CS 584-04: Machine Learning

Fall 2018 Assignment 2

Diego Martin Crespo A20432558

Question 1 (20 points)

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

We import the “itertools” library to get the results. Knowing that there are 7 items:

a) (1 point) What is the number of possible itemsets?
There are 127 possible itemsets (2^7-1).

b) (3 points) List all the possible 1-itemsets.
There are 7 possible 1-itemsets: [('A',), ('B',), ('C',), ('D',), ('E',), ('F',), ('G',)]

c) (3 points) List all the possible 2-itemsets.
There are 21 possible 2-itemsets : [('A', 'B'), ('A', 'C'), ('A', 'D'), ('A', 'E'), ('A', 'F'), ('A', 'G'), ('B', 'C'), ('B', 'D'), ('B', 'E'), ('B', 'F'), ('B', 'G'), ('C', 'D'), ('C', 'E'), ('C', 'F'), ('C', 'G'), ('D', 'E'), ('D', 'F'), ('D', 'G'), ('E', 'F'), ('E', 'G'), ('F', 'G')].

d) (3 points) List all the possible 3-itemsets.
There are 35 possible 3-itemsets: [('A', 'B', 'C'), ('A', 'B', 'D'), ('A', 'B', 'E'), ('A', 'B', 'F'), ('A', 'B', 'G'), ('A', 'C', 'D'), ('A', 'C', 'E'), ('A', 'C', 'F'), ('A', 'C', 'G'), ('A', 'D', 'E'), ('A', 'D', 'F'), ('A', 'D', 'G'), ('A', 'E', 'F'), ('A', 'E', 'G'), ('A', 'F', 'G'), ('B', 'C', 'D'), ('B', 'C', 'E'), ('B', 'C', 'F'), ('B', 'C', 'G'), ('B', 'D', 'E'), ('B', 'D', 'F'), ('B', 'D', 'G'), ('B', 'E', 'F'), ('B', 'E', 'G'), ('B', 'F', 'G'), ('C', 'D', 'E'), ('C', 'D', 'F'), ('C', 'D', 'G'), ('C', 'E', 'F'), ('C', 'E', 'G'), ('C', 'F', 'G'), ('D', 'E', 'F'), ('D', 'E', 'G'), ('D', 'F', 'G'), ('E', 'F', 'G')].

e) (3 points) List all the possible 4-itemsets.
There are 35 possible 4-itemsets: [('A', 'B', 'C', 'D'), ('A', 'B', 'C', 'E'), ('A', 'B', 'C', 'F'), ('A', 'B', 'C', 'G'), ('A', 'B', 'D', 'E'), ('A', 'B', 'D', 'F'), ('A', 'B', 'D', 'G'), ('A', 'B', 'E', 'F'), ('A', 'B', 'E', 'G'), ('A', 'B', 'F', 'G'), ('A', 'C', 'D', 'E'), ('A', 'C', 'D', 'F'), ('A', 'C', 'D', 'G'), ('A', 'C', 'E', 'F'), ('A', 'C', 'E', 'G'), ('A', 'C', 'F', 'G'), ('A', 'D', 'E', 'F'), ('A', 'D', 'E', 'G'), ('A', 'D', 'F', 'G'), ('A', 'E', 'F', 'G'), ('B', 'C', 'D', 'E'), ('B', 'C', 'D', 'F'), ('B', 'C', 'D', 'G'), ('B', 'C', 'E', 'F'), ('B', 'C', 'E', 'G'), ('B', 'C', 'F', 'G'), ('B', 'D', 'E', 'F'), ('B', 'D', 'E', 'G'), ('B', 'D', 'F', 'G'), ('B', 'E', 'F', 'G'), ('C', 'D', 'E', 'F'), ('C', 'D', 'E', 'G'), ('C', 'D', 'F', 'G'), ('C', 'E', 'F', 'G'), ('D', 'E', 'F', 'G')].

f) (3 points) List all the possible 5-itemsets.
There are 21 possible 5-itemsets: [('A', 'B', 'C', 'D', 'E'), ('A', 'B', 'C', 'D', 'F'), ('A', 'B', 'C', 'D', 'G'), ('A', 'B', 'C', 'E', 'F'), ('A', 'B', 'C', 'E', 'G'), ('A', 'B', 'C', 'F', 'G'), ('A', 'B', 'D', 'E', 'F'), ('A', 'B', 'D', 'E', 'G'), ('A', 'B', 'D', 'F', 'G'), ('A', 'B', 'E', 'F', 'G'), ('A', 'C', 'D', 'E', 'F'), ('A', 'C', 'D', 'E', 'G'), ('A', 'C', 'D', 'F', 'G'), ('A', 'C', 'E', 'F', 'G'), ('A', 'D', 'E', 'F', 'G'), ('B', 'C', 'D', 'E', 'F'), ('B', 'C', 'D', 'E', 'G'), ('B', 'C', 'D', 'F', 'G'), ('B', 'C', 'E', 'F', 'G'), ('B', 'D', 'E', 'F', 'G'), ('C', 'D', 'E', 'F', 'G')].

g) (3 points) List all the possible 6-itemsets.
There are 7 possible 6-itemsets: [('A', 'B', 'C', 'D', 'E', 'F'), ('A', 'B', 'C', 'D', 'E', 'G'), ('A', 'B', 'C', 'D', 'F', 'G'), ('A', 'B', 'C', 'E', 'F', 'G'), ('A', 'B', 'D', 'E', 'F', 'G'), ('A', 'C', 'D', 'E', 'F', 'G'), ('B', 'C', 'D', 'E', 'F', 'G')].

h) (1 point) List all the possible 7-itemsets.
There is 1 possible 7-itemsets: [('A', 'B', 'C', 'D', 'E', 'F', 'G')].

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

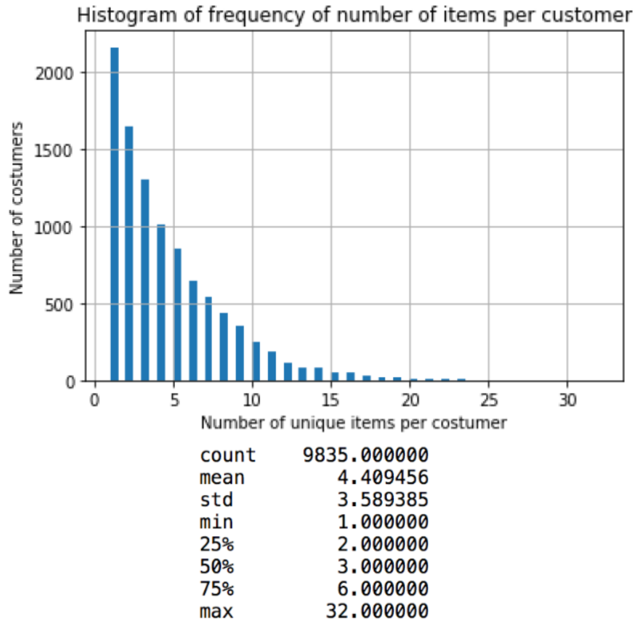
The data is already sorted in ascending order by Customer and then by Item. Also, all the items bought by each customer are all distinct.

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?
There are 9835 of customers. As the data set is sorted in ascending by customer, the number of customers will be the number associated to the last ordered customer.

b) (2 points) How many unique items in the market basket across all customers?
There are 169 of unique items.

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?

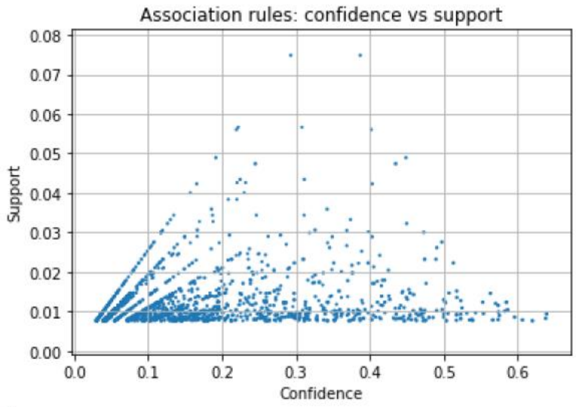


```
count    9835.000000
mean        4.409456
std         3.589385
min         1.000000
25%         2.000000
50%         3.000000
75%         6.000000
max        32.000000
```

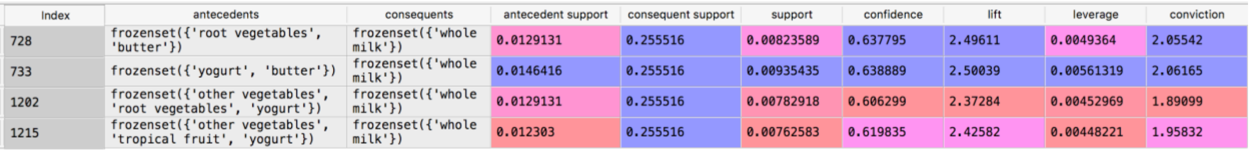
d) (5 points) Find out the *k*-itemsets which appeared in the market baskets of at least seventy five (75) customers. How many itemsets have you found? Also, what is the highest *k* value in your itemsets?
Using the apriory function with a "min_support" = 75/len(item per costumer).There are 524 possible itemsets. The highest k value is 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.
Using the association_rules function with a "min_threshold" = 0.01. There are 1228 association rules.

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.

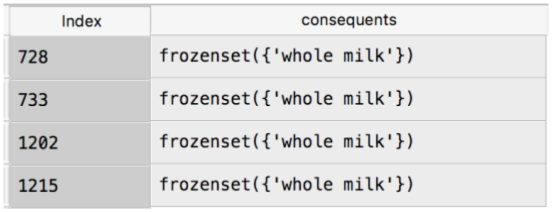


g) (5 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.



Index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
728	frozenset({'root vegetables', 'butter'})	frozenset({'whole milk'})	0.0129131	0.255516	0.00823589	0.637795	2.49611	0.0049364	2.05542
733	frozenset({'yogurt', 'butter'})	frozenset({'whole milk'})	0.0146416	0.255516	0.00935435	0.638889	2.50039	0.00561319	2.06165
1202	frozenset({'other vegetables', 'root vegetables', 'yogurt'})	frozenset({'whole milk'})	0.0129131	0.255516	0.00782918	0.606299	2.37284	0.00452969	1.89099
1215	frozenset({'other vegetables', 'tropical fruit', 'yogurt'})	frozenset({'whole milk'})	0.012303	0.255516	0.00762583	0.619835	2.42582	0.00448221	1.95832

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



Index	consequents
728	frozenset({'whole milk'})
733	frozenset({'whole milk'})
1202	frozenset({'whole milk'})
1215	frozenset({'whole milk'})

There are all the same 1-itemset “whole milk”.

Question 3 (20 points)

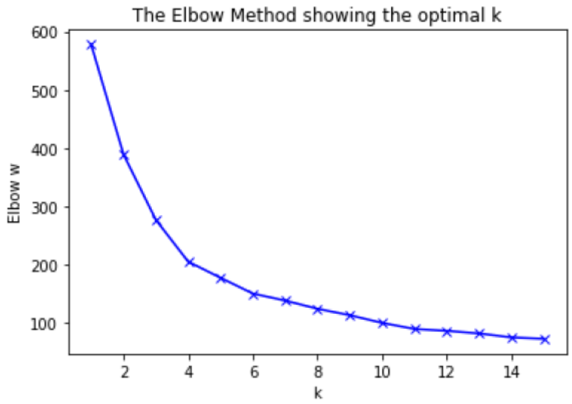
You are asked to write a Python program to calculate the Elbow value and the Silhouette value. For this question, you will use the CARS.CSV dataset to test your program. Here are the specifications for performing the respective analyses.

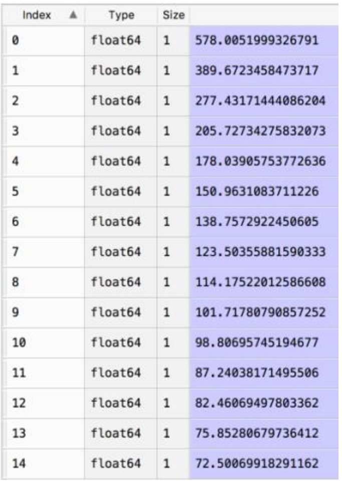
Clustering

- 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.

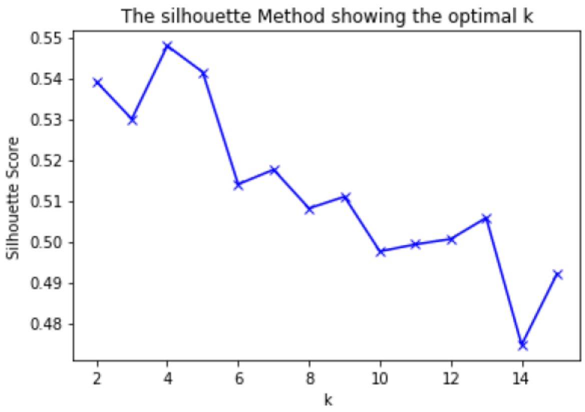
a) (15 points) List the Elbow values and the Silhouette values for your 1-cluster to 15-cluster solutions.

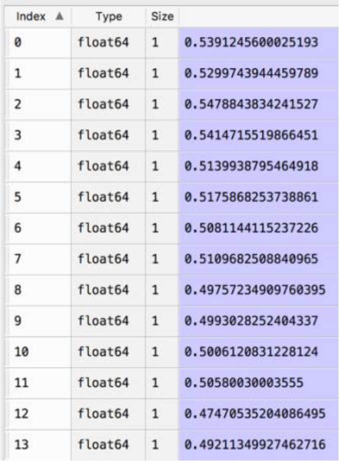




Index ▲	Type	Size	
0	float64	1	578.0051999326791
1	float64	1	389.6723458473717
2	float64	1	277.4317144086204
3	float64	1	205.7273427583207З
4	float64	1	178.0390575377263б
5	float64	1	150.9631083711226
6	float64	1	138.7572922450605
7	float64	1	123.50355881590333
8	float64	1	114.17522012586608
9	float64	1	101.71780790857252
10	float64	1	98.80695745194677
11	float64	1	87.24038171495506
12	float64	1	82.46069497803362
13	float64	1	75.85280679736412
14	float64	1	72.50069918291162

The index in the right previous picture correspond to the number of clusters – 1.





Index ▲	Type	Size	
0	float64	1	0.5391245600025193
1	float64	1	0.5299743944459789
2	float64	1	0.5478843834241527
3	float64	1	0.5414715519866451
4	float64	1	0.5139938795464918
5	float64	1	0.5175868253738861
6	float64	1	0.5081144115237226
7	float64	1	0.5109682508840965
8	float64	1	0.49757234909760395
9	float64	1	0.4993028252404337
10	float64	1	0.5006120831228124
11	float64	1	0.5058003000355
12	float64	1	0.47470535204086495
13	float64	1	0.49211349927462716

The index in the right previous picture correspond to the number of clusters – 2.

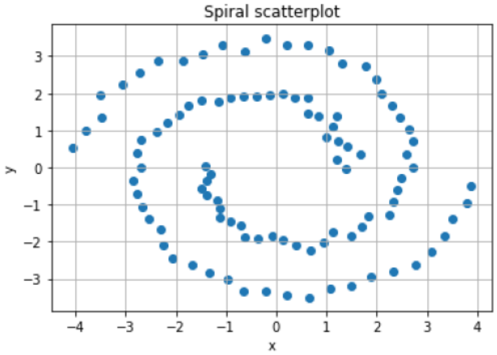
b) (5 points) Based on the Elbow values and the Silhouette values, what do you suggest for the number of clusters?

Seeing both previous graphs. At the silhouette graph, for 4 clusters we get the closest value to 1 of the score (0.54147), which means it is most appropriate cluster for the given data. On the other hand, if we have a look on the elbow graph the k=4 is at the “elbow” of the graph. This is also a good indicator that is the proper number of clusters.

Question 4 (30 points)

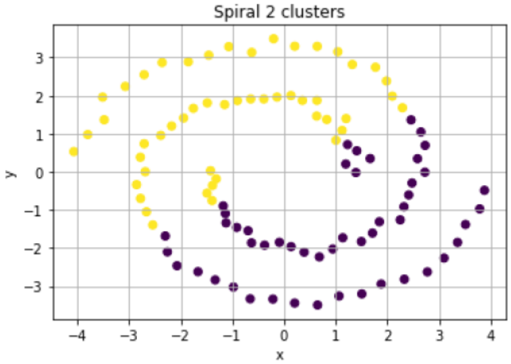
Apply the Spectral Clustering method to the Spiral.csv. Your input fields are x and y.

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



By visual inspection we could say there are two clusters, one for each arm of the spiral.

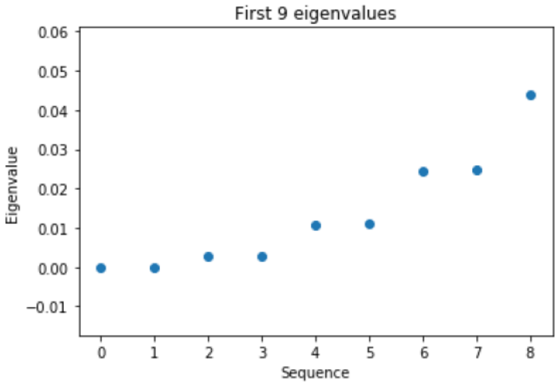
b) (5 points) Apply the K-mean algorithm directly using your number of clusters (in a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?



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

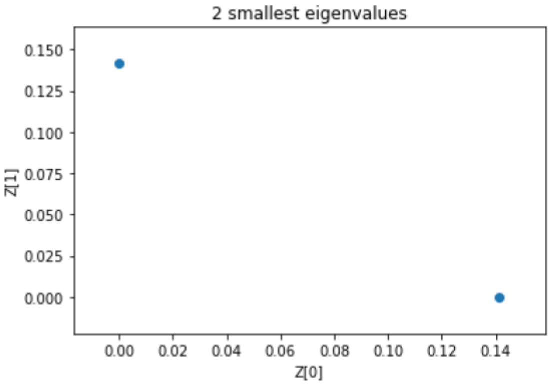
We choose 3 neighbors for example.

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 a) is appropriate?

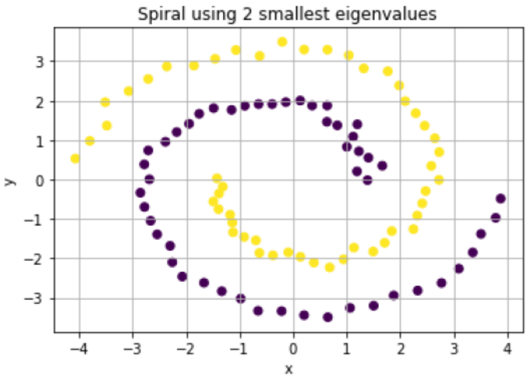


Yes, it is appropriate as there are no obvious jumps between them.

e) (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?



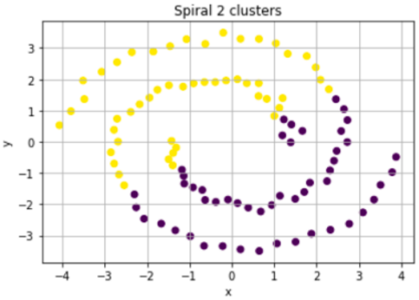
The previous picture shows the plot of the two smallest eigenvalues.

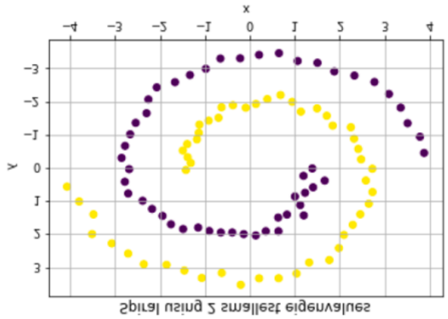


The previous picture shows the plot of the spiral regenerated using the two smallest eigenvalues.

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

The results are almost completely optimal. The spiral has been regenerated pretty well as it is shown in the following pictures (original on the left, regenerated on the right). The 2 clusters are well classified the two arms of the spiral


Spiral 2 clusters
y
x


Spiral using 2 smallest eigenvalues
y
x