Chapter 13

Clustering

13.1 Introduction

Clustering is used to classify the n cases into k groups. Discriminant analysis is a type of supervised classification while clustering is a type of unsupervised classification.

For k-means clustering, there are 4 steps.

1) Partition the n cases into k initial groups and find the means of each group. Alternatively, choose k initial seed points. These are groups of size 1 so the mean is equal to the seed point.

2) Compute distances between each case and each mean. Assign case to the cluster whose mean is the nearest.

3) Recalculate the mean of each cluster.

4) Go to 2) and repeat until no more reassignments occur.

Two problems with k-means clustering are i) there could be more or less than k clusters, and ii) two initial means could belong to the same cluster. Then the resulting clusters may be poorly differentiated.

Hierarchical clustering also has several steps. A distance is needed. Single linkage (or nearest neighbor) is the minimum distance between cases in cluster i and cases in cluster j. Complete linkage is the maximum distance between cases in cluster i and cases in cluster j. The average distance between clusters is also sometimes used.

1) Start with m = n clusters. Each case forms a cluster. Compute the distance matrix for the n clusters. Let $d_{U,V}$ be the smallest distance. Combine clusters U and V into a single cluster and set $m = n - 1$.

2) Repeat step 1) with the new m. Continue until there is a single cluster.

3) Plot the resulting clusters as a dendogram. Use the dendogram to select k reasonable clusters of cases.

13.2 Complements

Atkinson, Riani and Cerioli (2004, ch. 7) has some interesting ideas.

13.3 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

13.1*.

R/Splus Problems

Warning: Use the command *source("G:/mpack.txt")* **to download the programs. See Preface or Section 15.2.** Typing the name of the `mpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.