**K-MEANS ALGORITHM**

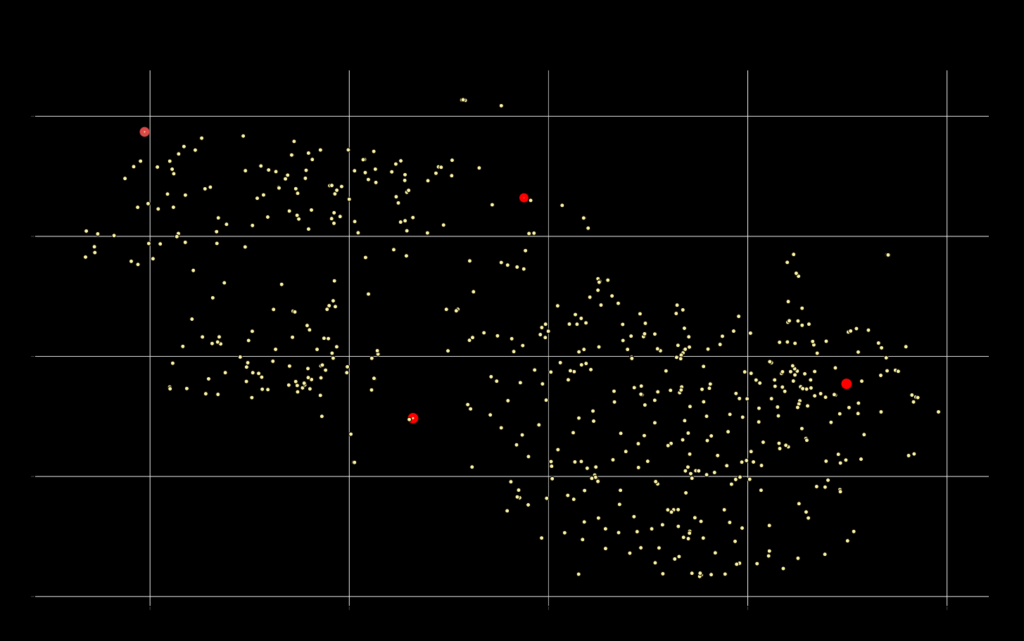
**1. What is k-means clustering?**

K-means clustering algorithm is an unsupervised machine learning algorithm for determining which group a certain object really belongs to. What it means by “being unsupervised” is that there are no prescribed labels in the data denoting its structure. The main idea is to assign each observation into the cluster with the nearest mean (centroid), serving as a prototype of the cluster.

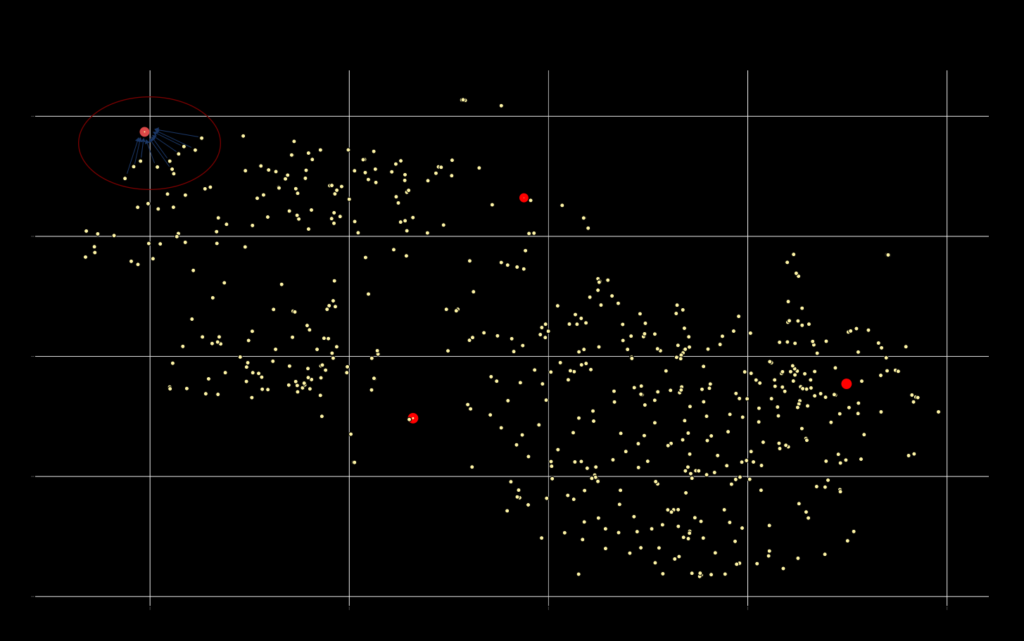
**2. How to perform k-means clustering**

Here are five simple steps for the k-means clustering algorithm and an example for illustration:

***Step 1:*** Visualize n data points and decide the number of clusters (k). Choose k random points on the graph as the centroids of each cluster. For this example, we would like to divide the data into 4 clusters, so we pick 4 random centroids.



***Step 2:*** Calculate the Euclidean distance between each data point and chosen clusters’ centroids. A point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other ones.



***Step 3***: After assigning all observations to the clusters, calculate the clustering score, by summing up all the Euclidean distances between each data point and the corresponding centroid.

Where:

k: the number of clusters

n: the number of points belonging to cluster j

cj: the centroid of cluster j

***Step 4:*** Define the new centroid of each cluster by calculating the mean of all points assigned to that cluster. Here’s the formula (n is the number of points assigned to that cluster):

***Step 5:*** Repeat from step 2 until the positions of the centroids no longer move and the assignments stay the same.

There you go: data points are now grouped into 4 different clusters. Using a simple idea of minimizing distances between data points to group them together, k-means clustering algorithm is extremely helpful for understanding the structure of the data, how observations are classified, and interpreting the story behind. K-means clustering has been widely used in data analysis, especially in life sciences, in analyzing thousands to millions of data points in single-cell RNA-seq and bulk RNA-seq experiments.

Note that the Euclidean metric measures the distance based on the vector connecting two points, and will cause some biases for data with different scales. For example, in RNA-seq data, gene expression values can range from as little as 0.001 to a thousand, stretching the data points along an axis. That is, the variable with the smaller scale will be easily dominated and play little in the convergence, as clusters will scatter along an axis only. For this reason, it is necessary to make sure that the variables are at the same scale before using k-means clustering.

**3. How to determine the optimal value of k?**

Note that before determining the number of clusters to assign the data into (the variable k), you should have an overview of the data and on what basis you want to group them. You can even apply a hierarchical clustering on the data first to briefly understand the structure of the data before choosing k by hand.

A well-known method to validate the number of clusters is the Elbow method, that is to run k-means clustering several times for a range of values of k (usually from 2 to 10) and pick out the value of k that causes sudden drop in the sum of squared distances. More specifically, for each value of k, we calculate the sum of squared distances (between each point and the corresponding centroid) and graph the results on a line chart. Choose the value where the sum of squares drops, giving an angle in the graph (a.k.a an elbow) - that is the optimal value of k.

