CAP- 5610 Machine Learning

Homework 4

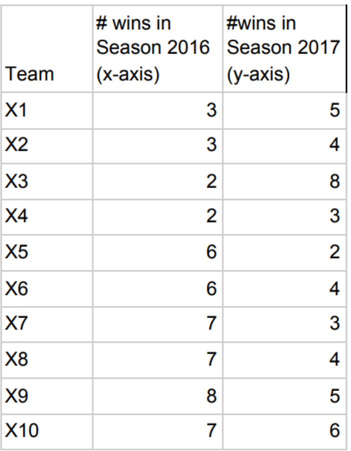
Name: Nabila Shahnaz Khan

NID: 5067496

Unsupervised learning

Task 1:

Suppose we have 10 college football teams X1 to X10. We want to cluster them into 2 groups. For each football team, we have two features: One is # wins in Season 2016, and the other is # wins in Season 2017.



Team	# wins in Season 2016 (x-axis)	#wins in Season 2017 (y-axis)
X1	3	5
X2	3	4
X3	2	8
X4	2	3
X5	6	2
X6	6	4
X7	7	3
X8	7	4
X9	8	5
X10	7	6

Q1) Initialize with two centroids, (4, 6) and (5, 4). Use Manhattan distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

Answer: After **1st iteration**, the results are:

Centroids: (4.0, 6.33), (5.57, 3.57)

Cluster 0: ('X1', 3.0, 5.0), ('X3', 2.0, 8.0), ('X10', 7.0, 6.0)

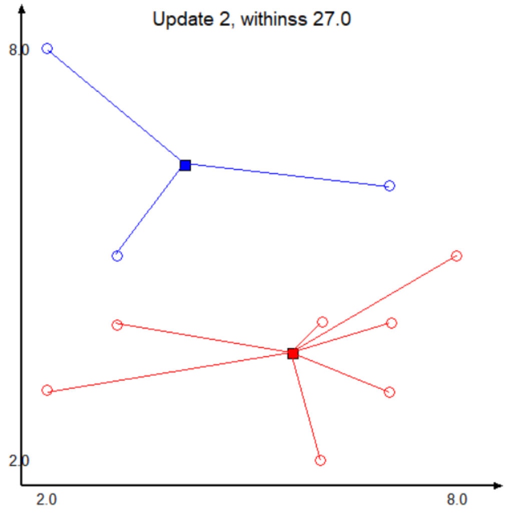
Cluster 1: ('X2', 3.0, 4.0), ('X4', 2.0, 3.0), ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0), ('X9', 8.0, 5.0)

The **final results** after applying K-Means are:

Centroids: (4.0, 6.33), (5.57, 3.57)

Cluster 0: ('X1', 3.0, 5.0), ('X3', 2.0, 8.0), ('X10', 7.0, 6.0)

Cluster 1: ('X2', 3.0, 4.0), ('X4', 2.0, 3.0), ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0), ('X9', 8.0, 5.0)



Q2) Initialize with two centroids, (4, 6) and (5, 4). Use Euclidean distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

Answer: After **1st iteration**, the results are:

Centroids: (2.5, 6.5), (5.75, 3.875)

Cluster 0: ('X1', 3.0, 5.0), ('X3', 2.0, 8.0)

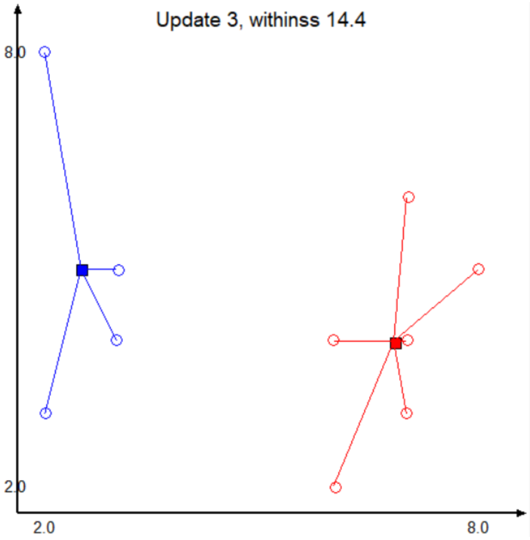
Cluster 1: ('X2', 3.0, 4.0), ('X4', 2.0, 3.0), ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0), ('X9', 8.0, 5.0), ('X10', 7.0, 6.0)

The **final results** after applying K-Means are:

Centroids: (2.5, 5.0), (6.83, 4.0)

Cluster 0: ('X1', 3.0, 5.0), ('X2', 3.0, 4.0), ('X3', 2.0, 8.0), ('X4', 2.0, 3.0)

Cluster 1: ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0), ('X9', 8.0, 5.0), ('X10', 7.0, 6.0)



Q3) Initialize with two centroids, (3, 3) and (8, 3). Use Manhattan distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

Answer: After **1st iteration**, the results are:

Centroids: (2.5, 5.0), (6.83, 4.0)

Cluster 0: ('X1', 3.0, 5.0), ('X2', 3.0, 4.0), ('X3', 2.0, 8.0), ('X4', 2.0, 3.0)

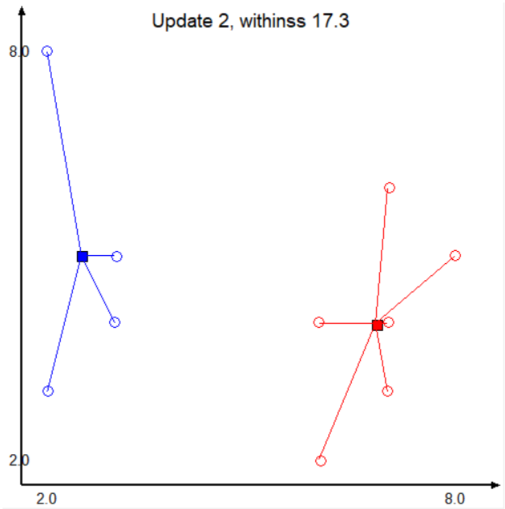
Cluster 1: ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0), ('X9', 8.0, 5.0), ('X10', 7.0, 6.0)]

The **final results** after applying K-Means are:

Centroids: (2.5, 5.0), (6.83, 4.0)

Cluster 0: ('X1', 3.0, 5.0), ('X2', 3.0, 4.0), ('X3', 2.0, 8.0), ('X4', 2.0, 3.0)

Cluster 1: ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0), ('X9', 8.0, 5.0), ('X10', 7.0, 6.0)]



Q4) Initialize with two centroids, (3, 2) and (4, 8). Use Manhattan distance as the distance metric. First, perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids. Second, please use K-Means to find two clusters.

Answer: After **1st iteration**, the results are:

Centroids: (4.86, 3.57), (5.67, 6.33)

Cluster 0: ('X1', 3.0, 5.0), ('X2', 3.0, 4.0), ('X4', 2.0, 3.0), ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0)

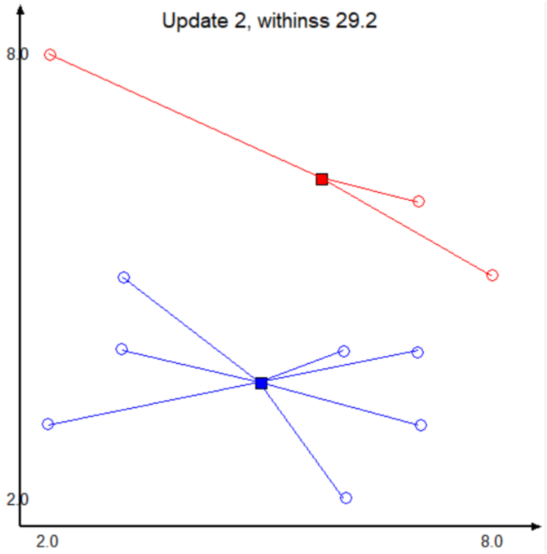
Cluster 1: ('X3', 2.0, 8.0), ('X9', 8.0, 5.0), ('X10', 7.0, 6.0)

The **final results** after applying K-Means are:

Centroids: (4.86, 3.57), (5.67, 6.33)

Cluster 0: ('X1', 3.0, 5.0), ('X2', 3.0, 4.0), ('X4', 2.0, 3.0), ('X5', 6.0, 2.0), ('X6', 6.0, 4.0), ('X7', 7.0, 3.0), ('X8', 7.0, 4.0)

Cluster 1: ('X3', 2.0, 8.0), ('X9', 8.0, 5.0), ('X10', 7.0, 6.0)



Task 2:

First, download the Iris data set from: https://archive.ics.uci.edu/ml/datasets/Iris. Then, implement the K-means algorithm. K-means algorithm computes the distance of a given data point pair. Replace the distance computation function with Euclidean distance, 1- Cosine similarity, and 1 – the Generalized Jarcard similarity (https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/jaccard.htm).

Q1) Run K-means clustering with Euclidean, Cosine and Jaccard similarity. Specify K= the number of categorical values of y (the variable of label). Compare the SSEs of Euclidean-K-means Cosine-K-means, Jaccard-K-means. Which method is better?

Answer: Here, the number of categories is 3, so K = 3. The SSEs of Euclidean-K-means, Cosine-K-means, Jaccard-K-means are given below:

Distance Calculation Metric	SSE
Euclidean-K-means	78.94
Cosine-K-means	680.8
Jaccard-K-means	79.19

According to the SSE values, Euclidean-K-means works best as it's SSE value is lowest.

Q2) Compare the accuracies of Euclidean-K-means Cosine-K-means, Jaccard-K-means. First, label each cluster with the label of the highest votes. Later, compute the accuracy of the Kmeans with respect to the three similarity metrics. Which metric is better?

Answer: Here, 'Iris-setosa' represents Cluster0, 'Iris-versicolor' represents Cluster1, 'Iris-virginica' represents Cluster2. The label of each cluster with the label of the highest votes are given below:

Distance Calculation Metric	Highest Votes
Euclidean-K-means	Cluster0: 38 Cluster1: 50 Cluster2: 62
Cosine-K-means	Cluster0: 38 Cluster1: 50 Cluster2: 62
Jaccard-K-means	Cluster0: 38 Cluster1: 50 Cluster2: 62

The accuracies of Euclidean-K-means Cosine-K-means, Jaccard-K-means are given below:

Distance Calculation Metric	Accuracy
Euclidean-K-means	89.33%
Cosine-K-means	33.33%
Jaccard-K-means	88%

According to the accuracy calculations, the Euclidean-K-means performs better than the other two.

Q3) Which of Euclidean-K-means, Cosine-K-means, Jaccard-K-means requires more iterations and times?
Answer: The number of iterations of Euclidean-K-means, Cosine-K-means, Jaccard-K-means are given below:

Distance Calculation Metric	# Iterations	Time (sec)
Euclidean-K-means	4	0.0149
Cosine-K-means	11	0.7128
Jaccard-K-means	7	0.0339

From the values of the table, it seems the Euclidean-K-means needs fewer iterations. Hence, Euclidean-K-means requires least amount of time to run.

Q4) Compare the SSEs of Euclidean-K-means Cosine-K-means, Jaccard-K-means with respect to the following three terminating conditions:
• when there is no change in centroid position
• when the SSE value increases in the next iteration
• when the maximum preset value (100) of iteration is complete
Which method requires more time or more iterations?
Answer:

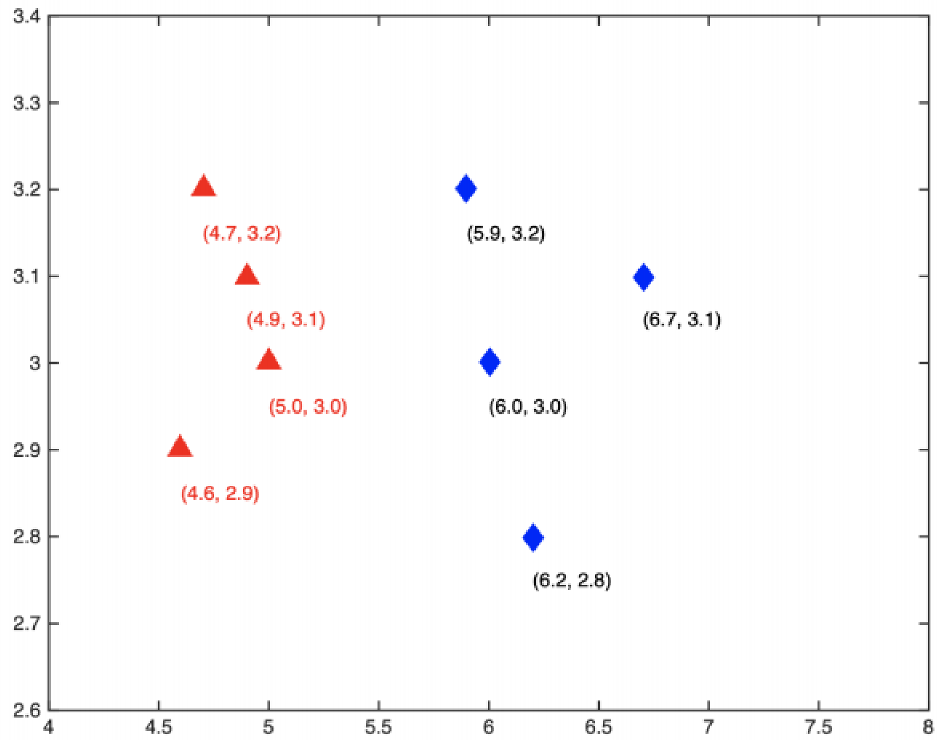
- ➢ When there is no change in centroid position
  - Euclidean SSE: SSE before iteration 0, SSE after iteration 133.26,
  - Cosine SSE: SSE before iteration 0, SSE after iteration 1374.58,
  - Jaccard SSE: SSE before iteration 0, SSE after iteration 134.84
- ➢ When the SSE value increases in the next iteration
  - Euclidean SSE: 78.94
  - Cosine SSE: 680.8,
  - Jaccard SSE: 79.19

➢ When the maximum preset value (100) of iteration is complete

- Euclidean SSE: 78.94
- Cosine SSE: 680.8,
- Jaccard SSE: 79.19

Task 3:

There are two clusters A (red) and B (blue), each has four members and plotted in Figure. The coordinates of each member are labeled in the figure. Compute the distance between two clusters using Euclidean distance.



A. What is the distance between the two farthest members? (round to four decimal places here, and next 2 problems);

Answer: The tow furthest members are: **(4.6,2.9)** and **(6.7,3.1).** The Euclidean distance between them is **2.1095**.

B. What is the distance between the two closest members?

Answer: The two closest members are **(5.0, 3.0)** and **(6.0, 3.0)**. The Euclidean distance between them is **1.0.**

C. What is the average distance between all pairs?

Answer: The centroid of cluster A is (4.8, 3.05) and cluster B is (6.2, 3.025). The average distance between all pairs is **1.4**.

D. Discuss which distance (A, B, C) is more robust to noises in this case?

Answer: Among the three distances of A, B, and C, the average distance between all pairs (calculated in C) is more robust to noises. Distance A only considers closest member while distance B only considers furthest members. They cannot represent all the members of the two clusters. Distance C considers all pairs, so the average distance has the capability to adjust the noise of the outliers while the other two measurement does not have this option.

Code Link:

Task 1: https://github.com/NabilaKhan/CAP-5610-Machine-Learning-/blob/main/CAP-5610-HW4-task1.ipynb

Task 2: https://github.com/NabilaKhan/CAP-5610-Machine-Learning-/blob/main/CAP-5610-HW4-task2.ipynb

[Please let me know if you are having any issue to find the code]