARS.CSV dataset to test your program. Here are the specifications for performing the respective analyses.

**Clustering**

* The input interval variables are Horsepower and Weight
* The distance metric is Euclidean
* The maximum number of clusters is 15
* Consider the silhouette\_score function for calculating the Silhouette value.

Please answer the following questions.

1. (15 points) List the Elbow values and the Silhouette values for your 1-cluster to 15-cluster solutions.
2. (5 points) Based on the Elbow values and the Silhouette values, what do you suggest for the number of clusters?

# Question 4 (30 points)

Apply the Spectral Clustering method to the Spiral.csv. Your input fields are x and y.

1. (5 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?
2. (5 points) Apply the K-mean algorithm directly using your number of clusters (in a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?
3. (5 points) Apply the nearest neighbor algorithm using the Euclidean distance. How many nearest neighbors will you use?
4. (5 points) Generate the sequence plot of the first nine eigenvalues, starting from the smallest eigenvalues. Based on this graph, do you think your number of nearest neighbors (in a) is appropriate?
5. (5 points) Apply the K-mean algorithm on your first two eigenvectors that correspond to the first two smallest eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?
6. (5 points) Comment on your spectral clustering results?