AN INTUITION TO

K-Means Clustering

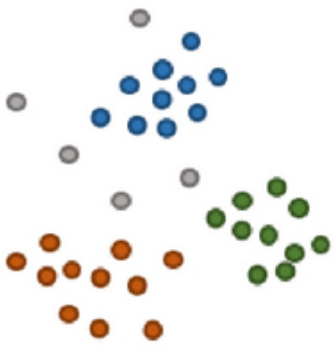
STARTING WITH UNSUPERVISED LEARNING

1.) WHAT IS UNSUPERVISED LEARNING

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. Unsupervised algorithms find patterns based only on input data. This technique is useful when we're not quite sure what to look for.

2.) CLUSTERING ALGORITHMS

Clustering Algorithms do the task of dividing the population or data points into a variety of groups such that data points within the same cluster are similar to other data points within the same cluster than those in other groups. Basically, the aim is to separate groups with similar traits and assign them into clusters.

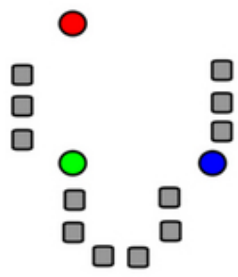


3.) K MEANS CLUSTERING

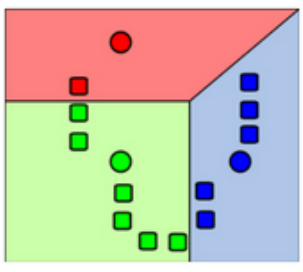
In this algorithm, we group the items into k clusters such that all items in the same cluster are as similar to each other as possible. And items not in the same cluster are as different as possible.

Distance measures(like Euclidean distance) are used to calculate similarity and dissimilarity between the data points. Each cluster has a centroid. Centroid can be thought as the point that is most representative of the cluster.

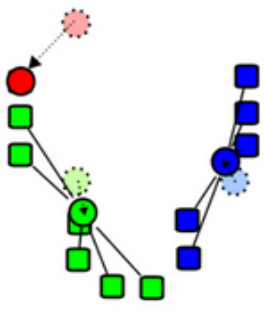
4.) HOW K-MEANS CLUSTERING WORKS



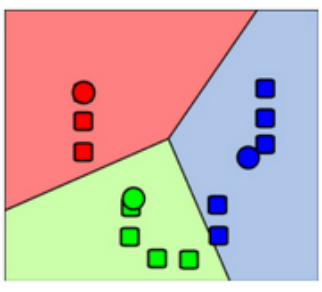
1. k initial "means" (in this case k=3) are randomly generated within the data domain.



2. k clusters are created by associating every observation with the nearest mean.



3. The centroid of each of the k clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function: