Homework

1 Clustering

Clustering is a generic term for a range of methods aimed at identifying groups in a set of data. Clustering techniques assign a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense and dissimilar in other groups. In other words, objects in one cluster are likely to be different when compared to objects grouped under another cluster.

This introduces the idea of (dis)similarity, which is crucial to an understanding of how many methods of cluster analysis work. Many measures of (dis)similarity can be defined, contributing to the many methods of cluster analysis available. The most commonly used methods of clustering are the *Hierarchical* clustering.

2 Hierarchical Clustering

Each case is initially treated as a single cluster so are observations belong to a single cluster. The two most similar cases are merged to form a cluster of two cases. To merge clusters a measure to determine how similar clusters are is needed. Similarity can be defined in different ways. For example, similarity of two clusters is measured by the smallest distance between two cases, one from each cluster. The two clusters merged are those for which this smallest distance is smallest.