TwoStep Cluster Analysis

The TwoStep Cluster Analysis procedure is an exploratory tool designed to reveal natural groupings (or clusters) within a dataset 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.

• 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 TwoStep algorithm allows you to analyze large data files.

Show me

Example. Retail and consumer product companies regularly apply clustering techniques to data that describe their customers' buying habits, gender, age, income level, etc. These companies tailor their marketing and product development strategies to each consumer group to increase sales and build brand loyalty.

Distance Measure. This selection determines how the similarity between two clusters is computed.

• Log-likelihood. The likelihood measure places a probability distribution on the variables. Continuous variables are assumed to be normally distributed, while categorical variables are assumed to be multinomial. All variables are assumed to be independent.

• Euclidean. The Euclidean measure is the "straight line" distance between two clusters. It can be used only when all of the variables are continuous.

Number of Clusters. This selection allows you to specify how the number of clusters is to be determined.

• Determine automatically. The procedure will automatically determine the "best" number of clusters, using the criterion specified in the Clustering Criterion group. Optionally, enter a positive integer specifying the maximum number of clusters that the procedure should consider.

• Specify fixed. Allows you to fix the number of clusters in the solution. Enter a positive integer.

Count of Continuous Variables. This group provides a summary of the continuous variable standardization specifications made in the Options dialog box. See the topic TwoStep Cluster Analysis Options for more information.

Clustering Criterion. This selection determines how the automatic clustering algorithm determines the number of clusters. Either the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) can be specified.