MSPA PREDICT 450-DL-55 LEC

Micro 1: Segmentation Analysis Kickstart

Darryl Buswell

1 What variables in the data file will you use as the "basis variables" for your segmentation analysis?

App Happy wants the analyst to undertake a 'general attitudinal post hoc segmentation analysis'. As such, attitudinal item variables will make up the basis variables for this analysis. Attitudinal item variables should be able to 'characterize' groups based on attitudes, therefore allowing the analyst to identify groups that might comprise of useful attitudinal segments.

From a review of the original survey, we find that Questions 24, 25 and 26 make the best candidates for basis variables. Question 24 relates to the respondents level of technological adoption or technological acceptance, Question 25 relates to personality characteristics, and Question 26 relates to purchasing behavior.

2 Will you treat your basis variables as continuous measures, categorical measures, or perhaps both?

The data used in cluster analysis can be of continuous or categorical type, however the type of data will influence the use of dissimilarity (or distance) metric. For continuous variables, the most common distance measure is the Euclidean distance, while for categorical or mixed type data, a Gower distance metric would be more appropriate (B. Everitt 2011).

All three of the identified basis variables employ a six level Likert scale: Agree Strongly, Agree, Agree Somewhat, Disagree Somewhat, Disagree, and Disagree Strongly. In that case, each variable can be considered to be of categorical type. Although we have the option of using a numeric equivalent of these variables, there are still obvious risks in assuming the data to be of continuous type.

Fortunately, the 'daisy' function as part of the 'cluster' package within R is able to compute distances using either a Euclidean or Gower distance metric (A. Struyf). As such, we will use the numeric equivalent of these variables from the numeric dataset, and leverage the Gower distance metric where possible.

3 What two (or more) clustering algorithms will you use?

Clustering algorithms can be separated into two main classes, hierarchical methods and non-hierarchical (or k-means clustering) methods. These methods are largely distinguished by how they derive a desired (or optimal) number of clusters. For hierarchical methods, the number of clusters is determined based on a dendrogram, while for non-hierarchical methods, the desired number of clusters is specified by the analyst in advance.

There are R packages available for both hierarchical and non-hierarchical clustering methods. For hierarchical clustering, the 'agnes' function as part of the 'cluster' package is able to perform unweighted pair-grouping, single linkage, or complete linkage clustering methods (Roudier). This package also has the benefit of being able to accept a dissimilarity matrix object directly, making its use alongside the 'daisy' function an attractive option.

For non-hierarchical clustering, the 'k-means' function as part of the 'stats' package is one option (Zurich). However, this function does not support the ability to directly pass a dissimilarity matrix object. An alternative may be the 'pam' function as part of the 'cluster' package (M. Maechler) which will in-fact accept a dissimilarity matrix object directly, providing a cluster partitioning around medoids.

4 How many cases in the data have one or more missing values on your basis variables?

Below includes a list of the number of observations (respondents) which recorded an NA within either the set of attitudinal (basis) variables, non-basis variables or within the dataset:

```
## [1] "Basis observations: 0"

## [1] "Non-basis observations: 585"

## [1] "Dataset observations: 585"
```

Reference

A. Struyf, & P. Rousseeuw, M. Hubert. “Dissimilarity Matrix Calculation.” http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/daisy.html.

B. Everitt, M. Leese, S. Landau. 2011. “Cluster Analysis.” John Wiley & Sons.

M. Maechler, & M. Studer. “Partitioning Around Medoids.” https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html.

Roudier, P. “Agglomerative Nesting (Hierarchical Clustering).” http://stat.ethz.ch/R-manual/R-patched/library/cluster/html/agnes.html.

Zurich, ETH. “K-Means Clustering.” https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html.