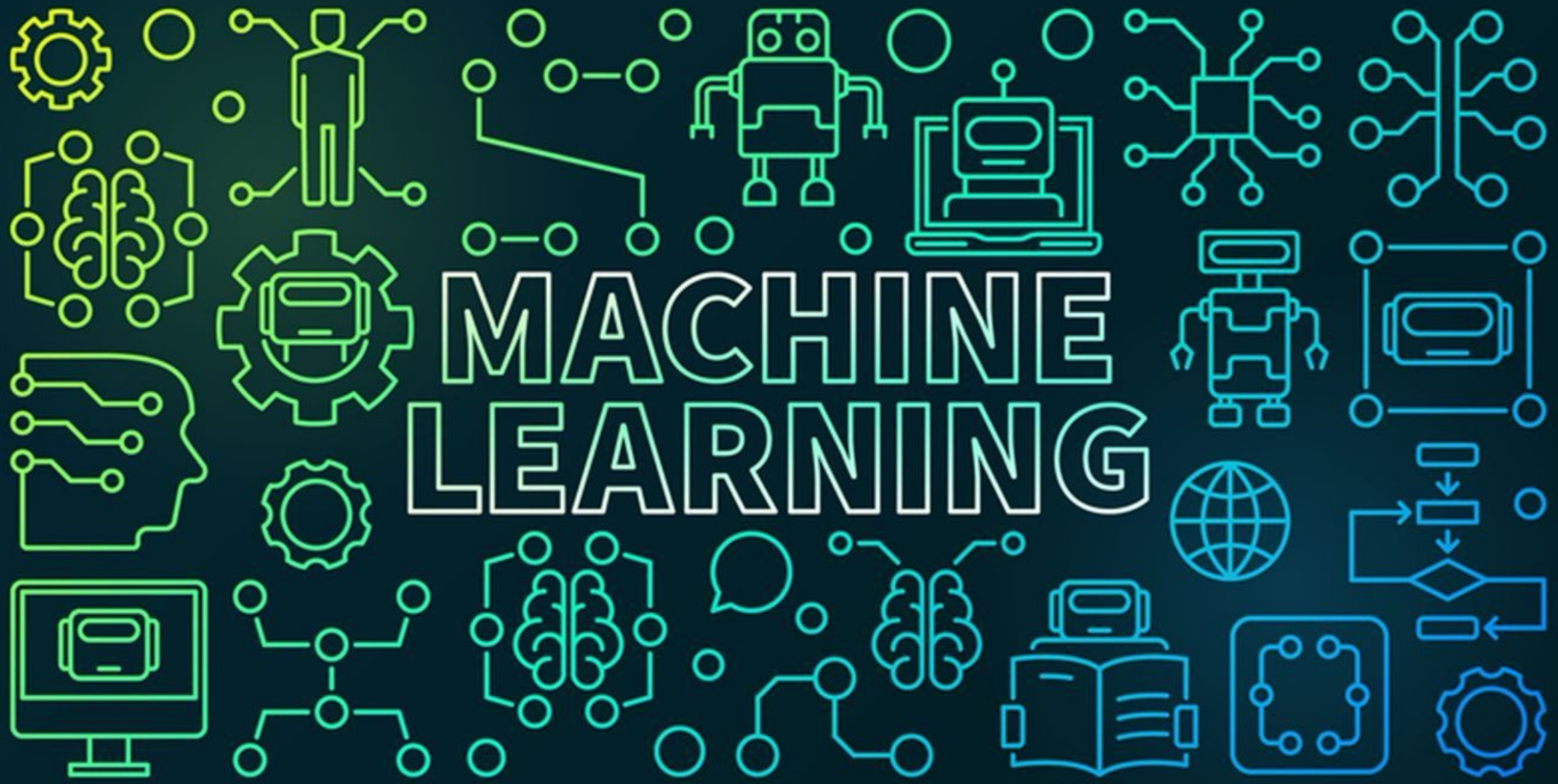


Muhammad Atif Saeed (Lecturer DS & AI)

What is K-means Clustering?

- K-means clustering is an unsupervised machine learning algorithm used to group a dataset into k clusters.
- It is an iterative algorithm that starts by randomly selecting k centroids in the dataset. After selecting the centroids, the entire dataset is divided into clusters based on the distance of the data points from the centroid.
- In the new clusters, the centroids are calculated by taking the mean of the data points.

K-means Clustering Algorithm

- First, we will select K random entries from the dataset and use them as centroids.
- Now, we will find the distance of each entry in the dataset from the centroids. You can use any distance metric such as euclidean distance, Manhattan distance, or squared euclidean distance.
- After finding the distance of each data entry from the centroids, we will start assigning the data points to clusters. We will assign each data point to the cluster with the centroid to which it has the least distance.
- After assigning the points to clusters, we will calculate the new centroid of the clusters. For this, we will use the mean of each data point in the same cluster as the new centroid. If the newly created centroids are the same as the centroids in the previous iteration, we will consider the current clusters to be final. Hence, we will stop the execution of the algorithm. If any of the newly created centroids is different from the centroids in the previous iteration, we will go to step 2.

Distance Functions

- There are various methods for calculating the distance between the new point and each training point

- The most commonly known methods are:
  - Euclidian (for continuous)
  - Manhattan (for continuous)
  - Hamming distance (for categorical)

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

K-Means Example (Cont.)

Point	Coordinates
A1	(2,10)
A2	(2,6)
A3	(11,11)
A4	(6,9)
A5	(6,4)
A6	(1,2)
A7	(5,10)
A8	(4,9)
A9	(10,12)
A10	(7,5)
A11	(9,11)
A12	(4,6)
A13	(3,10)
A14	(3,8)
A15	(6,11)

We are also given the information that we need to make 3 clusters. It means we are given K=3.We will solve this numerical on k-means clustering using the approach discussed below.

First, we will randomly choose 3 centroids from the given data. Let us consider A2 (2,6), A7 (5,10), and A15 (6,11) as the centroids of the initial clusters. Hence, we will consider that:

Centroid 1=(2,6) is associated with cluster 1.
Centroid 2=(5,10) is associated with cluster 2.
Centroid 3=(6,11) is associated with cluster 3.

K-Means Example (Cont.)

Point	Distance from Centroid 1 (2,6)	Distance from Centroid 2 (5,10)	Distance from Centroid 3 (6,11)	Assigned Cluster
A1 (2,10)	4	3	4.123106	Cluster 2
A2 (2,6)	0	5	6.403124	Cluster 1
A3 (11,11)	10.29563	6.082763	5	Cluster 3
A4 (6,9)	5	1.414214	2	Cluster 2
A5 (6,4)	4.472136	6.082763	7	Cluster 1
A6 (1,2)	4.123106	8.944272	10.29563	Cluster 1
A7 (5,10)	5	0	1.414214	Cluster 2
A8 (4,9)	3.605551	1.414214	2.828427	Cluster 2
A9 (10,12)	10	5.385165	4.123106	Cluster 3
A10 (7,5)	5.09902	5.385165	6.082763	Cluster 1
A11 (9,11)	8.602325	4.123106	3	Cluster 3
A12 (4,6)	2	4.123106	5.385165	Cluster 1
A13 (3,10)	4.123106	2	3.162278	Cluster 2
A14 (3,8)	2.236068	2.828427	4.242641	Cluster 1
A15 (6,11)	6.403124	1.414214	0	Cluster 3

In cluster 1, we have 6 points i.e. A2 (2,6), A5 (6,4), A6 (1,2), A10 (7,5), A12 (4,6), A14 (3,8). To calculate the new centroid for cluster 1, we will find the mean of the x and y coordinates of each point in the cluster. Hence, the new centroid for cluster 1 is (3.833, 5.167).

In cluster 2, we have 5 points i.e. A1 (2,10), A4 (6,9), A7 (5,10) , A8 (4,9), and A13 (3,10). Hence, the new centroid for cluster 2 is (4, 9.6)

In cluster 3, we have 4 points i.e. A3 (11,11), A9 (10,12), A11 (9,11), and A15 (6,11). Hence, the new centroid for cluster 3 is (9, 11.25).

K-Means Example (Cont.)

Point	Distance from Centroid 1 (3.833, 5.167)	Distance from centroid 2 (4, 9.6)	Distance from centroid 3 (9, 11.25)	Assigned Cluster
A1 (2,10)	5.169	2.040	7.111	Cluster 2
A2 (2,6)	2.013	4.118	8.750	Cluster 1
A3 (11,11)	9.241	7.139	2.016	Cluster 3
A4 (6,9)	4.403	2.088	3.750	Cluster 2
A5 (6,4)	2.461	5.946	7.846	Cluster 1
A6 (1,2)	4.249	8.171	12.230	Cluster 1
A7 (5,10)	4.972	1.077	4.191	Cluster 2
A8 (4,9)	3.837	0.600	5.483	Cluster 2
A9 (10,12)	9.204	6.462	1.250	Cluster 3
A10 (7,5)	3.171	5.492	6.562	Cluster 1
A11 (9,11)	7.792	5.192	0.250	Cluster 3
A12 (4,6)	0.850	3.600	7.250	Cluster 1
A13 (3,10)	4.904	1.077	6.129	Cluster 2
A14 (3,8)	2.953	1.887	6.824	Cluster 2
A15 (6,11)	6.223	2.441	3.010	Cluster 2

In cluster 1, we have 5 points i.e. A2 (2,6), A5 (6,4), A6 (1,2), A10 (7,5), and A12 (4,6). To calculate the new centroid for cluster 1, we will find the mean of the x and y coordinates of each point in the cluster. Hence, the new centroid for cluster 1 is (4, 4.6).

In cluster 2, we have 7 points i.e. A1 (2,10), A4 (6,9), A7 (5,10) , A8 (4,9), A13 (3,10), A14 (3,8), and A15 (6,11). Hence, the new centroid for cluster 2 is (4.143, 9.571)

In cluster 3, we have 3 points i.e. A3 (11,11), A9 (10,12), and A11 (9,11). Hence, the new centroid for cluster 3 is (10, 11.333).

K-Means Example (Cont.)

Point	Distance from Centroid 1 (4, 4.6)	Distance from centroid 2 (4.143, 9.571)	Distance from centroid 3 (10, 11.333)	Assigned Cluster
A1 (2,10)	5.758	2.186	8.110	Cluster 2
A2 (2,6)	2.441	4.165	9.615	Cluster 1
A3 (11,11)	9.485	7.004	1.054	Cluster 3
A4 (6,9)	4.833	1.943	4.631	Cluster 2
A5 (6,4)	2.088	5.872	8.353	Cluster 1
A6 (1,2)	3.970	8.197	12.966	Cluster 1
A7 (5,10)	5.492	0.958	5.175	Cluster 2
A8 (4,9)	4.400	0.589	6.438	Cluster 2
A9 (10,12)	9.527	6.341	0.667	Cluster 3
A10 (7,5)	3.027	5.390	7.008	Cluster 1
A11 (9,11)	8.122	5.063	1.054	Cluster 3
A12 (4,6)	1.400	3.574	8.028	Cluster 1
A13 (3,10)	5.492	1.221	7.126	Cluster 2
A14 (3,8)	3.544	1.943	7.753	Cluster 2
A15 (6,11)	6.705	2.343	4.014	Cluster 2

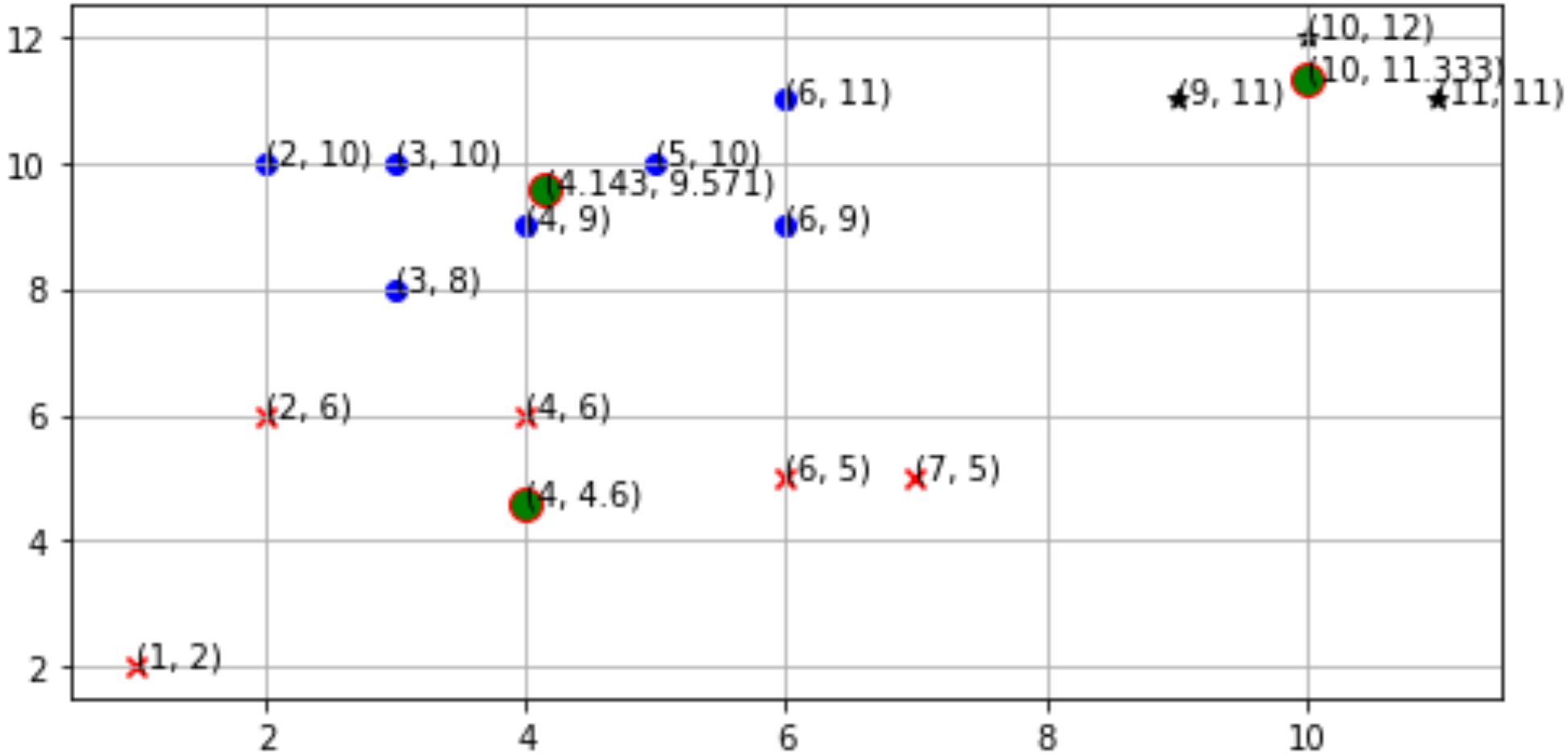
In cluster 1, we have 5 points i.e. A2 (2,6), A5 (6,4), A6 (1,2), A10 (7,5), and A12 (4,6). To calculate the new centroid for cluster 1, we will find the mean of the x and y coordinates of each point in the cluster. Hence, the new centroid for cluster 1 is (4, 4.6).

In cluster 2, we have 7 points i.e. A1 (2,10), A4 (6,9), A7 (5,10) , A8 (4,9), A13 (3,10), A14 (3,8), and A15 (6,11). Hence, the new centroid for cluster 2 is (4.143, 9.571)

In cluster 3, we have 3 points i.e. A3 (11,11), A9 (10,12), and A11 (9,11). Hence, the new centroid for cluster 3 is (10, 11.333).

Here, you can observe that no point has changed its cluster compared to the previous iteration. Due to this, the centroid also remains constant. Therefore, we will say that the clusters have been stabilized.

Plot for K-Means Clustering



Challenges with K-Means Clustering

- Choosing the Right Number of Clusters (k): One of the biggest challenges is deciding how many clusters to use.
- Sensitive to Initial Centroids: The final clusters can vary depending on the initial random placement of centroids.
- Non-Spherical Clusters: K-Means assumes that the clusters are spherical and equally sized. This can be a problem when the actual clusters in the data are of different shapes or densities.
- Outliers: K-Means is sensitive to outliers, which can distort the centroid and, ultimately, the clusters.

Applications of K-means Clustering

- Document Classification: Using k-means clustering, we can divide documents into various clusters based on their content, topics, and tags.
- Customer segmentation: Supermarkets and e-commerce websites divide their customers into various clusters based on their transaction data and demography. This helps the business to target appropriate customers with relevant products to increase sales.
- Cyber profiling: In cyber profiling, we collect data from individuals as well as groups to identify their relationships. With k-means clustering, we can easily make clusters of people based on their connection to each other to identify any available patterns.
- Image segmentation: We can use k-means clustering to perform image segmentation by grouping similar pixels into clusters.
- Fraud detection in banking and insurance: By using historical data on frauds, banks and insurance agencies can predict potential frauds by the application of k-means clustering.

Advantages of K-means Clustering

- Easy to implement: K-means clustering is an iterable algorithm and a relatively simple algorithm. In fact, we can also perform k-means clustering manually as we did in the numerical example.
- Scalability: We can use k-means clustering for even 10 records or even 10 million records in a dataset. It will give us results in both cases.
- Convergence: The k-means clustering algorithm is guaranteed to give us results. It guarantees convergence. Thus, we will get the result of the execution of the algorithm for sure.
- Generalization: K-means clustering doesn’t apply to a specific problem. From numerical data to text documents, you can use the k-means clustering algorithm on any dataset to perform clustering. It can also be applied to datasets of different sizes having entirely different distributions in the dataset. Hence, this algorithm is completely generalized.
- Choice of centroids: You can warm-start the choice of centroids in an easy manner. Hence, the algorithm allows you to choose and assign centroids that fit well with the dataset.