ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

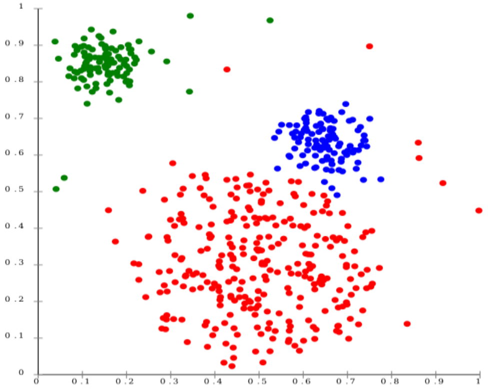
U LISBOA
UNIVERSIDADE
DE LISBOA

Carlos J. Costa

CLUSTERS ANALYSIS

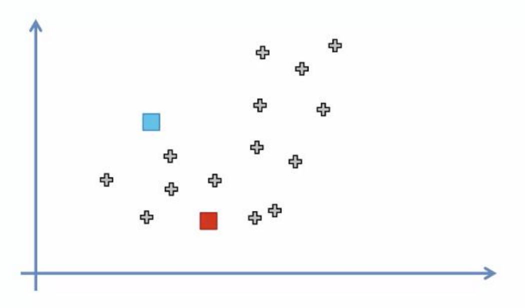
Cluster Analysis

- Cluster analysis is a multivariate method
- aims to classify a sample of subjects (or objects) into several different groups such that similar subjects are placed in the same group
- based on a set of measured variables

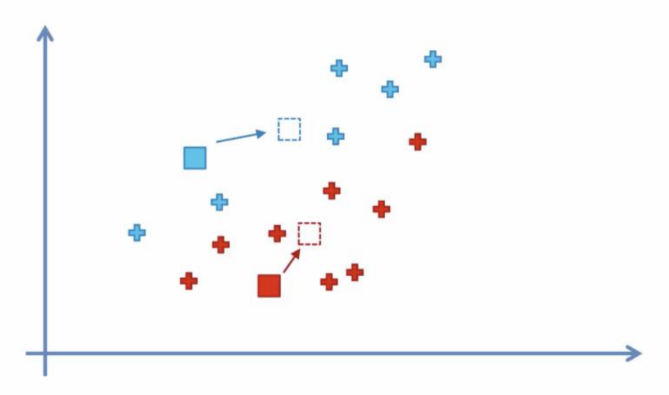


K-means Clustering

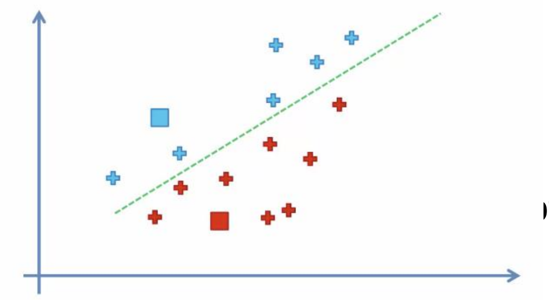
- 1.Select K (i.e. 2) random points as cluster centres called centroids



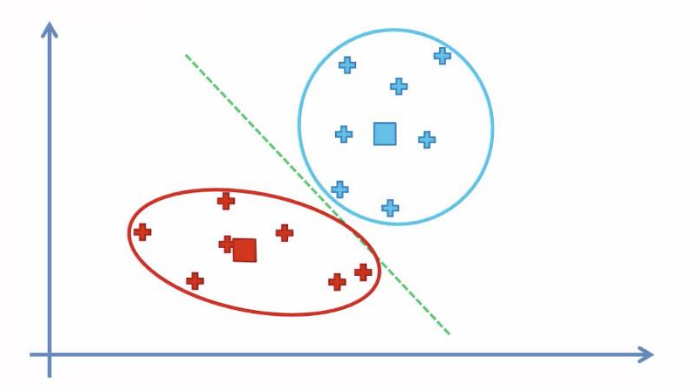
- 3. Determine the new cluster centre by computing the average of the assigned points



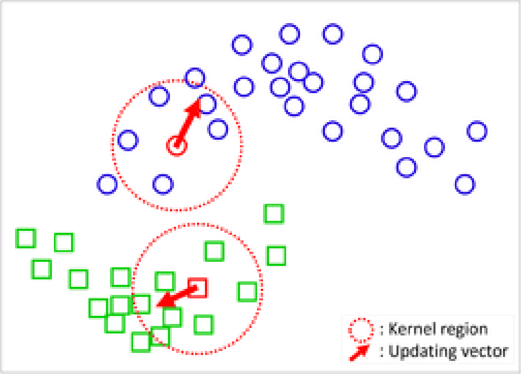
- 2. Assign each data point to the closest cluster by calculating its distance with respect to each centroid



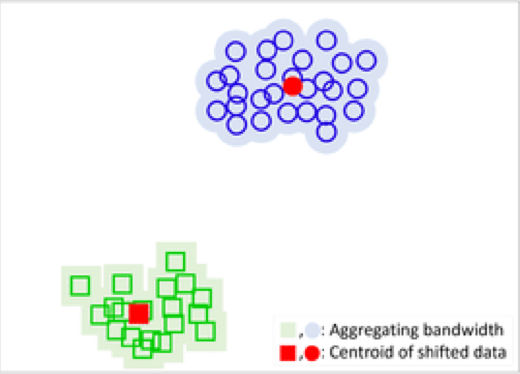
- 4. I f the
clu



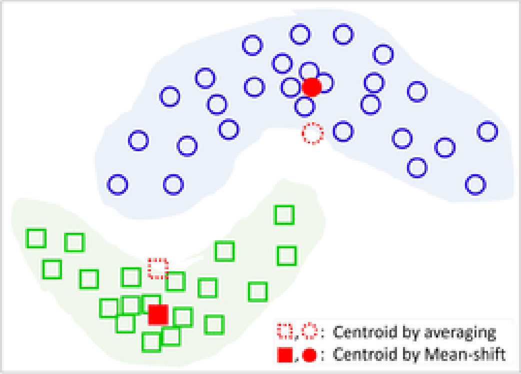
Means Shift Clustering



Updates (shifts) all data point toward
high density region until all the points converge



Aggregate the nearby shifted data points
into a cluster whose centroid is their average

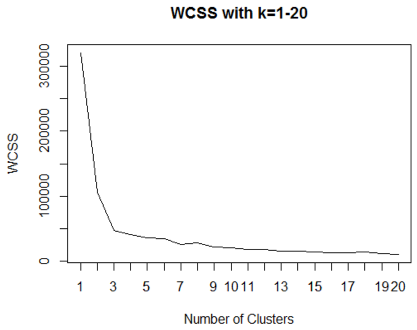


Assign the original data into the according clusters,
But keep the centroid calculated with shifted data

WCSS

- Within-Cluster-Sum-of-Squares (WCSS)- Implicit **objective function in k-Means** measures sum of distances of observations from their cluster centroids.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$



Yi is centroid for observation Xi.

- Given that k-Means has no in-built preference for right number of clusters, following are some of the common ways k can be selected:
  - Domain Knowledge
  - Rule of Thumb
  - Elbow-Method using WCSS
  - Cluster Quality using Silhouette Coefficient