1 Linkage Methods

- To determine which observations will form the next cluster, we need to come up with a method for finding the distance between an existing cluster and individual observations, since once a cluster has been formed, we'll determine which observation will join it based on the distance between the cluster and the observation.

- Some of the methods that have been proposed to do this are to take the minimum distance between an observation and any member of the cluster, to take the maximum distance, to take the average distance, or to use some kind of measure that minimizes the distances between observations within the cluster.

- Each of these methods will reveal certain types of structure within the data.

- Using the *minimum* tends to find clusters that are drawn out and "snake"-like, while using the *maximum* tends to find compact clusters.

- Using the mean is a compromise between those methods.

- One method that tends to produce clusters of more equal size is known as ***Ward's method***. It attempts to form clusters keeping the distances within the clusters as small as possible, and is often useful when the other methods find clusters with only a few observations.

The following was written for SPSS users

- Having selected how we will measure distance, we must now choose the clustering algorithm, i.e. the rules that govern between which points distances are measured to determine cluster membership.
- There are many methods available, the criteria used differ and hence different classifications may be obtained for the same data. This is important since it tells us that, although cluster analysis may provide an objective method for the clustering of cases, there can be subjectivity in the choice of method.
- The linkage distances are calculated by SPSS. The goal of the clustering algorithm is to join objects together into successively larger clusters, using some measure of similarity or distance. SPSS provides seven clustering algorithms, the most commonly used one being ***Ward's method***.

1.1 Nearest neighbour method

(Also known as the single linkage method).
In this method the distance between two clusters is defined to be the distance between the two closest members, or neighbours. This method is relatively simple but is often criticised because it doesnt take account of cluster structure and can result in a problem called chaining whereby clusters end up being long and straggly. However, it is better than the other methods when the natural clusters are not spherical or elliptical in shape.

1.2 Furthest neighbour method

(Also known as the complete linkage method).
In this case the distance between two clusters is defined to be the maximum distance between members i.e. the distance between the two subjects that are furthest apart. This method tends to produce compact clusters of similar size but, as for the nearest neighbour method, does not take account of cluster structure. It is also quite sensitive to outliers.

1.3 Average (between groups) linkage method

(sometimes referred to as UPGMA).
The distance between two clusters is calculated as the average distance between all pairs of subjects in the two clusters. This is considered to be a fairly robust method.

1.4 Centroid method

Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is also fairly robust.

1.5 Wards method

In this method all possible pairs of clusters are combined and the sum of the squared distances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

1.6 Summary

-
-