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

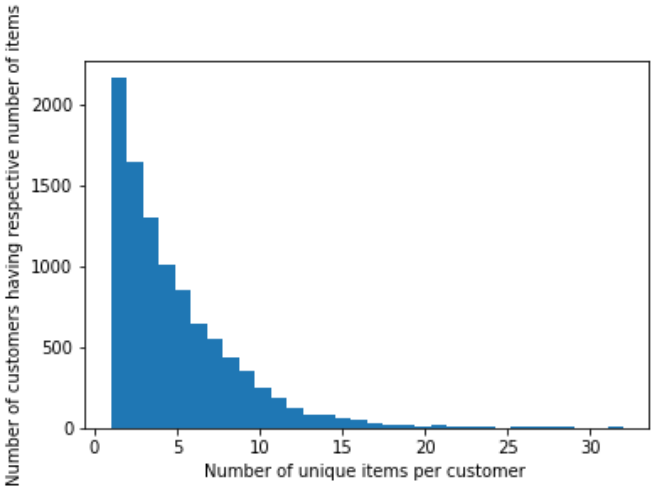
The data is already sorted in ascending order by Customer and then by Item. Also, all the items bought by each customer are all distinct.

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?

Ans:

1. **Histogram: There are 169 unique items across all customers.**



2. **Median, 25th percentile, 75th percentile:**

Median	**3**
25th Percentile	**2**
75th Percentile	**6**

```
median, the 25th percentile and the 75th percentile in above histogram is
```

```
:
```

	0
count	9835.000000
mean	4.409456
std	3.589385
min	1.000000
25%	2.000000
50%	3.000000
75%	6.000000
max	32.000000

---

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?

Ans:

1. **Min support= 75/9835**
   As we are asked to find k itemsets which can be found in at least 75 customers amongst 9835.
2. **524 frequent Itemsets can be found. I applied appriori function with min support of 75/9835 to calculate number of frequent itemsets.**
3. **Largest k value amongst itemsets = 4**
   There are maximum 4 items in itemset which satisfies condition of min support.

---

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.

Ans:

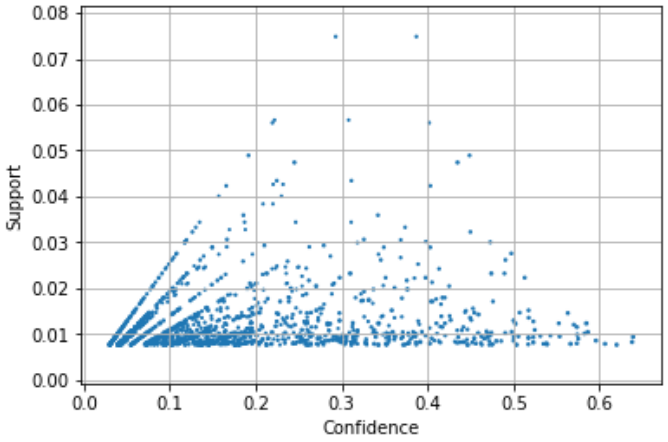
There are 1228 association rules having Confidence metrics at least 1%(0.01)
I applied appriori function and for metric is chosen as Confidence with min_threshold value = 0.01

---

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.

Ans:

```
plt.figure(figsize=(6,4))
plt.scatter(assoc_rules['confidence'], assoc_rules['support'], s = assoc_rules['lift'])
plt.grid(True)
plt.xlabel("Confidence")
plt.ylabel("Support")
plt.show()
```



Every marker has size of corresponding lift matrices value.

---

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

Ans:

antecedents	consequents	Support	Lift
(root vegetables, butter)	(whole milk)	0.008236	2.496107
(yogurt, butter)	(whole milk)	0.009354	2.500387
(yogurt, root vegetables, other vegetables)	(whole milk)	0.007829	2.372842
(tropical fruit, yogurt, other vegetables)	(whole milk)	0.007626	2.425816

```
assoc_rules_60 = association_rules(frequent_itemsets, metric = "confidence", min_threshold = 0.6)
assoc_rules_60
```

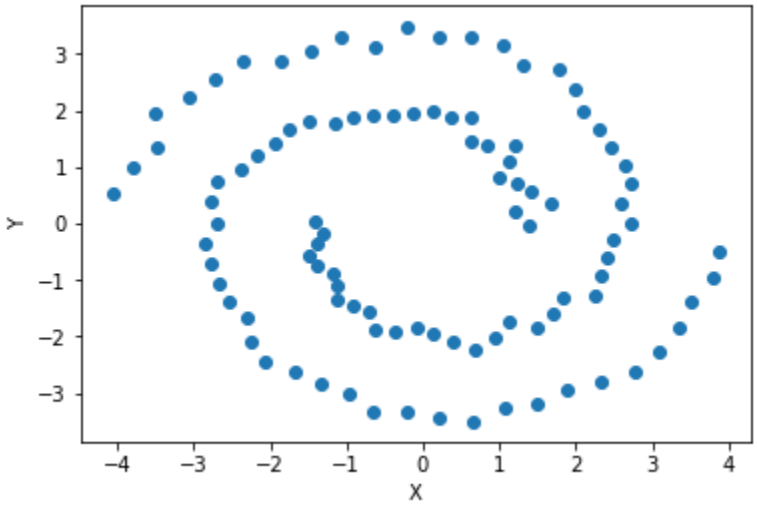
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(root vegetables, butter)	(whole milk)	0.012913	0.255516	0.008236	0.637795	2.496107	0.004936	2.055423
1	(yogurt, butter)	(whole milk)	0.014642	0.255516	0.009354	0.638889	2.500387	0.005613	2.061648
2	(yogurt, root vegetables, other vegetables)	(whole milk)	0.012913	0.255516	0.007829	0.606299	2.372842	0.004530	1.890989
3	(tropical fruit, yogurt, other vegetables)	(whole milk)	0.012303	0.255516	0.007626	0.619835	2.425816	0.004482	1.958317

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?

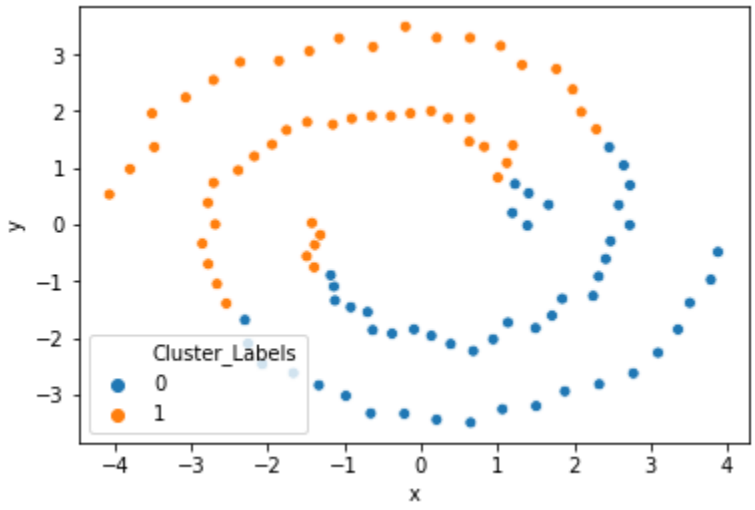
Ans: As there are 2 spirals present, we can say that there are 2 clusters present in scatter plot



---

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?

Ans:



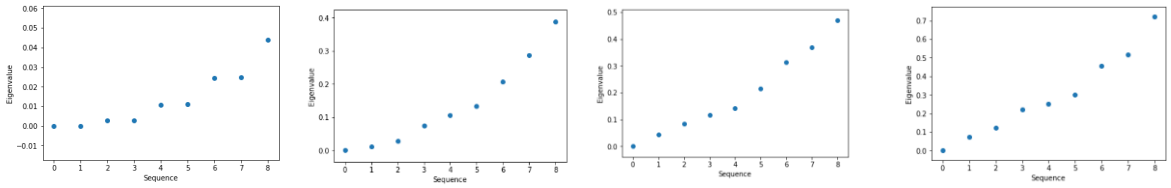
Above graph is a result of applying kmeans for 2 cluster solutions. But these are not the clusters we desired.

---

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.

Ans:

- **I tried using different values of k neighbors (3,7,8,9) in which I saw that there is a jump at point 1 to 2 (which means 3 graph starts with 0) which is occurring again and again in each sequence plot of eigenvalues.**
- **As discussed in class, this observation confirms that the 3 number of neighbors is appropriate solution.**

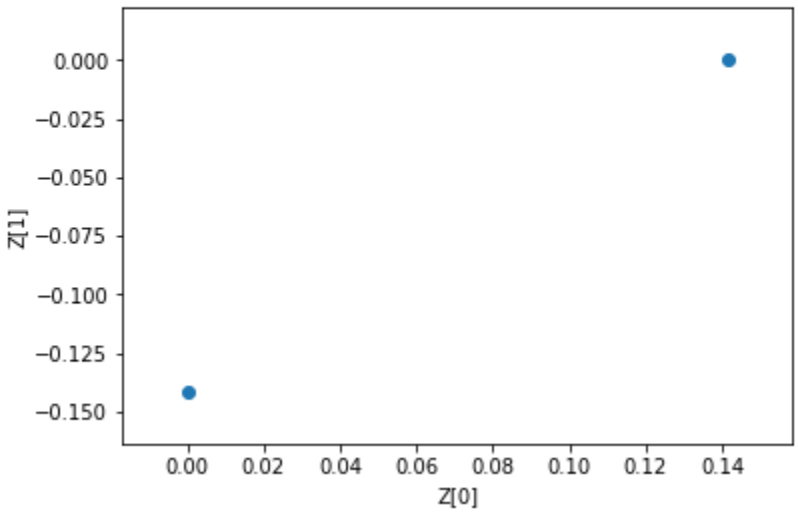


---

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.

Ans:

```
1st eigenvectors: [[1.41421356e-01 7.87731025e-13]]
2nd eigenvectors: [[ 7.77267140e-13 -1.41421356e-01]]
```



Mean of 1st eigenvector	0.0707106781
Mean of 2nd eigenvector	-0.0707106781
Std deviation of 1st eigenvector	0.0707106781
Std deviation of 2nd eigenvector	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?

Ans:

After applying kmeans algorithm on 1st two eigenvectors that correspond to the 1st two smallest eigenvalues we got below graph having 2 spiral clusters.