

CS 584-04: Machine Learning

Autumn 2019 Assignment 2

Question 1 (50 points)

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

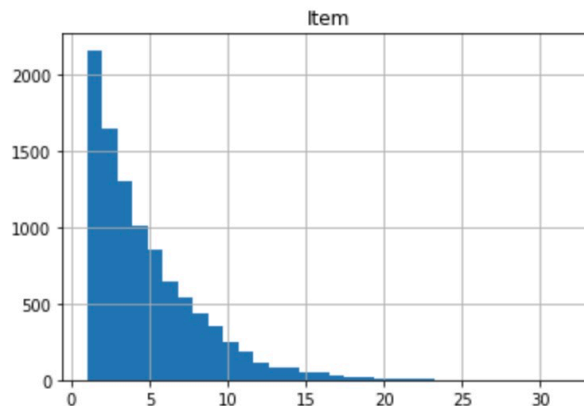
1. Customer: Customer Identifier
2. Item: Name of Product Purchased

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) (10 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?

Solution:-



Median:- 3

25th percentile:- 2

75th percentile:- 6

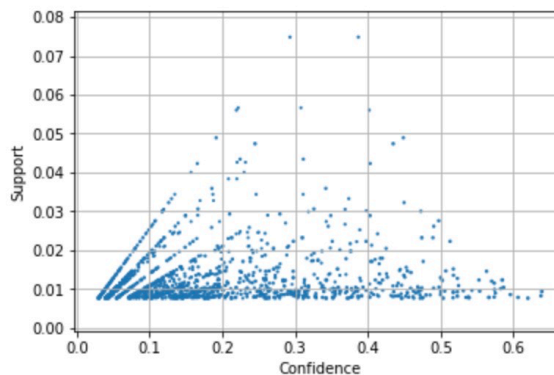
- b) (10 points) If you are interested in the k -itemsets which can be found in the market baskets of at least seventy five (75) customers. How many itemsets can you find? Also, what is the largest k value among your itemsets?

Solution:- We can find 524 itemsets. Largest k -value among the itemsets is 4. Since the itemset is ordered in increasing order of the number of items in itemsets, we check the number of items in the last itemset.

- c) (10 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. Also, you **do not** need to show those rules.

Solution:- We found 1228 association rules

- d) (10 points) Graph the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (c). Please use the Lift metrics to indicate the size of the marker.



- e) (10 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.

	antecedents	consequents	
0	(butter, root vegetables)	(whole milk)	
1	(butter, yogurt)	(whole milk)	
2	(yogurt, other vegetables, root vegetables)	(whole milk)	
3	(tropical fruit, other vegetables, yogurt)	(whole milk)	

	antecedent support	consequent support	support	confidence	lift
0	0.012913	0.255516	0.008236	0.637795	2.496107
1	0.014642	0.255516	0.009354	0.638889	2.500387
2	0.012913	0.255516	0.007829	0.606299	2.372842
3	0.012303	0.255516	0.007626	0.619835	2.425816

	leverage	conviction
0	0.004936	2.055423
1	0.005613	2.061648
2	0.004530	1.890989
3	0.004482	1.958317

SUPPORT:-

0	0.008236
1	0.009354
2	0.007829
3	0.007626

LIFT:-

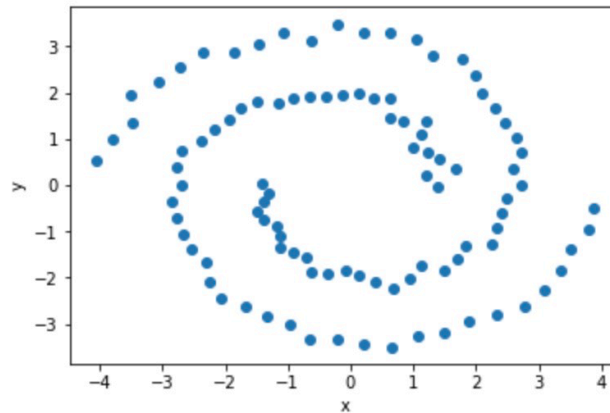
0	2.496107
1	2.500387
2	2.372842
3	2.425816

Question 2 (50 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) (10 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?

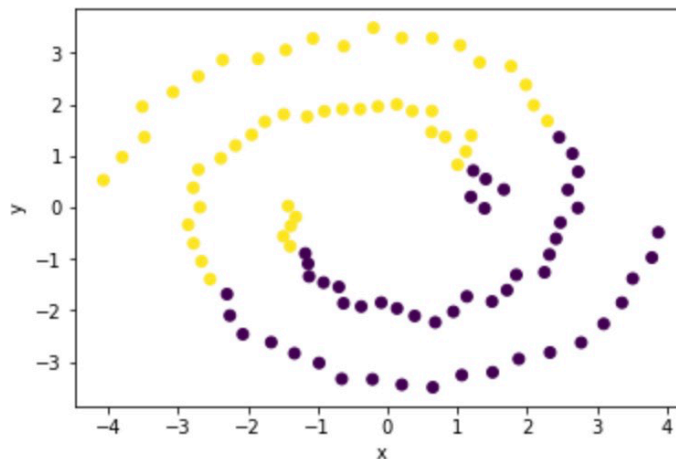
Solution:-



We can say there are 2 clusters by visual inspection.

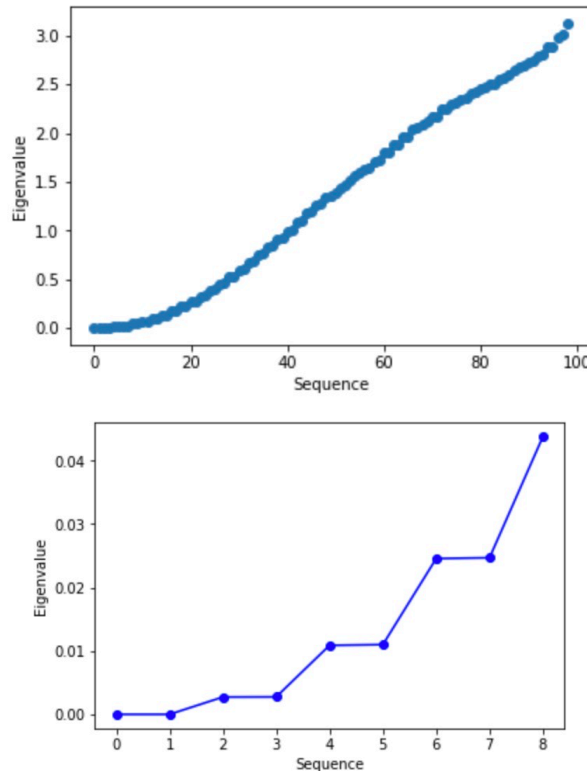
- b) (10 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?

Solution:-



- c) (10 points) Apply the nearest neighbor algorithm using the Euclidean distance. How many nearest neighbors will you use? Remember that you may need to try a couple of values first and use the eigenvalue plot to validate your choice.

Solution:-



We will be using 3 nearest neighbors after trying all the points which shows visible jump.

- d) (10 points) Retrieve the first two eigenvectors that correspond to the first two smallest eigenvalues. Display up to ten decimal places the means and the standard deviation of these two eigenvectors. Also, plot the first eigenvector on the horizontal axis and the second eigenvector on the vertical axis.

Solution:-

For First eigenvector:-

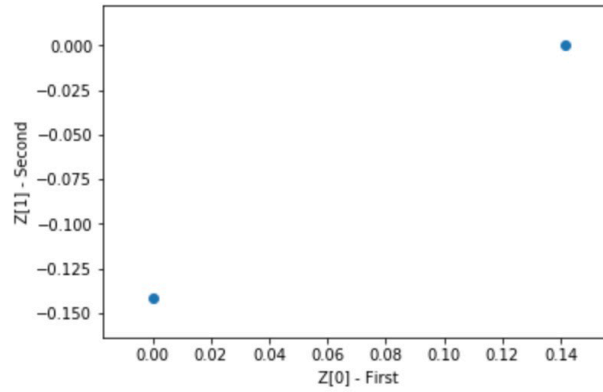
Mean = 0.0707106781

Std. Deviation = 0.0707106781

For Second Minimum eigenvalue:-

Mean = -0.0707106781

Std. Deviation:- 0.0707106781



- e) (10 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?

Solution:-

