Unsupervised Learning : K-means Algorithm

Table of Contents

- Introduction
- Design
- Implementation
- Test
- Conclusion

Introduction

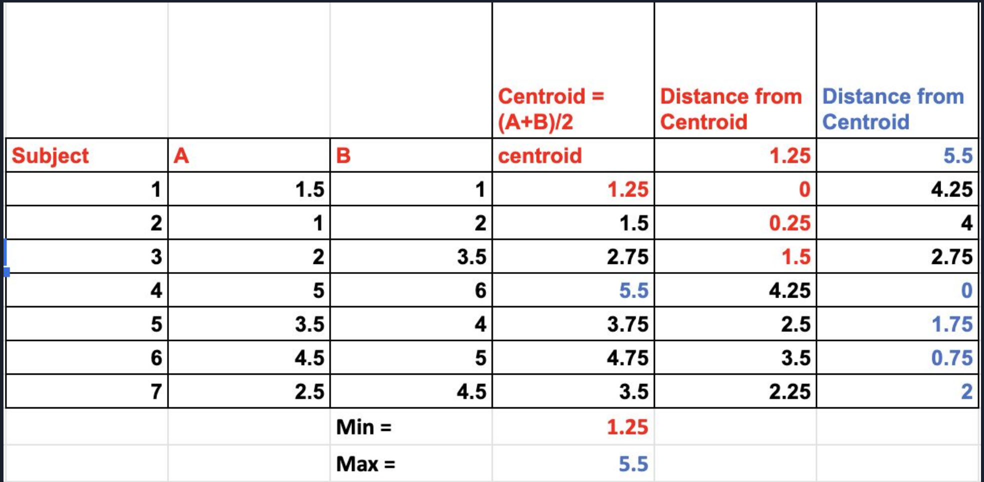
- In this assignment we will using k-means.
- K-means is an unsupervised learning algorithm that is used to get structure out of unstructured data.
- The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means.

Design

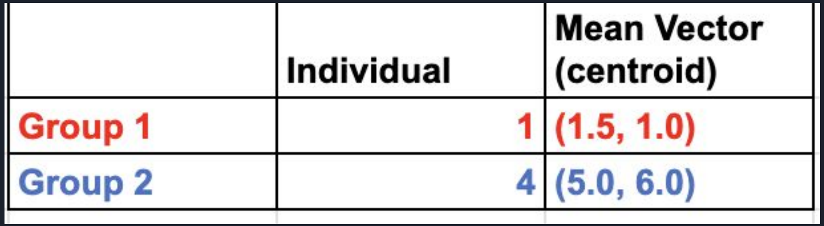
- The data to be used has to be in matrix format .
- In this example, our assumption is that we will have 2 clusters
- The next step would be to calculate two centroids so that we can partition our dataset into two groups
- Next we start clustering our individual data into clusters depending on the distance from the individual data centroid to the cluster centroid .
- The mean vector is recalculated each time a new member is added.
- But we cannot yet be sure that each individual has been assigned to the right cluster. So, we compare each individual's distance to its own cluster mean and to that of the opposite cluster.

Implementation

- Select the number of clusters needed to solve the problem.
- And then calculate the min and max Centroid
- centroid_x = (x1 + x2 + x3 + ... xn) / n
- centroid_y = (y1 + y2 + y3 + ... yn) / n



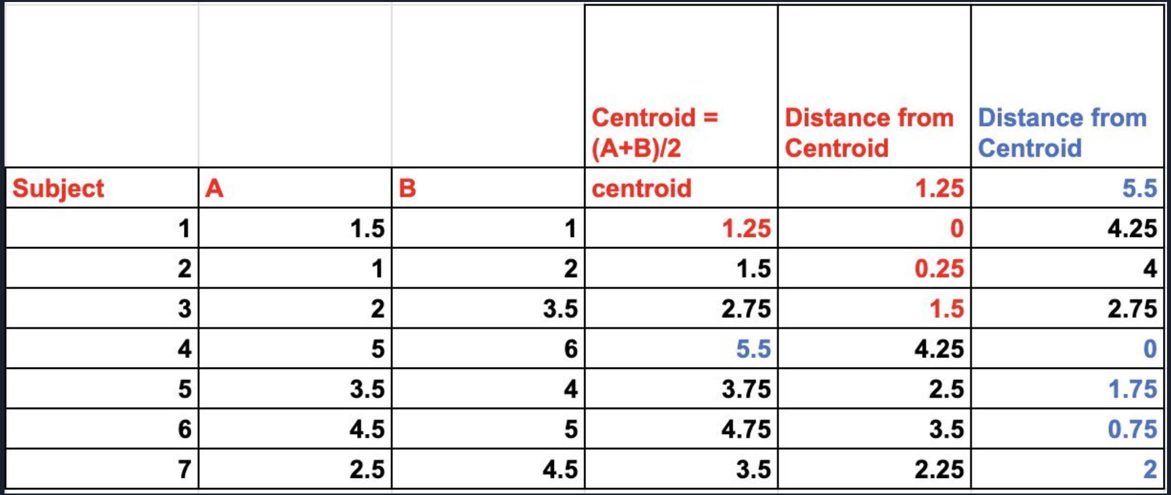
			Centroid = (A+B)/2	Distance from Centroid	Distance from Centroid
Subject	A	B	centroid	1.25	5.5
1	1.5	1	1.25	0	4.25
2	1	2	1.5	0.25	4
3	2	3.5	2.75	1.5	2.75
4	5	6	5.5	4.25	0
5	3.5	4	3.75	2.5	1.75
6	4.5	5	4.75	3.5	0.75
7	2.5	4.5	3.5	2.25	2
		Min =	1.25		
		Max =	5.5		



	Individual	Mean Vector (centroid)
Group 1	1	(1.5, 1.0)
Group 2	4	(5.0, 6.0)

Implementation

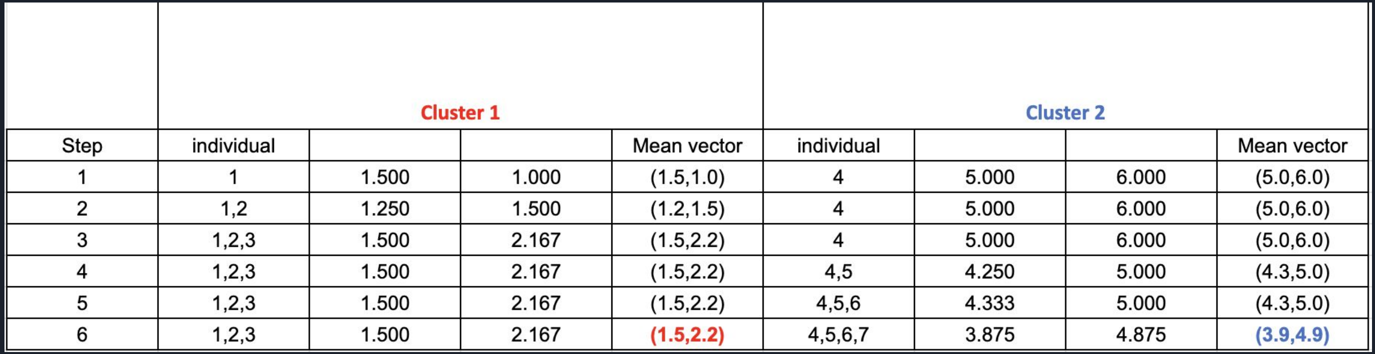
- Calculate the distance of each subject and the 2 centroids
- The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean.
    - Note: Subject 3 has the same distance to both clusters. We can randomly allocate it to any cluster.



			Centroid = (A+B)/2	Distance from Centroid	Distance from Centroid
Subject	A	B	centroid	1.25	5.5
1	1.5	1	1.25	0	4.25
2	1	2	1.5	0.25	4
3	2	3.5	2.75	1.5	2.75
4	5	6	5.5	4.25	0
5	3.5	4	3.75	2.5	1.75
6	4.5	5	4.75	3.5	0.75
7	2.5	4.5	3.5	2.25	2

Implementation

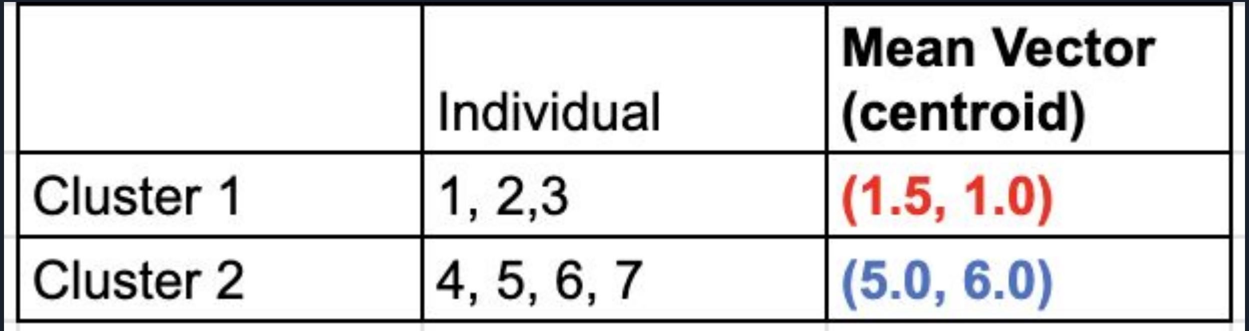
- Next we calculate the mean vector
- We also recalculate it everytime we add new member to a cluster



	Cluster 1				Cluster 2			
Step	individual			Mean vector	individual			Mean vector
1	1	1.500	1.000	(1.5,1.0)	4	5.000	6.000	(5.0,6.0)
2	1,2	1.250	1.500	(1.2,1.5)	4	5.000	6.000	(5.0,6.0)
3	1,2,3	1.500	2.167	(1.5,2.2)	4	5.000	6.000	(5.0,6.0)
4	1,2,3	1.500	2.167	(1.5,2.2)	4,5	4.250	5.000	(4.3,5.0)
5	1,2,3	1.500	2.167	(1.5,2.2)	4,5,6	4.333	5.000	(4.3,5.0)
6	1,2,3	1.500	2.167	**(1.5,2.2)**	4,5,6,7	3.875	4.875	**(3.9,4.9)**

Implementation

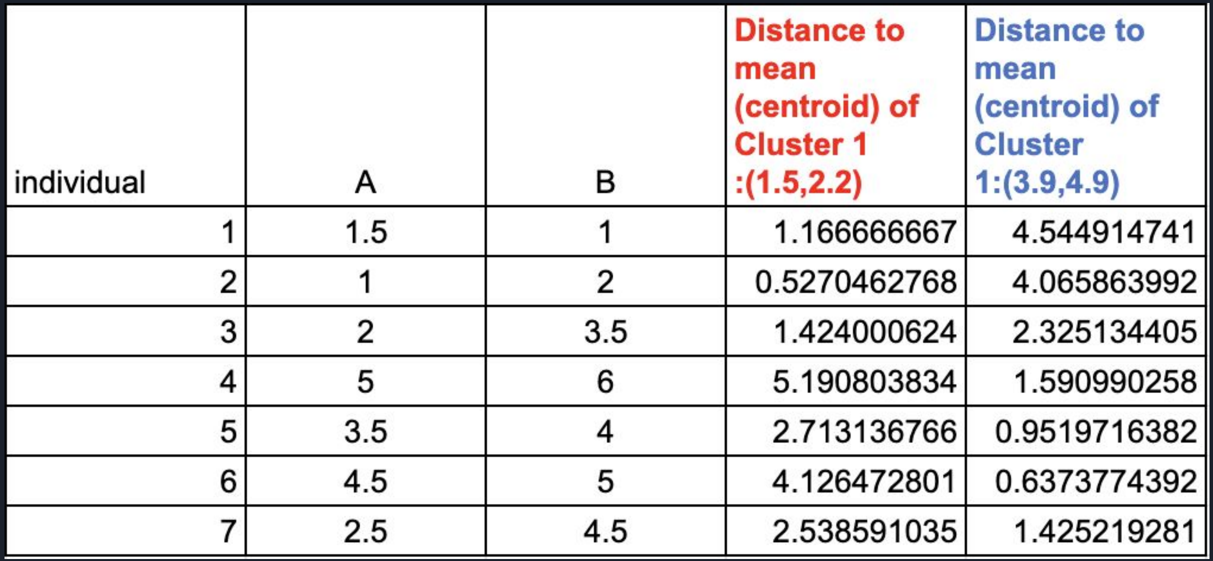
- We then move on to check the result of our previous mean vector calculation and placement
- We recalculate the mean vector , So, we compare each individual’s distance to its own cluster mean and to that of the opposite cluster.



	Individual	**Mean Vector (centroid)**
Cluster 1	1, 2,3	**(1.5, 1.0)**
Cluster 2	4, 5, 6, 7	**(5.0, 6.0)**

Implementation

- Now we calculate the new mean vector distance and compare each individual’s distance to the cluster, if there is any discrepancy we have to relocate to the closest cluster and then we do recalculate again, else we are done.



individual	A	B	Distance to mean (centroid) of Cluster 1 :(1.5,2.2)	Distance to mean (centroid) of Cluster 1:(3.9,4.9)
1	1.5	1	1.166666667	4.544914741
2	1	2	0.5270462768	4.065863992
3	2	3.5	1.424000624	2.325134405
4	5	6	5.190803834	1.590990258
5	3.5	4	2.713136766	0.9519716382
6	4.5	5	4.126472801	0.6373774392
7	2.5	4.5	2.538591035	1.425219281

Conclusion

- From the results we have we can see that the initial clustering we did was correct and that all assigned data to each clustered was assumed correctly. Our data now has been structured from a previously unstructured data and can be used to make inferences.

Bibliography

- https://hc.labnet.sfbu.edu/~henry/sfbu/course/data_science/algorithm/slide/k-means_example.html
- https://github.com/Alami64/Machine-Learning/upload/main/Unsupervised%20Learning