Data Engineering
COMP2031/8031

- Topic Coordinator: Dr Mehwish Nasim
- Office: 3.13 Tonsley

Flinders
UNIVERSITY

Unsupervised Learning

Flinders UNIVERSITY

K-means

- defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

where:

- x_i is a data point belonging to the cluster C_k
- μ_k is the mean value of the points assigned to the cluster C_k

K-means

- Each observation (x_i) is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers (μk) is minimized.
- We define the total within-cluster variation as follows:

$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

- The *total within-cluster sum of square* measures the compactness (i.e goodness) of the clustering and we want it to be as small as possible.

Algorithm

- Specify the number of clusters (K) to be created (by the analyst)
- Select randomly k objects from the data set as the initial cluster centers or means
- Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
- For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a Kth cluster is a vector of length *p* containing the means of all variables for the observations in the kth cluster; *p* is the number of variables.
- Iteratively minimize the total within sum of square (Eq. 7). That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

Output

The output of `kmeans` is a list with several bits of information. The most important being:

- `cluster` : A vector of integers (from 1:k) indicating the cluster to which each point is allocated.
- `centers` : A matrix of cluster centers.
- `totss` : The total sum of squares.
- `withinss` : Vector of within-cluster sum of squares, one component per cluster.
- `tot.withinss` : Total within-cluster sum of squares, i.e. sum(withinss).
- `betweenss` : The between-cluster sum of squares, i.e. $totss-tot.withinss$.
- `size` : The number of points in each cluster.

Script

- Check the R script provided with this lecture

Other examples

Flinders UNIVERSITY

- K-means clustering with Twitter Data
- http://rstudio-pubs-static.s3.amazonaws.com/5983_af66eca6775f4528a72b8e243a6ecf2d.html