K-Means

- K-Means clustering is an simple and most popular unsupervised learning algorithm.
- K in K-means represent number of clusters in a given data.
- So what is clustering?
- Clustering is a technique which tries to divide the entire data into no of homogeneous cluster/groups based on similarity of the points.
- It always try to keep points in a group as similar as possible and points from different group as dissimilar as possible.
- Clustering is used in various fields like image recognition, pattern analysis, medical informatics, genomics, data compression, customer segmentation, Document clustering, recommendation engines etc...

K-Means

- Clustering: Types, Clustering can be broadly divided into two subgroups:
- Hard clustering: Here each data object or point either belongs to a cluster completely or not.
- Soft clustering: Here a data point can belong to more than one cluster with some probability or likelihood value.

K-Means

- **Clustering Algorithm: Types, Clustering algorithms can be broadly divided into multiple types:**
- **Connectivity-based clustering: Data points that are closer in the data space are more similar than to data points farther away. Then clusters are formed by connecting data points based on their distances and it is also called as hierarchical clustering(we discus more about this when we read about hierarchical clustering)**
- **Centroid-based clustering: Here clusters are represented by central vector or centroid(It can be a point in a dataset or not) We will be discussing more about this using K-Means algorithm.**

K-Means

- **Distribution-based clustering:** Clustering is based on the notion of how probable is it for a data point to belong to a certain distribution, such as the Gaussian distribution, for example.
- **Density-based methods:** search the data space for areas of varied density of data points. Clusters are defined as areas of higher density within the data space compared to other regions. Data points in the sparse areas are usually considered to be noise and/or border points. DBSCAN and OPTICS are some prominent density based clustering.

K-Means

- K-Means is a centroid based clustering algorithm.
- K-means uses an iterative refinement method to produce its final clustering based on the number of clusters defined by the user(K in K-means is the number or clusters).
- In this algorithm each point in a dataset will be in a group which is close to the centroid.
- We calculate the distance between each point and the centroids and choose the one which has the minimum distance.
- This is an iterative process where we change the centroids in each iteration and group the points.
- We try to minimize the distance between the data points, it is called local optimal solutions.
- We stop the iteration process when we see there is no change of centroid values or it reaches the maximum iteration number given by the user.

K-Means

- The Main objective of K-Means algorithm is to minimize the sum of distances between the data points with its respective cluster centroid, Known as intra-cluster variance
- Steps:
- 1. Choose the number of clusters (K)
- 2. Randomly select K points from the dataset as centroids.
- 3. Assign the data points to its nearest centroid using any distance measure algorithms.
- 4. Find the intra-cluster variance using the below formula
- sum(j)(sum(i)((X(ji) - C(j))^2)) (sum(j) is no of clusters, sum(i) is no of points in that cluster, C is the centroid of that cluster(J), X(ji) is the point X in the cluster(J))
- 5. Recalculate the centroids using the mean of all the points in the cluster and that mean will be the new centroid value.
- 6. Repeat step 3, 4, and 5. And stop based on the given criteria.

K-Means

- Stopping Criterias for the algorithm
- One way is when your centroid values doesnt change much from the previous centroids even after multiple iterations or getting same centroid value even after multiple iterations, that means algorithm is not learning any new patters and need to stop.
- Or when you see same data points in the same cluster even after multiple iterations.
- Or when the user gives the no of iterations count and when we reach that, we stop.

K-Means

- What if we don't know how many clusters are there in the given data and how to find the K?
- There is no universal answer for this, and although the optimal number of centroids or clusters is not known a priori, different approaches exist to try to estimate it.
- One most used approach is we try with different K values and calculate the intra-cluster variance.
- choosing the K value at which an increase will cause a very small decrease in the intra-cluster variance, while a decrease will sharply increase the error sum.'This point that defines the optimal number of clusters is known as the “elbow point”, and can be used as a visual measure to find the best pick for the value of K.

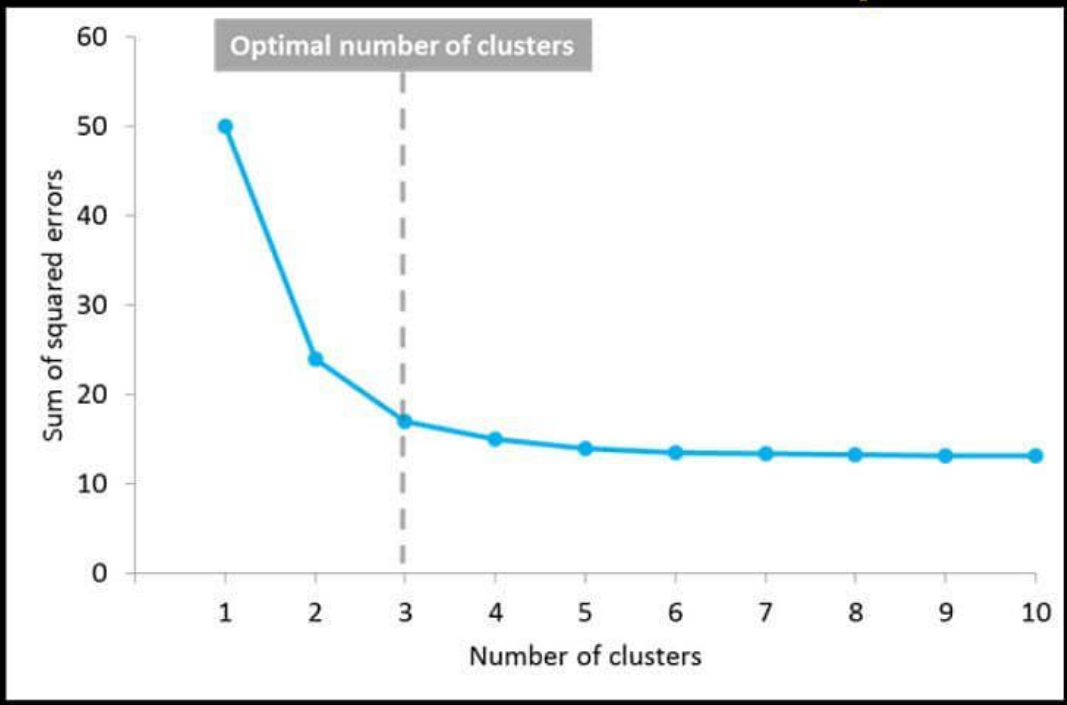
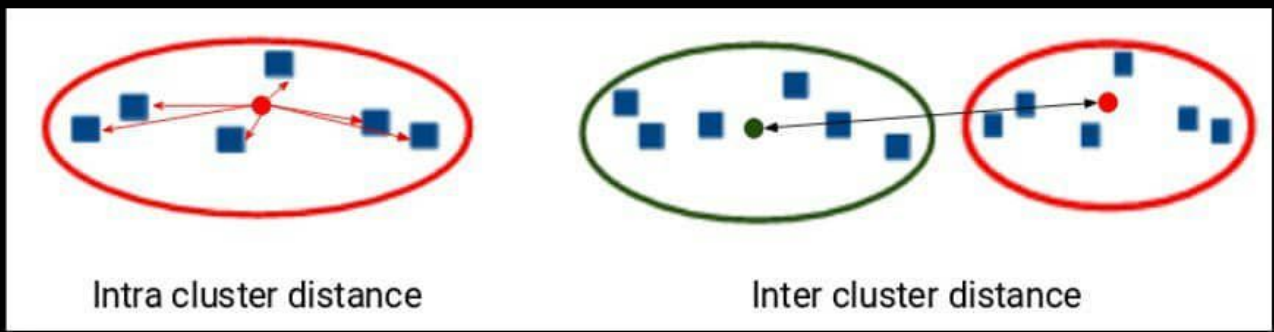


Image source - KdNuggets

K-Means

- How to evaluate clusters?
- If data has less dimensions like 2 or 3 then we can visualize the clusters and can say whether model did well or not. But in real world we don't see that 2 or 3 dimensions data always. So how to check if your model did well or not?
- One way is we already discussed is Intra-cluster distance. If we want data points in each cluster is as simila as possible then distance between them should be as low as possible. And we call this as inertia.
- And one more method is Dunn Index where we try to find the ratio of minimum Inter-cluster distance and max of Intra-cluster distance.

$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

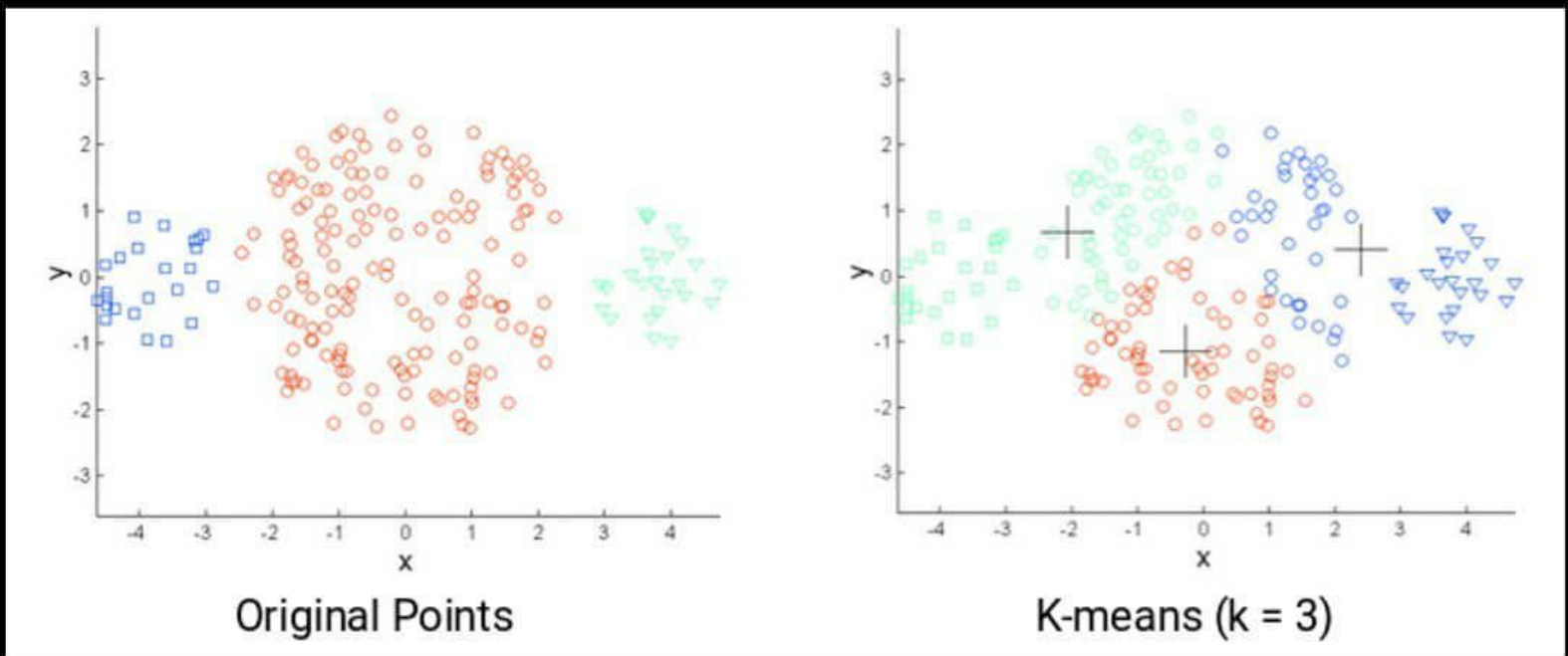


K-Means

- We always try to maximize the Dunn Index because if the min(Inter-cluster distance) is maximum i.e the small distance between the clusters should be high.
- And the max(intra-cluster distance) is mimimum i.e the clusters are very close.
- So if both the conditions satisfy then our model is very good.

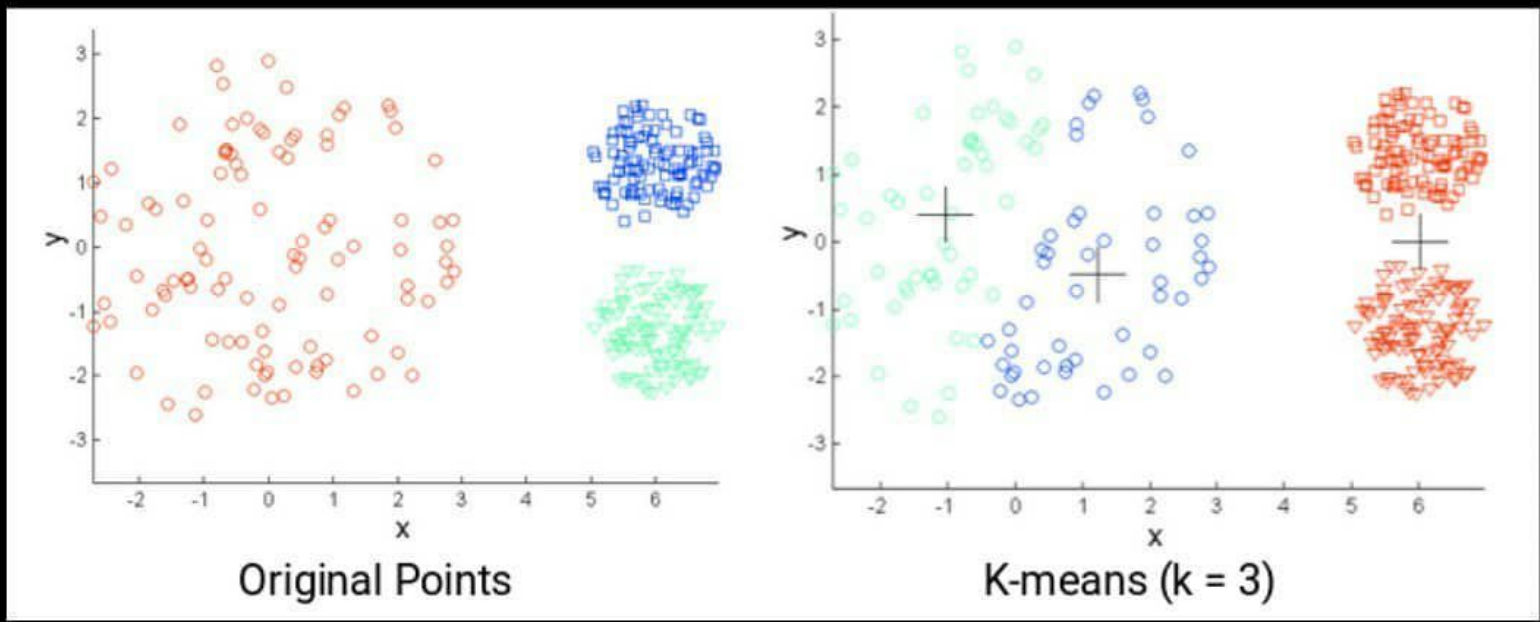
K-Means

- Downsides of K-Means
- One common challenge is the clusters are of unequal size or density.
- check the below image for more clarification.
- This is how k-means clusters the data when we have different sizes



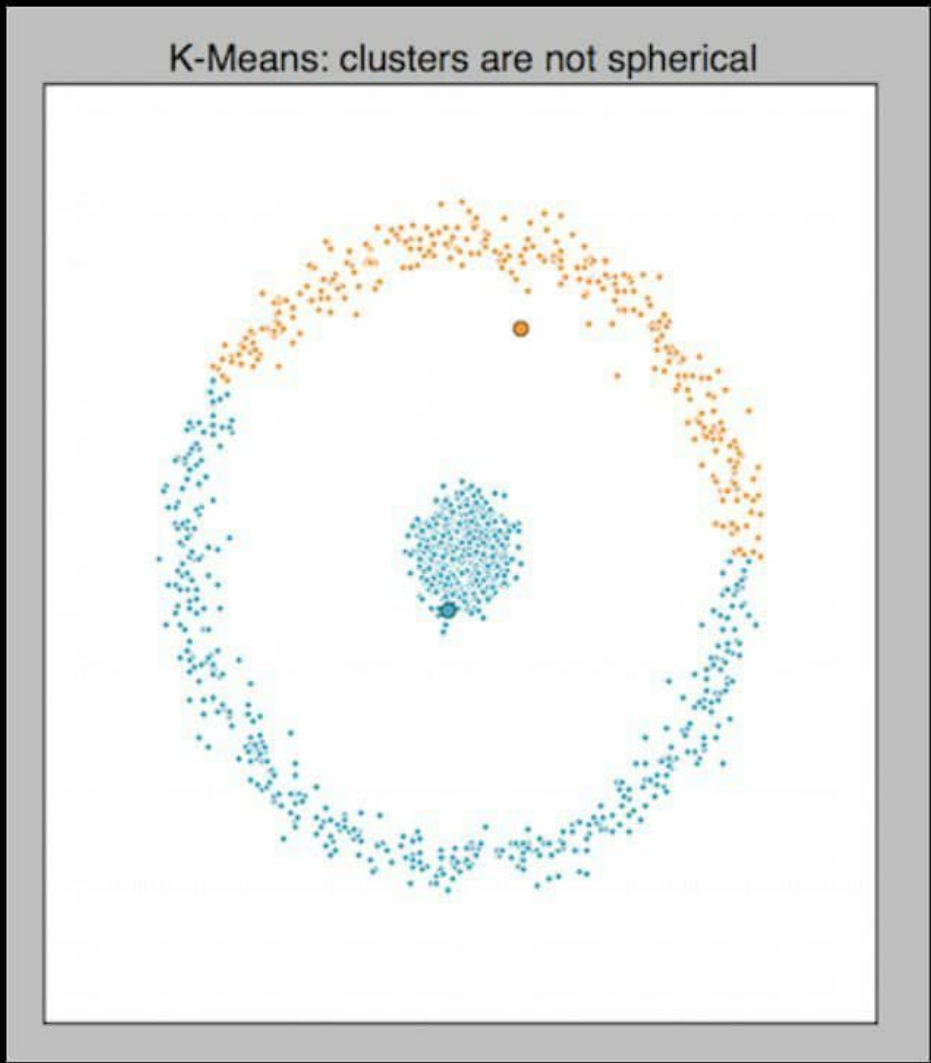
K-Means

- When the densities are different.
- Here, the points in the red cluster are spread out whereas the points in the remaining clusters are closely packed together.



K-Means

- When clusters are non-spherical
- This is how k-means clusters the data when we have sperical shapes.



K-Means

- When There are outliers in the data.
- If outliers are present in the dataset, they can influence clusterings results and change the outcome. The dataset should be pre-processed before applying k-means to detect and remove any outliers.
- And another problem is we always get different clusters because of random initialization of centroids.
- To overcome the above problem we use K-means++ algorithm.