K-MEANS CLUSTERING – OVERVIEW (Steinley, 2006)

The K-means method is designed to partition two-way, two-mode data (that is, N objects each having measurements on P variables) into K classes ($C_1, C_2, \ldots, C_K$), where C_k is the set of n_k objects in cluster k, and K is given. If $\mathbf{X}_{N\times P} = \{x_{ij}\}_{N\times P}$ denotes the $N \times P$ data matrix, the K-means method constructs these partitions so that the squared Euclidean distance between the row vector for any object and the centroid vector of its respective cluster is at least as small as the distances to the centroids of the remaining clusters. The centroid of cluster C_k is a point in P-dimensional space found by averaging the values on each variable over the objects within the cluster. For instance, the centroid value for the jth variable in cluster C_k is

$$\bar{x}_j^{(k)} = \frac{1}{n_k}\sum_{i\in C_k} x_{ij}, \tag{2}$$

and the complete centroid vector for cluster C_k is given by

$$\bar{\mathbf{x}}^{(k)} = (\bar{x}_1^{(k)}, \bar{x}_2^{(k)}, \ldots, \bar{x}_P^{(k)})'. \tag{3}$$

SUMMARY of k-MEANS ALGORITHM (from Steinley, 2006)

(1) K initial seeds are defined by P-dimensional vectors $(s_1^{(k)}, \ldots, s_P^{(k)})$, for $1 \leq k \leq K$, and the squared Euclidean distance, $d^2(i, k)$, between the ith object and the kth seed vector is obtained:

$$d^2(i,k) = \sum_{j=1}^{P} (x_{ij} - s_j^{(k)})^2. \tag{4}$$

Objects are allocated to the cluster where (4) is minimum.

(2) After initial object allocation, cluster centroids are obtained for each cluster as described by (3), then objects are compared to each centroid (using $d^2(i, k)$) and moved to the cluster whose centroid is closest.

(3) New centroids are calculated with the updated cluster membership (by calculating the centroids after all objects have been assigned).

(4) Steps 2 and 3 are repeated until no objects can be moved between clusters.

MEASURING CLUSTER GOODNESS:

When attempting to find a 'good' partitioning of an object through the iterative method just described, it is of interest to note that we are also attempting to minimize a particular loss criterion, the error sum of squares (SSE):

$$SSE = \sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2. \tag{5}$$

Späth (1980, p. 72) noted that at times, but probably rarely in practice, *SSE* (also referred to as 'squared error distortion' in the pattern recognition literature; Gersho & Gray, 1992) may be further minimized by single object reallocation from one cluster to another. After the initial *K*-means algorithm is performed, a final inspection is made between all points and centroids. If there is an object within C_k such that

$$\frac{n_k}{n_k - 1} d^2(i,k) > \frac{n_{k^*}}{n_{k^*} + 1} d^2(i,k^*), \tag{6}$$

then move the *i*th object from C_k to cluster C_{k^*}, and *SSE* is reduced (see Späth, 1980, p. 72).