Contents

1 Agenda for Today's Class

- Review of Important Topics
- Review of K-Means Clustering (SPSS Exercise)
- Two-Step Clustering
- Review of Regression (Optional for Math Science Students)

2 Important Topics

- **Multi-collinearity**: Multi-collinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response. Examples of pairs of multi-collinear predictors are years of education and income, height and weight of a person, and assessed value and square footage of a house.
- **Consequences of high multicollinearity**: Multi-collinearity leads to decreased reliability and predictive power of statistical models, and hence, very often, confusing and misleading results.
- Multicollinearity will be dealt with in a future component of this course: Variable Selection Procedures.

3 Two-Step Cluster Analysis

When you have a really large data set or you need a clustering procedure that can rapidly form clusters on the basis of either categorical or continuous data, neither of the previous two procedures are entirely appropriate. Hierarchical clustering requires a matrix of distances between all pairs of cases, and k-means requires shuffling cases in and out of clusters and knowing the number of clusters in advance.

The Two-Step Cluster Analysis procedure was designed for such applications. The name two-step clustering is already an indication that the algorithm is based on a two-stage approach

- In the first stage, the algorithm undertakes a procedure that is very similar to the k-means algorithm.
- Based on these results, the two-step procedure conducts a modified hierarchical agglomerative clustering procedure that combines the objects sequentially to form homogenous clusters.

The Two-Step Cluster Analysis is an exploratory tool designed to reveal natural groupings (or clusters) within a data set that would otherwise not be apparent. The algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- Handling of categorical and continuous variables. By assuming variables to be independent, a joint ***multinomial-normal distribution*** can be placed on categorical and continuous variables. (Interesting, but not examinable).
- Automatic selection of number of clusters. By comparing the values of a ***model-choice criterion*** across different clustering solutions, the procedure can automatically determine the optimal number of clusters.
- Scalability. By constructing a ***cluster features*** (CF) tree that summarizes the records, the Two-Step algorithm allows you to analyze large data files. The Two-Step Cluster Analysis requires only one pass of data (which is important for very large data files).

3.1 Pre-clustering

In two-step clustering, to make large problems tractable, in the first step, cases are assigned to ***preclusters***. In the second step, the preclusters are clustered using the hierarchical clustering algorithm. You can specify the number of clusters you want or let the algorithm decide based on preselected criteria.

In general, the larger the number of sub-clusters produced by the pre-cluster step, the more accurate the final result is. However, too many sub-clusters will slow down the clustering during the second step.

The maximum number of sub-clusters should be carefully chosen so that it is large enough to produce accurate results and small enough not to slow down the second step clustering.

3.1.1 Step 1: Preclustering: Making Little Clusters

The first step of the two-step procedure is formation of preclusters. The goal of preclustering is to reduce the size of the matrix that contains distances between all possible pairs of cases. Preclusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed precluster or start a new precluster. When preclustering is complete, all cases in the same precluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of preclusters.

3.1.2 Step 2: Hierarchical Clustering of Preclusters

In the second step, SPSS uses the standard hierarchical clustering algorithm on the preclusters. Forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters.

4 Important Considerations for Two-Step Clustering

4.1 Cluster Features Tree

Two-Step Cluster Analysis is done by building a so-called ***cluster feature tree*** whose ***leaves*** represent distinct objects in the dataset. The procedure can handle categorical and continuous variables simultaneously and offers the user the flexibility to specify the cluster numbers as well as the maximum number of clusters, or to allow the technique to automatically choose the number of clusters on the basis of statistical evaluation criteria.

Additionally, the procedure indicates each variables importance for the construction of a specific cluster. These desirable features make the somewhat less popular two-step clustering a viable alternative to the traditional methods.

4.2 Types of Data

The Two-Step procedure works with both continuous and categorical variables. Cases represent objects to be clustered, and the variables represent attributes upon which the clustering is based.

4.3 Case Order

Note that the cluster features tree and the final solution may depend on the order of objects (or cases). To minimize order effects, randomly order the cases. It is recommended to obtain several different solutions with cases sorted in different random orders to verify the stability of a given solution. In situations where this is difficult due to extremely large file sizes, multiple runs with a sample of cases sorted in different random orders might be substituted.

5 SPSS Implementation

- To implement a Two-Step Cluster Analysis in SPSS, you use the following options: ***Analyze*** > ***Classify*** > ***TwoStep Cluster***.
- **Distance Measure** Log likelihood distance measures are the default; Euclidean distance can be used if all variables are continuous. (Log likelihood distance measures are not part of course).
- **Count of Continuous Variables** Continuous variables are standardized by default. The variables are standardized so that they all contribute equally to the distance or similarity between cases.
- **Number of clusters** You can specify the number of clusters, or you can let the algorithm select the optimal number based on either the Schwarz Bayesian criterion (BIC) or the Akaike information criterion (AIC).
- **Clustering Criterion** BIC and AIC are offered with the default being BIC.

5.1 Graphical Outputs

The lower part of the output indicates the quality of the cluster solution. The silhouette measure of cohesion and separation is a measure of the clustering solutions overall goodness-of-fit. It is essentially based on the average distances between the objects and can vary between -1 and +1. Specifically, a silhouette measure of less than 0.20 indicates a poor solution quality, a measure between 0.20 and 0.50 a fair solution, whereas values of more than 0.50 indicate a good solution . In our case, the measure indicates a satisfactory ("fair") cluster quality. Consequently, you can proceed with the analysis by double-clicking on the output. This will open up the model viewer , an evaluation tool that graphically presents the structure of the revealed clusters.

The model viewer provides us with two windows: the main view, which initially shows a model summary (left-hand side), and an auxiliary view, which initially features the cluster sizes (right-hand side). At the bottom of each window, you can request different information, such as an overview of the cluster structure and the overall variable importance.

6 More on Two-Step Clustering

6.1 Step 1: Pre-clustering: Making Little Clusters

The first step of the two-step procedure is formation of pre-clusters. The goal of pre-clustering is to reduce the size of the Distance matrix (the matrix that contains distances between all possible pairs of cases). Pre-clusters are just clusters of the original cases that are used in place of the raw data in the hierarchical clustering. As a case is read, the algorithm decides, based on a distance measure, if the current case should be merged with a previously formed pre-cluster or start a new precluster.

When preclustering is complete, all cases in the same precluster are treated as a single entity. The size of the distance matrix is no longer dependent on the number of cases but on the number of preclusters.

6.2 Step 2: Hierarchical Clustering of Preclusters

In the second step, SPSS uses the standard hierarchical clustering algorithm on the preclusters. Forming clusters hierarchically lets you explore a range of solutions with different numbers of clusters. Tip: The Options dialog box lets you control the number of preclusters. Large numbers of preclusters give better results because the cases are more similar in a precluster; however, forming many preclusters slows the algorithm.

7 Examining the Number of Clusters