CHAPTER 2: CLUSTER ANALYSIS

Data used in Chapter 2 are provided in ASCII (*.txt). The SPSS (v10.0) syntax for carrying out the analysis presented in the book is also given (*.sps) together with the SPSS output file (*.spo). SPSS output files (*.spo) can only be viewed in SPSS, but output files are also provided in pdf format.

Note that the first line of each syntax file will need to be edited to give the correct location of the data file.

The SPSS syntax given for the examples below uses either nearest neighbour (single linkage) or farthest neighbour (complete linkage). You should experiment with different methods.

The dendrograms and icicle plots given in Chapter 2 were redrawn for the book. You may wish to do the same, as the SPSS graphics, particularly for icicle plots, leave much to be desired

archer.txt

The data are similarities between 24 carvings of Persian archers. The similarity matrix is given in Table 2.13. The file *archer.txt* contains the lower triangle of the similarity matrix. Syntax for reading the similarity matrix into SPSS and carrying out (farthest neighbour or Complete) Cluster Analysis is given in *archerCLA.sps* and the SPSS output is in *archerCLA.spo*. To carry out nearest neighbour Cluster Analysis (or any other method) replace 'COMPLETE' by 'SINGLE' (or other method) in the CLUSTER command.

Table 2.13 gives the similarities and the text explains how they are derived from the data matrix in Table 2.12. Figure 2.13 shows the dendrogam for complete linkage. Figures 2.12 and 2.14 illustrate a single archer and the set of 24 archers respectively. Figure 2.14 is an example of an alternative simple method of clustering for these data.

The similarities between the carvings are also analysed in Chapter 3 using MDS.

dialect.txt

The data are from a study by the University of Leeds on English dialects. The similarities in dialects between pairs of 25 villages are given in Table 2.9. The file *dialect.txt* contains the lower triangle of the similarity matrix. Syntax for reading the similarity matrix into SPSS and carrying out (nearest neighbour) Cluster Analysis scaling is given in file *dialectCLA.sps*. The file *dialectCLA.spo* gives output for both the nearest neighbour and for the farthest neighbour Cluster Analyses. To carry out farthest neighbour Cluster Analysis (or any other method) replace 'SINGLE' by 'COMPLETE' (or other method) in the CLUSTER command.

This is the main example used to explain CLA. The similarities between villages are given in Table 2.9 and a map of the villages in Figure 2.6. Dendrograms and an icicle plot are given in Figures 2.7, 2.9, and 2.8 respectively.

The similarities between the villages are also analysed in Chapter 3 using MDS.

attabsim.txt

The data are similarities between attitude to abortion response patterns. Each of 379 respondents agreed or disagreed with each of four items or statements about abortion, thus giving one of 16 possible response patterns. The response patterns and their frequencies are shown in Table 2.15 and the similarities calculated from them are shown in Table 2.16. The full matrix of similarities is given in *attabsim.txt*. The file *attabsim.sps* gives the syntax for (farthest neighbour) Cluster Analysis. To carry out nearest neighbour Cluster Analysis (or any other method) replace 'COMPLETE' by 'SINGLE' (or other method) in the CLUSTER command.

The response patterns and their frequencies are shown in Table 2.15 and the similarities calculated from them are shown in Table 2.16.

The response patterns are also analysed in Chapters 8 and 10 using latent variable and latent class models.

educ.txt

This is an example where variables, rather than cases are clustered. The data are correlations between variables relating to home and school circumstances of children. The file *educ.txt* contains the full matrix of correlations, which we use as similarities.

The variables are coded as follows:

X1=Parental circumstances in 1964
X2=Details of class teacher in 1964
X3=School-parent interaction in 1964
X4=Girl's attitude in 1964
X5=Test score in 1964
X6=Type of school in 1968
X7=Parental circumstances in 1968
X8=School-parent interaction in 1968
X9=Test score in 1968

The syntax for reading these correlations/similarities, and for single (nearest neighbour) and for complete (farthest neighbour) linkage cluster analysis are given in *educCLA.sps*. The file *educCLA.spo* (and *educCLA.pdf*) gives the output for nearest neighbour and for farthest neighbour cluster analysis, including both icicle plots and dendrograms.

The correlation matrix, and a more elegant version of the nearest neighbour dendrogram are given in Table 2.17 and Figure 2.17.