Introduction to statistics and data analysis

Block III: Multivariate data analysis

Raúl Benítez
raul.benitez@upc.edu

Automatic Control Department
Universitat Politècnica de Catalunya

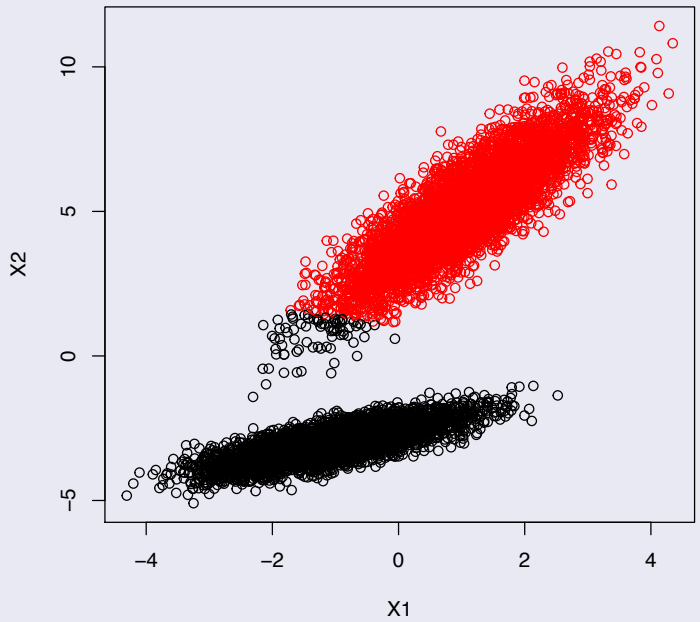
kmeans clustering (1/2 hour)

kmeans clustering

1. Specify the number of clusters k to be found in data.
2. Set initial values for the cluster centroids $\mu_1^0, \ldots, \mu_k^0$ (at random or prior knowledge).
3. Assign each observation to the nearest cluster (euclidean distance).
4. Recompute the centroid of each cluster from the assigned observations $\mu_1^1, \ldots, \mu_k^1$.
5. Repeat steps 3-4 until no change in the centroids. Provide final clustering $\mu_1^n, \ldots, \mu_k^n$, where n is the number of iterations.

Example: kmeans in R

2D features:



Hierarchical clustering (1/2 hour)

Hierarchical clustering

Sequence of partitions of the data into a set of clusters. They can be either agglomerative or Divisive:

- Divisive: Start with all observations in one cluster and slplit the clusters sequencially.
- Agglomerative: Start with as many cluster as observations and group them according to a cluster to cluster distance (linkage).

Consider two clusters r and s with n_r and n_s observations. A common cluster-cluster distances is:

$$d_{min}(r,s) = \min\{d(x_i^r, x_j^s)\},\ i = 1,\ldots,n_r; j = 1,\ldots,n_s$$

where $d(x_i^r, x_j^s)$ is the euclidean distance between observation i in cluster r and j in cluster s.

Hierarchical clustering measures

Other commonly used cluster-cluster distances are:

$$d_{max}(r,s) = \max\{d(x_i^r, x_j^s)\},\ i = 1,\ldots,n_r; j = 1,\ldots,n_s$$

$$d_{max}(r,s) = \frac{1}{n_s n_r}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} d(x_i^r, x_j^s)$$

Agglomerative Hierarchical clustering algorithm

1. Specify the number of clusters k to be found in data.
2. Initizlize the number of clusters to the number of observations N.
3. Group the nearest two clusters based on a cluster-to-cluster distance.
4. Recompute the centroid of each cluster from the assigned observations $\mu_1^1, \ldots, \mu_{N-1}^1$.
5. Repeat steps 3-4 until the number of clusters is k.

Example: Hierarchical clustering in R

Plot dendrogram: