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.

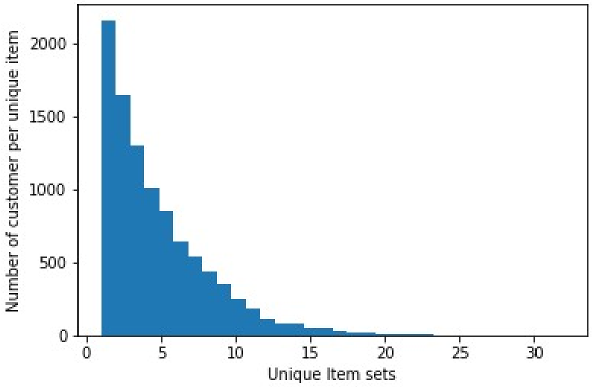
After you have imported the CSV file, please discover association rules using this dataset.

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?

Explanation: I have created the dataset such that we get number of customers for distinct item sets. From the data frame we infer that there can be itemset with maximum 32 items.

Median	3
25th percentile	2
75th percentile	6

Histogram of Unique items vs. No. of customers per unique items is as follows:



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?

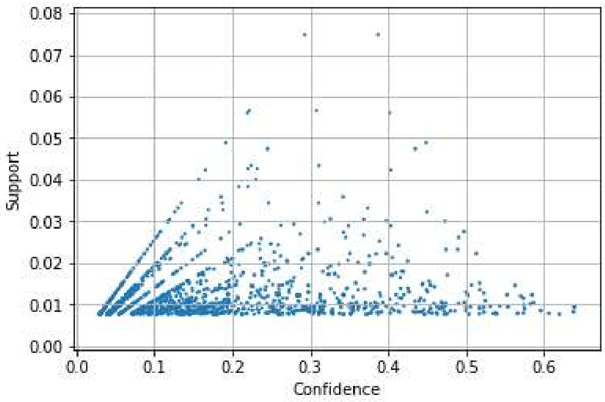
Explanation: I have chosen parameters of Apriori function used in the code based on apriori documentation links provided in the .ppt files given by professor. Link to the documentation: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
hence, min_support = 75/total_cust, here total_cust = total number of customers and max_len = 32. Therefore, Number of item sets are: 524, Largest k value among the item sets is: 4

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.

Explanation: I have set min_threshold = 0.01 hence, number of association rules are: 1228

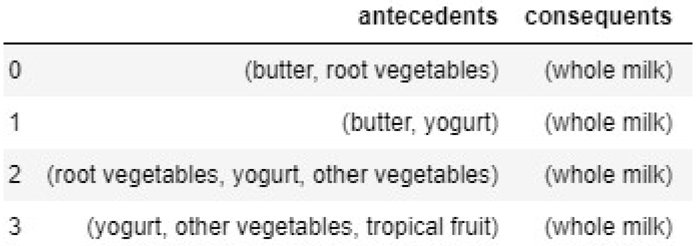
d) (10 points) Graph the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you found in (c). Please use the Lift metrics to indicate the size of the marker.

Explanation: The graph of Confidence vs. Support. It has been plotted using the s = assoc_rules['lift'] which indicates the size of marker.



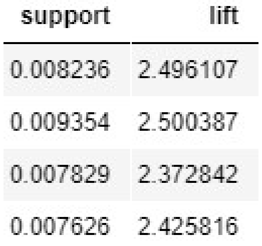
e) (10 points) List the rules whose Confidence metrics are at least 60%. Please include their Support and Lift metrics.

Rules whose Confidence metrics are at least 60%:



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

Support and Lift metrices:



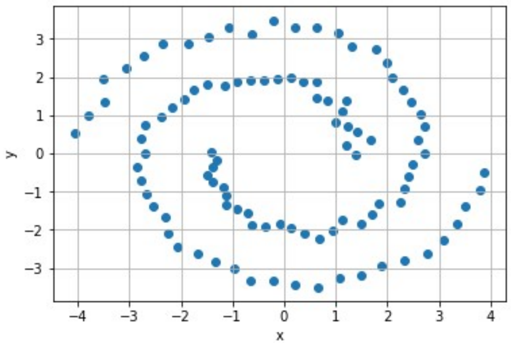
support	lift
0.008236	2.496107
0.009354	2.500387
0.007829	2.372842
0.007626	2.425816

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.

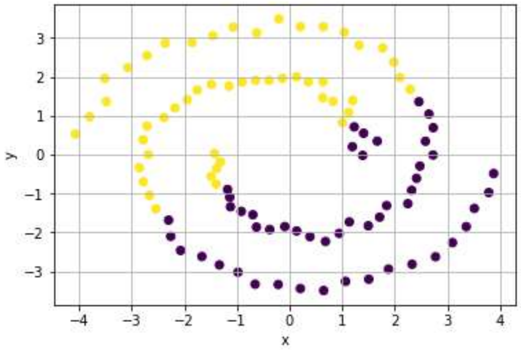
a) (10 points) Generate a scatterplot of y (vertical axis) versus x (horizontal axis). How many clusters will you say by visual inspection?

Explanation: It can be seen that there are two clusters in the spiral graph growing in spiral fashion from inward to outward.



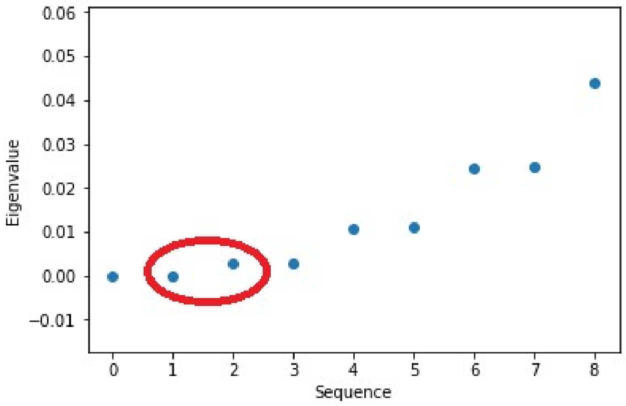
b) (10 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme?

Explanation: From above scatter plot I see that there are 2 clusters, hence I have used n_clusters = 2. But it can be seen from the graph that cluster formation is inappropriate. The cluster formation has taken place diagonally instead of dividing the datapoints in two distinct spiral clusters.



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.

Explanation: For the given dataset.i.e. Spiral.csv the we can see that the data is connected but not necessarily compact hence K-mean algorithm is not the best choice to cluster this data. Hence, we use Spiral clustering technique for clustering based on transformation by dimensionality reduction. The obvious jump in the



graph is between 1 and 2, therefore I can confirm that using *three as nearest neighbors* is the best choice.

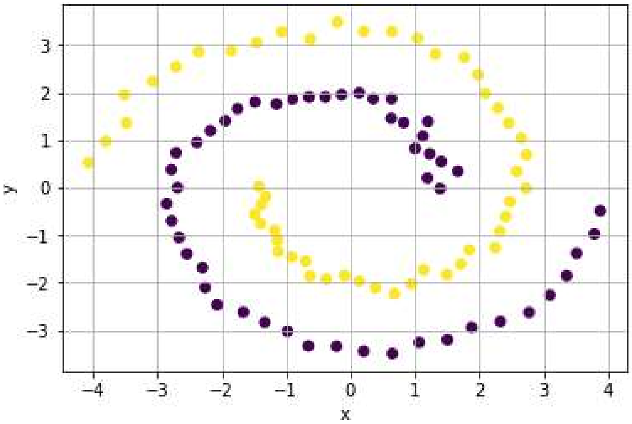
d) (10 points) Retrieve the first <u>two</u> eigenvectors that correspond to the first two smallest eigenvalues. Display up to ten decimal places the means and the standard deviation of these two eigenvectors. Also, plot the first eigenvector on the horizontal axis and the second eigenvector on the vertical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

Explanation: As explained in class we use small eigenvalues instead of large eigenvalues, because noise can creep in very fast for large eigenvalues if the dataset is very large. So, we use eigenvectors corresponding to small eigenvalues.

	First Eigenvector	Second Eigenvector
Value	1.41421356e-01 7.87731025e-13	7.77267140e-13 -1.41421356e-01
Mean	0.0707106781	-0.0707106781
Standard Deviation	0.0707106781	0.0707106781

e) (10 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?

- Explanation: As explained in Week4 .ppt slides: *Usually, the number of eigenvectors selected is the same as the number of fields used for clustering in the original training data.* Hence, I have used first <u>two</u> eigenvectors that correspond to the first two smallest eigenvalues and after applying the K-mean algorithm we obtain the proper clustering of the data.


3
2
1
0
−1
−2
−3
y
−4
−3
−2
−1
0
1
2
3
4
x