K-Means Clustering

Tuesday, 14 November 2023 10:55 PM

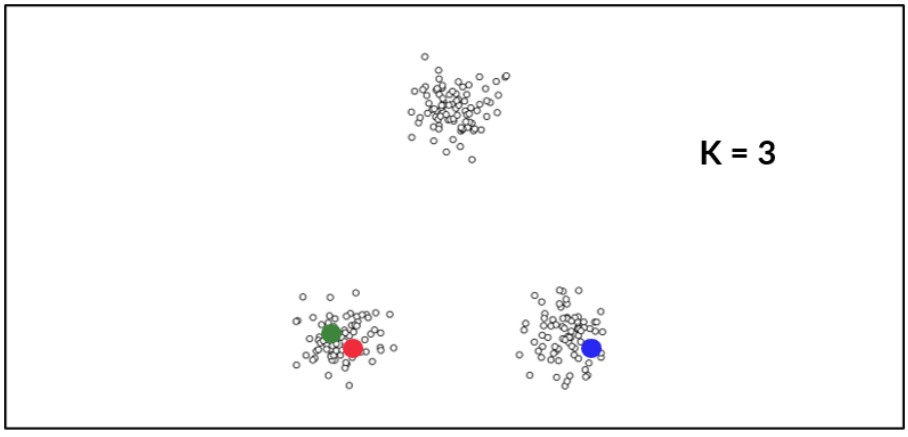
Suppose we have a dataset of N samples with P columns.

Assumptions : There are K-clusters available in the data.

So our objective is to find K-clusters present in the data. We assign each of these samples/observations to one of the clusters.

Algorithm →

- Input :
  - X : The input data with N rows, P columns.
  - K : The number of clusters to be identified.
- Initialization :
  - Each cluster K is represented by a cluster mean m_K.
  - Each cluster mean m_K is called a Centroid.
  - There are K-cluster means $m_1, m_2 \ldots m_K$.



All the observations 'X'

The three coloured dots are initial cluster means for $K=3$.

After initialization, the algorithm will repeat the following 2 steps →

- Step 1 – Assigning each observation to the nearest cluster means.
- Step 2 → Updating cluster means based on clusters to which observations are assigned.

These 2 steps will repeat until mean cluster assignments no longer change.

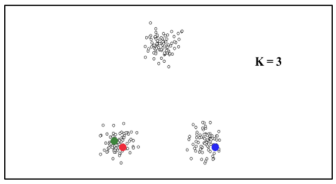


Fig. 1 - **Initial cluster means**

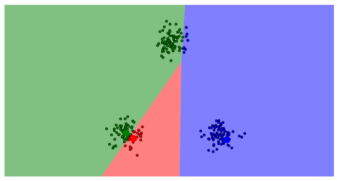


Fig. 2 – **Observations assigned to clusters**

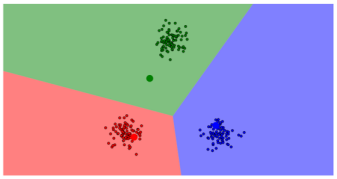


Fig. 3 - **Re-assigned clusters with updated cluster means**

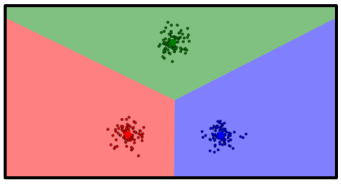


Fig. 4 - **Convergence**

Mathematical Intuition of the Algorithm →

Step 1 → Assignment Step

$$a_i = \text{argmin}_k \, d(x_i, m_k), \quad i = 1, \ldots, N$$

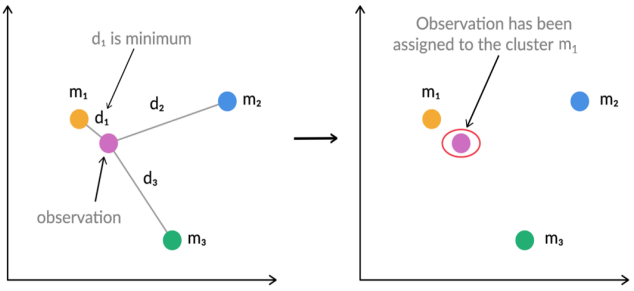
a_i : cluster assignment for observation i; it can be $1, 2, 3, \ldots K$

… have $d(x_i, m_k)$

$d(\cdot,\cdot)$ represents distance, [illegible] measures the distance between observation x_i and each cluster mean.

Based on the distance the argmin operator will assign an observation to that cluster for which the distance between the obser. and the respective cluster is minimum.

We use Euclidean Distance to calculate the distance here.



Similarly all the points will be assigned to one one of the clusters in this step.

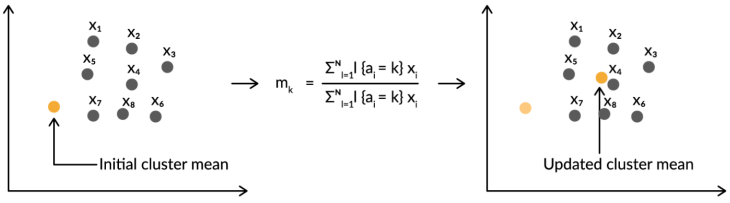
2→ Update Cluster Means

Suppose we stored all the cluster assignments in $\vec{a}$.

$$\vec{a} = [a_1\ a_2\ a_3\ a_4\ \ldots\ a_n]$$

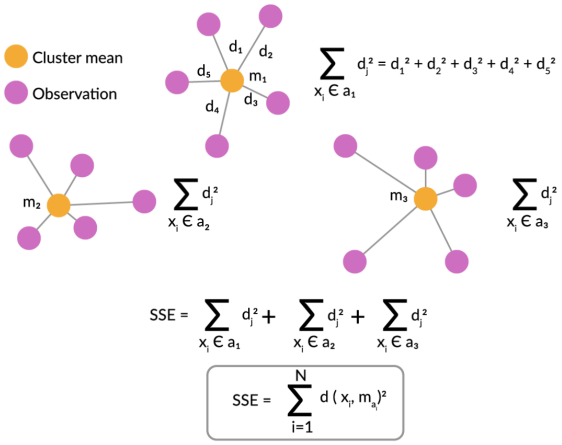
We take $\vec{a}$ as input and try to revise cluster means based on cluster assignments. We compute the average of all the observations

assigned to cluster K and make this average as new cluster mean.



Objective of K-Means

Minimize the 'Within cluster Sum of squares' (WCSS). It is the sum of squared distances for each observation to its assigned cluster mean.



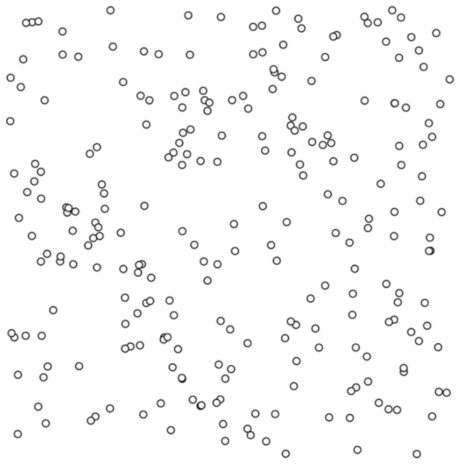
Effect of Initial Cluster Means on the final Cluster formed →

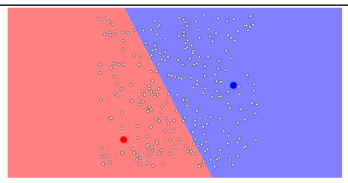
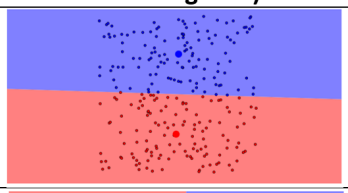
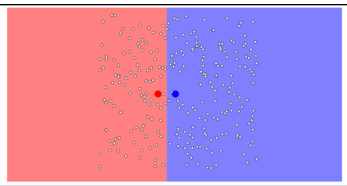
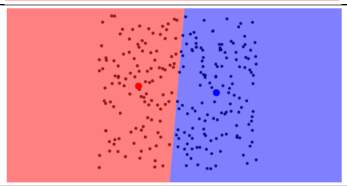
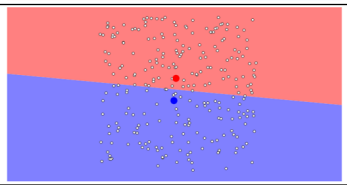
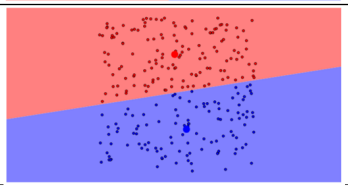
- K-Means is an iterative optimization algorithm.

→ It is sensitive to initial values for the cluster means

→ It only returns a local solution. The value of WCSS is not necessarily lowest possible

→ For better results, use multiple starting values for the cluster means. Then keep the solution that obtains the lowest WCSS.

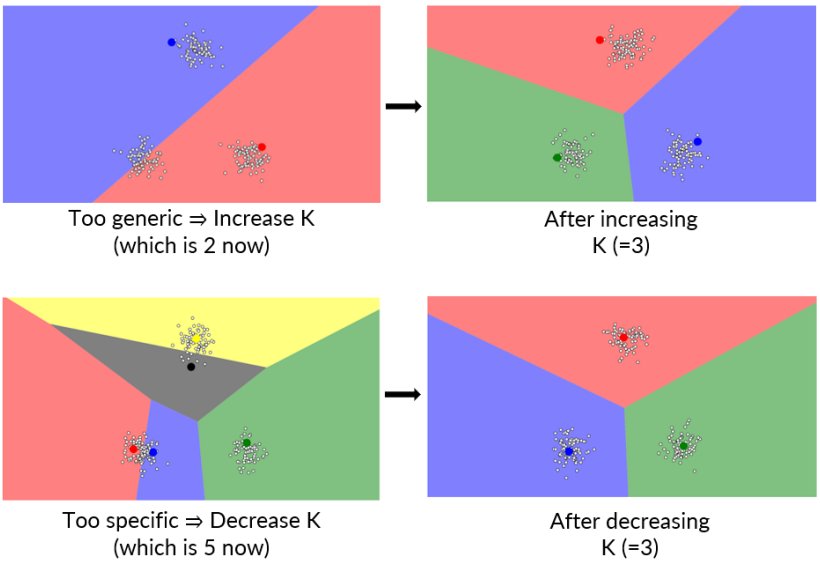


Case	Initial Cluster Means	Final Clusters (After Convergence)
Case 1		
Case 2		
Case 3		

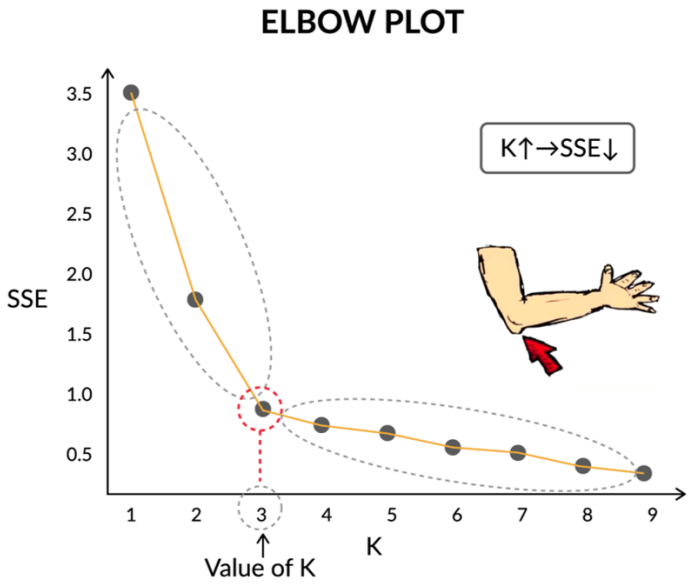
Selecting 'K' →

Two methods to select optimal K →

- Elbow Method.
- Silhouette Method.



* The Elbow Method →



Run the K-Means clustering algorithm for multiple values of K and plot the corresponding WCSS values. You get a curve called 'Elbow plot'.

* Pick the K-value at which the elbow bends.