CS 484: Introduction to Machine Learning

Autumn 2020 Assignment 2

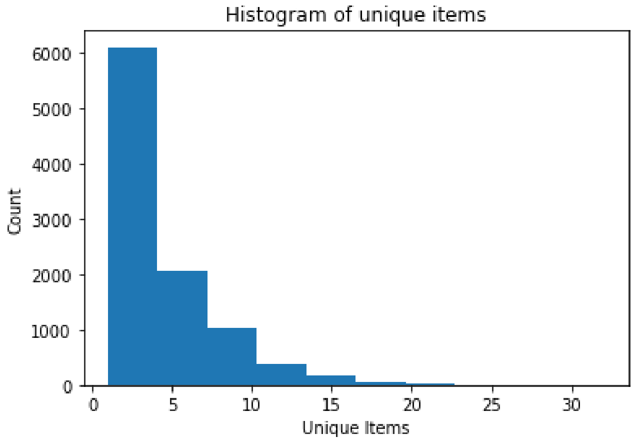
Question 1 (35 points)

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

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

After you have imported the CSV file, please discover association rules using this dataset. For your information, the observations have been sorted in ascending order by Customer and then by Item. Also, duplicated items for each customer have been removed.

a) (5 points) Create a data frame that contains the number of unique items in each customer's market basket. Draw a histogram of the number of unique items. What are the 25th, 50th, and the 75th percentiles of the histogram?



25th percentile: 2

50th percentile: 3

75th percentile: 6

b) (10 points) We are only interested in the *k*-itemsets that can be found in the market baskets of at least seventy five (75) customers. How many itemsets can we find? Also, what is the largest *k* value among our itemsets?

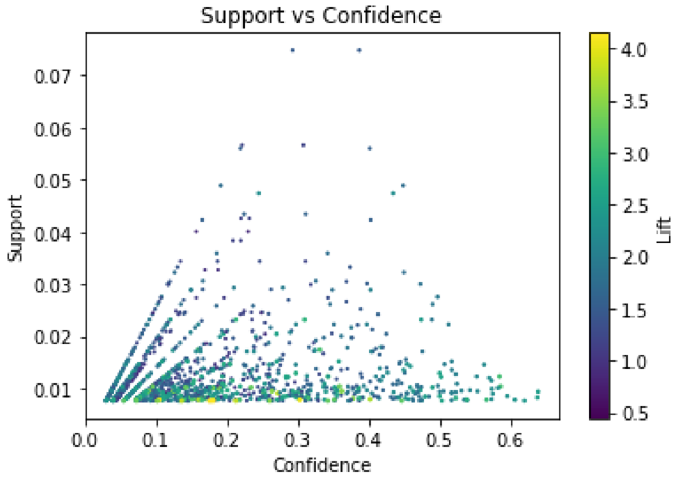
There are 524 item sets we can find.

The largest k value among the item sets is 4.

c) (10 points) Find out the association rules whose Confidence metrics are greater than or equal to 1%. How many association rules can we find? Please be reminded that a rule must have a non-empty antecedent and a non-empty consequent. Please **do not** display those rules in your answer.

1228 association rules can be found with confidence >= 1%.

d) (5 points) Plot the Support metrics on the vertical axis against the Confidence metrics on the horizontal axis for the rules you have found in (c). Please use the Lift metrics to indicate the size of the marker.



e) (5 points) List the rules whose Confidence metrics are greater than or equal to 60%. Please include their Support and Lift metrics.

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

Question 2 (10 points)

In the breakfast café example in Chapter 4 of *A Practitioner's Guide to Machine Learning*, you will take the three most frequent categories as the initial centroids. What is your final cluster solution? How does this final cluster solution compare with that of the example?

C1 = Coffee, C2 = Tea, C3 = Juice

Beverage	Coffee	Hot Chocolate	Juice	Milk	Soda	Spring Water	Tea
Frequency	**90**	**20**	**40**	**25**	**5**	**10**	**70**
Distance from c1, Coffee	**0**	**0.0611**	**0.0361**	**0.0511**	**0.2111**	**0.1111**	**0.0254**
Distance from c2, Tea	**0.0254**	**0.0643**	**0.0393**	**0.0543**	**0.2143**	**0.1143**	**0**
Distance from c3, Juice	**0.0361**	**0.0750**	**0**	**0.0650**	**0.2250**	**0.1250**	**0.0393**
Cluster Membership	**1**	**1**	**3**	**1**	**1**	**1**	**2**

This results in Coffee, Hot Chocolate, Milk, Soda, and Spring Water being in one cluster where Coffee is the centroid. Tea is the only one in its cluster where Tea is the centroid. Juice is the only one in its cluster where Juice is the centroid. Comparing with that of the example, they are very similar. In the textbook example, the result was Soda and Tea in their own clusters, while the remaining five are in cluster where Coffee is the centroid. Similar in the sense that both results in two beverages being in their own clusters, while the remaining five are in the cluster where coffee is the centroid.

Question 3 (10 points)

Let $\{a_{ij}, i, j = 1, \dots, n\}$ denote the matrix elements of the Adjacency matrix. Suppose λ is an eigenvalue and $\mathbf{v}$ the corresponding eigenvector of the Laplacian matrix. Prove mathematically that $\lambda = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\left(v_i - v_j\right)^2$ where $\{v_i, i = 1, \dots, n\}$ denote the elements of the eigenvector $\mathbf{v}$. If possible, please type your answer using an Equation Editor (Press **Alt** and **=** together in Microsoft Word) or similar application.

With $\{a_{ij}, i, j = 1, \dots, n\}$ denoting the matrix elements of adjacency matrix and λ is an eigenvalue and v the corresponding eigenvector of Laplacian matrix. Then, let L be the Laplacian matrix, $L = D - A$

Thus $v\lambda = Lv$ and multiplying v^t to the left, we get that $v^t v\lambda = v^t Lv$ but because $v^t v = 1$, it simplifies to $\lambda = v^t Lv$

Expanding L, we get that $\lambda = v^t Dv - v^t Av$

Expressing $D, v, A,$ in terms of their elements, d_{ij}, v_i, a_{ij} respectively

We get that $\lambda = \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}v_iv_j - \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_iv_j$

When i does not equal j, $d_{ij} = 0$ **so can simplify to** $\lambda = \sum_{i=1}^{n} d_{ii}v_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_iv_j$ **(1)**

With $d_{ii} = \sum_{j=1}^{n} a_{ij}$ **by definition, we get that** $\sum_{i=1}^{n} d_{ii}v_i^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2$

With symmetric property, $\boldsymbol{a_{ij}} = \boldsymbol{a_{ji}}$ **and switching indices**

we can get that $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_j^2$

Thus, $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2 = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_j^2\right)$

And, in terms of D, $\sum_{i=1}^{n} d_{ii}v_i^2 = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_j^2\right)$

Substituting back to (1):

We get: $\lambda = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_j^2\right) - \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_iv_j$

Which expanded, $\lambda = \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_i^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_iv_j + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_j^2\right)$

Can be simplified to $\lambda = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\left(v_i - v_j\right)^2$

Thus $\boldsymbol{\lambda = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\left(v_i - v_j\right)^2}$ **where** $\{\boldsymbol{v_i}, \boldsymbol{i} = \mathbf{1}, \dots, \boldsymbol{n}\}$ **denote the elements of the eigenvector** $\mathbf{v}$**.**

Question 4 (10 points)

Suppose Cluster 0 contains observations {-2, -1, 1, 2, 3} and Cluster 1 contains observations {4, 5, 7, 8}.

a) (4 points) Calculate the Silhouette Width of the observation 2 (i.e., the value -1) in Cluster 0.

a = mean distance between observation and all other points in the same cluster = 10/4 = 2.5

b = smallest mean distance of the observation to all points in any other cluster = 7

S = (b – a) / max(a, b) = 0.6428571428571429

Thus, the silhouette width of observation 2 in cluster 0 is 0.6428571. Calculated in python file.

b) (4 points) Calculate the cluster-wise Davies-Bouldin value of Cluster 0 (i.e., R_0) and Cluster 1 (i.e., R_1).

$S_k = \frac{1}{n_k}\sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{c}_k)$ and calculating in python we get:

Cluster 0 has **1.68** cluster-wise Davies-Bouldin value

Cluster 1 has **1.5** cluster-wise Davies-Bouldin value

c) (2 points) What is the Davies-Bouldin Index of this two-cluster solution?

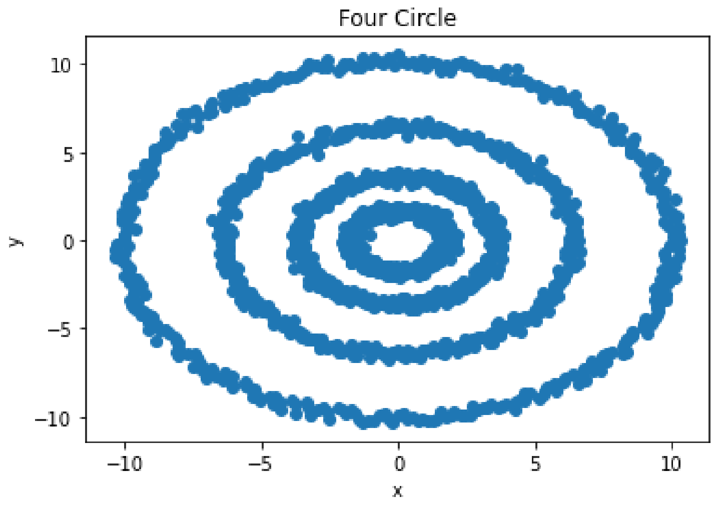
$$M_{kl} = d(\mathbf{c}_k, \mathbf{c}_l) \rightarrow R_{kl} = \frac{S_k + S_l}{M_{kl}} \rightarrow R_k = \max_{\substack{1\le l\le K \\ l\ne k}} R_{kl} \rightarrow DB = \frac{1}{K}\sum_{k=1}^{K} R_k$$

Davies-Bouldin index: 0.2944 – calculated in python file

Question 5 (35 points)

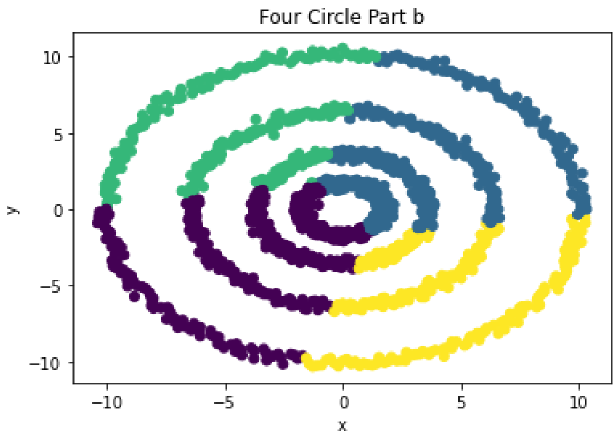
Apply the Spectral Clustering method to the FourCircle.csv. Your input fields are x and y. Wherever needed, specify random_state = None in calling the KMeans function.

a) (5 points) Plot y on the vertical axis versus x on the horizontal axis. Based on your visual inspection, how many connected clusters are there?



Looking at the graph, there seems to be 4 clusters because there are 4 circles.

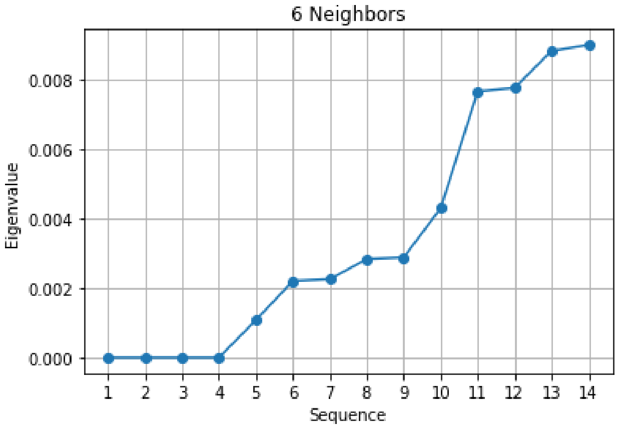
b) (5 points) Apply the K-mean algorithm directly using your number of clusters that you think in (a). Regenerate the scatterplot using the K-mean cluster identifiers to control the color scheme. Please comment on this K-mean result.



Looking at the image, the KMean cluster doesn't apply the correct cluster with the 4 circles. Rather than each circle being a cluster, it's parts of the 4 circles being in a cluster.

c) (10 points) Apply the nearest neighbor algorithm using the Euclidean distance. We will consider the number of neighbors from 1 to 15. What is the smallest number of neighbors that we should use to discover the clusters correctly? Remember that you may need to iterate between parts c), d), and e) a couple of values and use the eigenvalue plot to validate your choice.

After trying multiple values, I found 6 to be the smallest number of neighbors to use by looking at the graphs from 1 to 15 neighbors. The graph below shows a steep increase at 4. Going below, n = 5, the steep increase is at 5, thus, 6 is the smallest number neighbor to discover the clusters correctly.

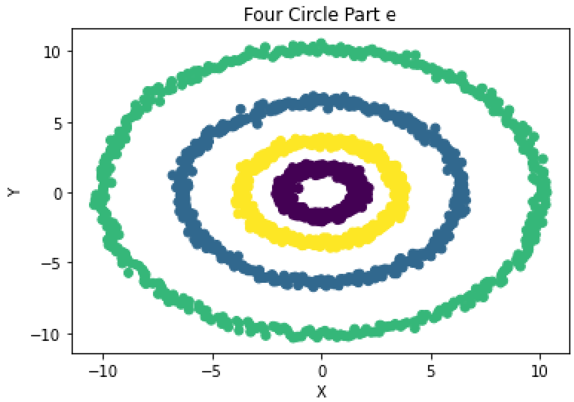


d) (5 points) Using your choice of the number of neighbors in (c), calculate the Adjacency matrix, the Degree matrix, and finally the Laplacian matrix. How many eigenvalues do you determine are practically zero? Please display values of the "zero" eigenvalues in scientific notation.

4 eigenvalues are practically zero.

[-2.82310149e-15, -1.21658580e-15, -3.76692065e-16, 2.68402137e-16]

e) (10 points) Apply the K-mean algorithm on the eigenvectors that correspond to your "practically" zero eigenvalues. The number of clusters is the number of your "practically" zero eigenvalues. Regenerate the scatterplot using the K-mean cluster identifier to control the color scheme.


Four Circle Part e
Y
X