CS 584: Machine Learning

Pradyot Mayank(A20405826)
Spring 2019 Assignment 2

Question 1 (20 points)

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

a) (1 point) What is the number of possible itemsets?
Given Items: ['A', 'B', 'C', 'D', 'E', 'F', 'G']
Number of Items: 7
Number of possible itemset: 2^m-1=2^7-1=127

b) (3 points) List all the possible 1-itemsets.
1.('A') 2. ('B') 3. ('C') 4. ('D') 5. ('E') 6. ('F') 7. ('G')
No. of possible 1-itemsets=7

c) (3 points) List all the possible 2-itemsets.
1. ('A', 'B') 2. ('A', 'C') 3. ('A', 'D') 4. ('A', 'E') 5. ('A', 'F') 6. ('A', 'G') 7. ('B', 'C') 8. ('B', 'D') 9. ('B', 'E') 10. ('B', 'F') 11. ('B', 'G') 12. ('C', 'D') 13. ('C', 'E') 14. ('C', 'F') 15. ('C', 'G') 16. ('D', 'E') 17. ('D', 'F') 18. ('D', 'G') 19. ('E', 'F') 20. ('E', 'G') 21. ('F', 'G')
No. of possible 2-itemsets=21

d) (3 points) List all the possible 3-itemsets.
1. ('A', 'B', 'C') 2. ('A', 'B', 'D') 3. ('A', 'B', 'E') 4. ('A', 'B', 'F') 5. ('A', 'B', 'G') 6. ('A', 'C', 'D') 7. ('A', 'C', 'E') 8. ('A', 'C', 'F') 9. ('A', 'C', 'G') 10. ('A', 'D', 'E') 11. ('A', 'D', 'F') 12. ('A', 'D', 'G') 13. ('A', 'E', 'F') 14. ('A', 'E', 'G') 15. ('A', 'F', 'G') 16. ('B', 'C', 'D') 17. ('B', 'C', 'E') 18. ('B', 'C', 'F') 19. ('B', 'C', 'G') 20. ('B', 'D', 'E') 21. ('B', 'D', 'F') 22. ('B', 'D', 'G') 23. ('B', 'E', 'F') 24. ('B', 'E', 'G') 25. ('B', 'F', 'G') 26. ('C', 'D', 'E') 27. ('C', 'D', 'F') 28. ('C', 'D', 'G') 29. ('C', 'E', 'F') 30. ('C', 'E', 'G') 31. ('C', 'F', 'G') 32. ('D', 'E', 'F') 33. ('D', 'E', 'G') 34. ('D', 'F', 'G') 35. ('E', 'F', 'G')
No. of possible 3-itemsets=35

e) (3 points) List all the possible 4-itemsets.
1. ('A', 'B', 'C', 'D') 2. ('A', 'B', 'C', 'E') 3. ('A', 'B', 'C', 'F') 4. ('A', 'B', 'C', 'G') 5. ('A', 'B', 'D', 'E') 6. ('A', 'B', 'D', 'F') 7. ('A', 'B', 'D', 'G') 8. ('A', 'B', 'E', 'F') 9. ('A', 'B', 'E', 'G') 10. ('A', 'B', 'F', 'G') 11. ('A', 'C', 'D', 'E') 12. ('A', 'C', 'D', 'F') 13. ('A', 'C', 'D', 'G') 14. ('A', 'C', 'E', 'F') 15. ('A', 'C', 'E', 'G') 16. ('A', 'C', 'F', 'G') 17. ('A', 'D', 'E', 'F') 18. ('A', 'D', 'E', 'G') 19. ('A', 'D', 'F', 'G') 20. ('A', 'E', 'F', 'G') 21. ('B', 'C', 'D', 'E') 22. ('B', 'C', 'D', 'F') 23. ('B', 'C', 'D', 'G') 24. ('B', 'C', 'E', 'F') 25. ('B', 'C', 'E', 'G') 26. ('B', 'C', 'F', 'G') 27. ('B', 'D', 'E', 'F') 28. ('B', 'D', 'E', 'G') 29. ('B', 'D', 'F', 'G') 30. ('B', 'E', 'F', 'G') 31. ('C', 'D', 'E', 'F') 32. ('C', 'D', 'E', 'G') 33. ('C', 'D', 'F', 'G') 34. ('C', 'E', 'F', 'G') 35. ('D', 'E', 'F', 'G')
No. of possible 4-itemsets=35

f) (3 points) List all the possible 5-itemsets.
1. ('A', 'B', 'C', 'D', 'E') 2. ('A', 'B', 'C', 'D', 'F') 3. ('A', 'B', 'C', 'D', 'G') 4. ('A', 'B', 'C', 'E', 'F') 5. ('A', 'B', 'C', 'E', 'G') 6. ('A', 'B', 'C', 'F', 'G') 7. ('A', 'B', 'D', 'E', 'F') 8. ('A', 'B', 'D', 'E', 'G') 9. ('A', 'B', 'D', 'F', 'G') 10. ('A', 'B', 'E', 'F', 'G') 11. ('A', 'C', 'D', 'E', 'F') 12. ('A', 'C', 'D', 'E', 'G') 13. ('A', 'C', 'D', 'F', 'G') 14. ('A', 'C', 'E', 'F', 'G') 15. ('A', 'D', 'E', 'F', 'G') 16. ('B', 'C', 'D', 'E', 'F') 17. ('B', 'C', 'D', 'E', 'G') 18. ('B', 'C', 'D', 'F', 'G') 19. ('B', 'C', 'E', 'F', 'G') 20. ('B', 'D', 'E', 'F', 'G') 21. ('C', 'D', 'E', 'F', 'G')
No. of possible 5-itemsets=21

g) (3 points) List all the possible 6-itemsets.
1. ('A', 'B', 'C', 'D', 'E', 'F') 2. ('A', 'B', 'C', 'D', 'E', 'G') 3. ('A', 'B', 'C', 'D', 'F', 'G') 4. ('A', 'B', 'C', 'E', 'F', 'G') 5. ('A', 'B', 'D', 'E', 'F', 'G') 6. ('A', 'C', 'D', 'E', 'F', 'G') 7. ('B', 'C', 'D', 'E', 'F', 'G')
No. of possible 6-itemsets=7

h) (1 point) List all the possible 7-itemsets.
1. ('A', 'B', 'C', 'D', 'E', 'F', 'G')
No. of possible 7-itemsets=1

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

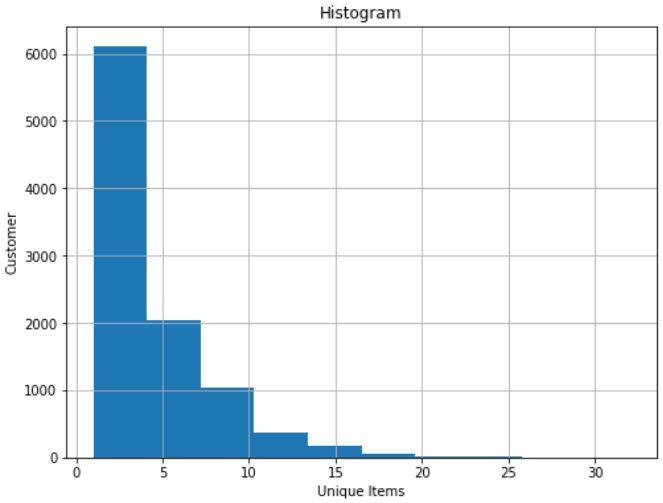
The data is already sorted in ascending order by Customer and then by Item. Also, all the items bought by each customer are all distinct.

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

a) (2 points) How many customers in this market basket data?
Unique Customer: [1 2 3 ... 9833 9834 9835]
Total number of Customer: 9835

b) (2 points) How many unique items in the market basket across all customers?
Total number of Unique Items: 169

c) (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?



Median in the histogram: 15.0
25th percentile in the histogram: 8.0
75th percentile in the histogram: 22.0

d) (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?

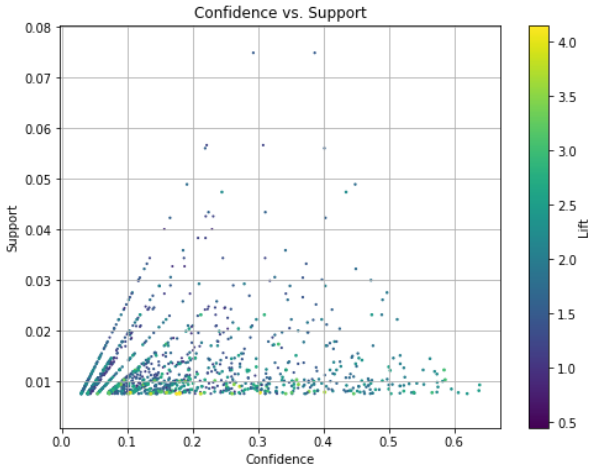
Total number of itemset: 524

The highest value of k in the itemset: 4

e) (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.

Total number of association rules: 1228

f) (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.



g) (5 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.

Antecents	Consequents	Confidence	Support	Lift
(root vegetables, butter)	(whole milk)	0.637795	0.00823589	2.49611
(yogurt, butter)	(whole milk)	0.638889	0.00935435	2.50039
(other vegetables, yogurt, root vegetables)	(whole milk)	0.606299	0.00782918	2.37284
(other vegetables, yogurt, tropical fruit)	(whole milk)	0.619835	0.00762583	2.42582

h) (1 point) What similarities do you find among the consequents that appeared in (g)?

Consequents (i.e whole milk) are the same for all the antecedents.

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.

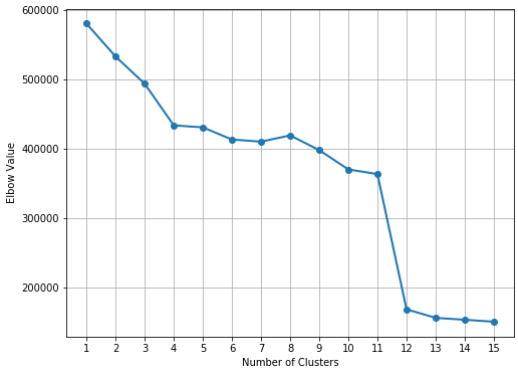
Clustering

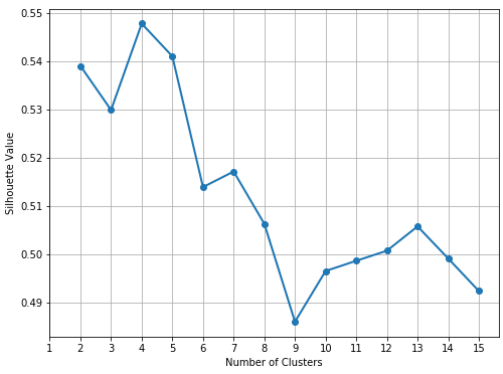
- The input interval variables are Horsepower and Weight
- The distance metric is Euclidean
- The maximum number of clusters is 15
- Consider the `silhouette_score` function for calculating the Silhouette value
- Specify random_state = 60616 in calling the KMeans function

Please answer the following questions.

a) (15 points) List the Elbow values and the Silhouette values for your 1-cluster to 15-cluster solutions.

No of Clusters	Elbow Value	Silhouette Value
1	579857.954	nan
2	532455.272	0.539
3	493218.081	0.530
4	433215.815	0.548
5	430290.457	0.541
6	412804.931	0.514
7	409729.742	0.517
8	418744.248	0.506
9	397493.532	0.486
10	369702.705	0.497
11	362959.003	0.499
12	168058.092	0.501
13	155749.416	0.506
14	153006.554	0.499
15	150220.900	0.492





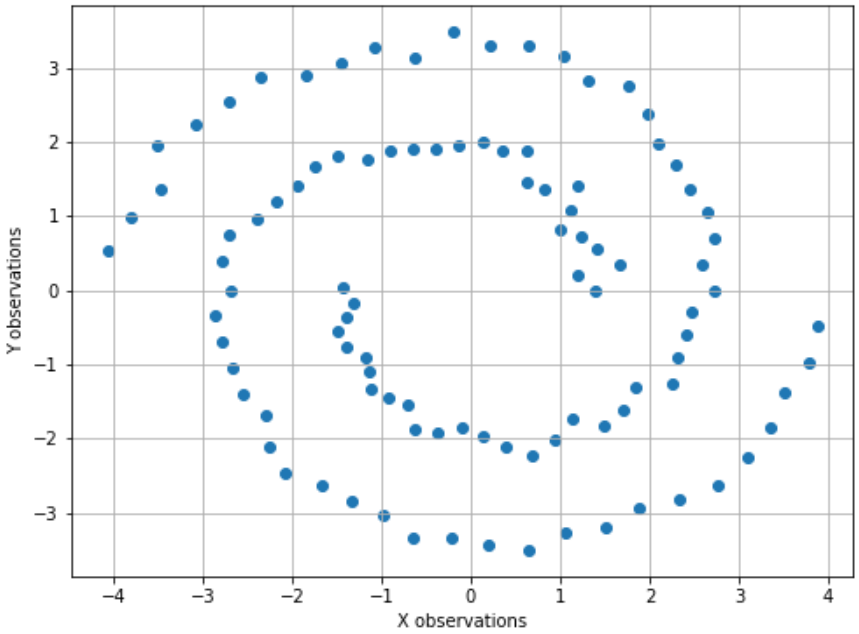
b) (5 points) Based on the Elbow values and the Silhouette values, what do you suggest for the number of clusters?

Based on the Elbow and Silhouette Values the optimum number of clusters is 4 as we can see the Silhouette Value is the maximum when the no. of cluster is 4 and an Elbow is also formed at that point.

Question 4 (30 points)

Apply the Spectral Clustering method to the Spiral.csv. Your input fields are x and y. Wherever needed, specify random_state = 60616 in calling the KMeans function.

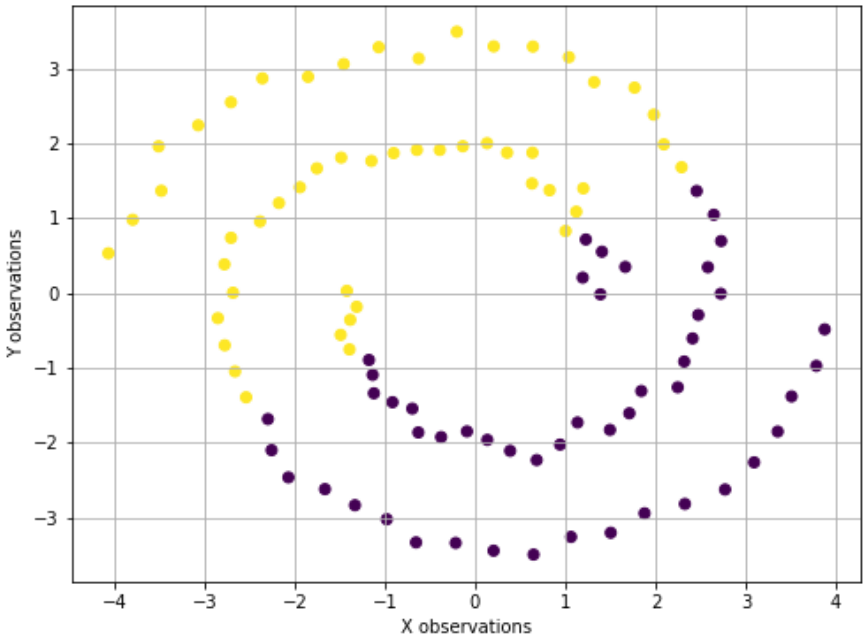
a) (5 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?



By observing or visualizing the graph we can say that there are 2 clusters.

Number of Clusters: 2

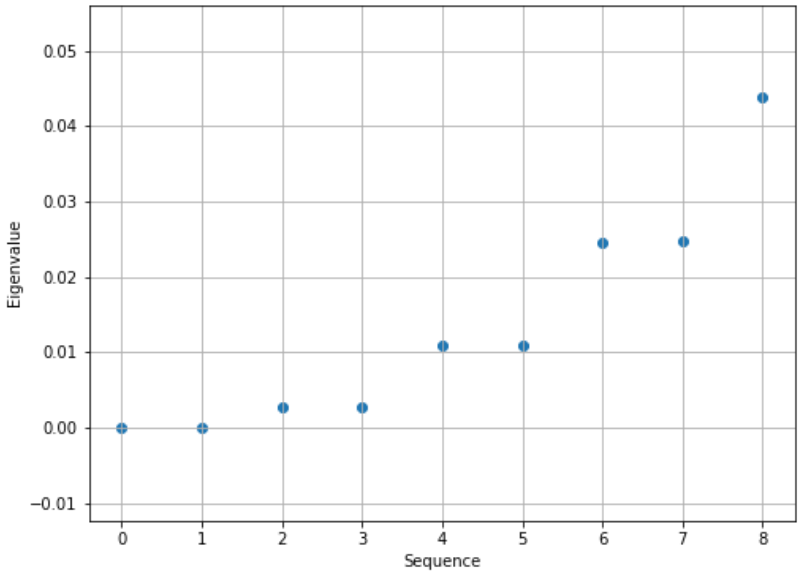
b) (5 points) Apply the K-mean algorithm directly using your number of clusters (in a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?



c) (5 points) Apply the nearest neighbor algorithm using the Euclidean distance. How many nearest neighbors will you use?

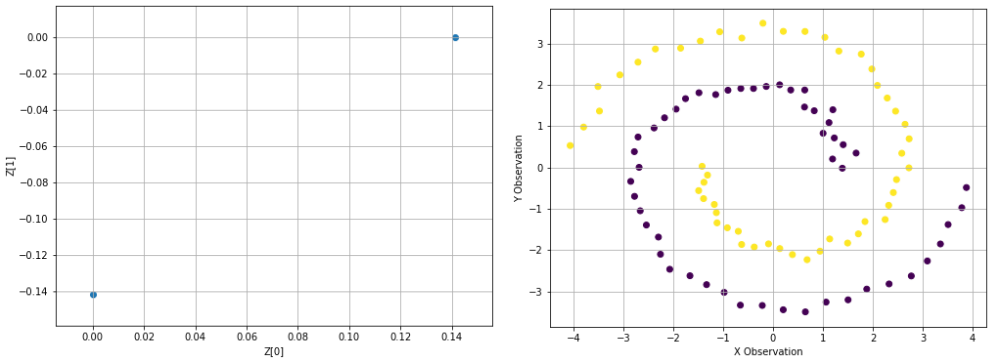
We will use 3 nearest neighbors.

d) (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 c) is appropriate?



The above graph confirms that the three nearest neighbors' solution is appropriate.

e) (5 points) Apply the K-mean algorithm on your first <u>two</u> eigenvectors that correspond to the first two smallest eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?



f) (5 points) Comment on your spectral clustering results?

Based on the observation from the Eigenvalue graph, the first jump happens at 4 but when we choose the nearest neighbors 4 the clustering was incorrect, so we choose the nearest value less than 4 which is 3, and we can see the spectral clustering is correct.