Contents

1 Review of Last Class

- Cluster analysis is a convenient method for identifying homogenous groups of objects called clusters. Objects (or cases, observations) in a specific cluster share many characteristics, but are very dissimilar to objects not belonging to that cluster.

- There are three cluster analysis approaches: hierarchical methods, partitioning methods (more precisely, k-means), and two-step clustering, which is largely a combination of the first two methods. In the last class we looked as hierarchical clustering analysis.

- Each of these procedures follows a different approach to grouping the most similar objects into a cluster and to determining each objects cluster membership.

- Some approaches most notably hierarchical methods require us to specify how similar or different objects are in order to identify different clusters. Most software packages, such as SPSS, calculate a measure of (dis)similarity by estimating the distance between pairs of objects. Objects with smaller distances between one another are more similar, whereas objects with larger distances are more dissimilar.

- An important problem in the application of cluster analysis is the decision regarding how many clusters should be derived from the data. This question is explored in the next step of the analysis. Sometimes, however, number of segments that have to be derived from the data will be known in advance.

- By choosing a specific clustering procedure, we determine how clusters are to be formed. (This always involves optimizing some kind of criterion, such as minimizing the within-cluster variance (i.e., the clustering variables overall variance of objects in a specific cluster), or maximizing the distance between the objects or clusters). The procedure could also address the question of how to determine the (dis)similarity between objects in a newly formed cluster and the remaining objects in the dataset.

- Hierarchical clustering procedures are characterized by the tree-like structure established in the course of the analysis. Most hierarchical techniques fall into a category called agglomerative clustering. In this category, clusters are consecutively formed from objects. Initially, this type of procedure starts with each object representing an individual cluster. These clusters are then sequentially merged according to their similarity. First, the two most similar clusters (i.e., those with the smallest distance between them) are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on. This allows a hierarchy of clusters to be established from the bottom up.

- A cluster hierarchy can also be generated top-down. In this divisive clustering, all objects are initially merged into a single cluster, which is then gradually split up. Divisive procedures are quite rarely used in practice. We therefore concentrate on the agglomerative clustering procedures.

- This means that if an object is assigned to a certain cluster, there is no possibility of reassigning this object to another cluster. This is an important distinction between these types of clustering and partitioning methods such as ***k-means***.

- In statistics, the occurrence of several variables in a multiple regression model are **closely correlated** to one another, and carrying the same information, more or less. Multicollinearity can cause strange results when attempting to study how well individual independent variables contribute to an understanding of the dependent variable, often undermining the analysis.
- In many analysis tasks, the variables under consideration are measured on different scales or levels. This would clearly distort any clustering analysis results. We can resolve this problem by ***standardizing*** the data prior to the analysis.
- Different standardization methods are available, such as the simple ***z standardization***, which re-scales each variable to have a mean of 0 and a standard deviation of 1.
- In most situations, however, ***standardization by range***(e.g., to a range of 0 to 1 or -1 to 1) is preferable. We recommend standardizing the data in general, even though this procedure can potentially reduce or inflate the variables influence on the clustering solution.