

Statistical Data Mining, October 2001

Practicals Information

Getting Started

You will be using S-PLUS 6 for Windows, and we will assume that you have used it or S-PLUS 2000 before. If not, please ask for some extra start-up help.

You will need both to enter commands in the commands window and to run scripts from script windows.

Open a commands window (if one is not already open) by clicking on the button  on the upper toolbar.

We will be using the following libraries:

SDM datasets and functions for this course.

MASS *Modern Applied Statistics with S-PLUS* — Venables & Ripley

nnet Neural networks, multinomial models

class Classification — kNN and LVQ

rpart Recursive partitioning, by Terry Therneau and Beth Atkinson

The middle three ship with S-PLUS: see the notes for where to get copies of the other two. You can load libraries by the Load Library... dialog box from the File menu or by `library(SDM)`

in the commands windows or from a scripts window (see below).

If you want to use menus, you will find that MASS, nnet and rpart have menu/dialog-box interfaces (for rpart contributed by Patrick Aboyoun). You can make the menus available by running (once for a particular working directory) by

```
addMassMenus()
```

```
addNnetMenus()
```

```
addRpartMenus()
```

Explore to see what is there: GUIs are supposed to be intuitive! If you want to remove them, use `removeMassMenus()` and so on.

Scripts

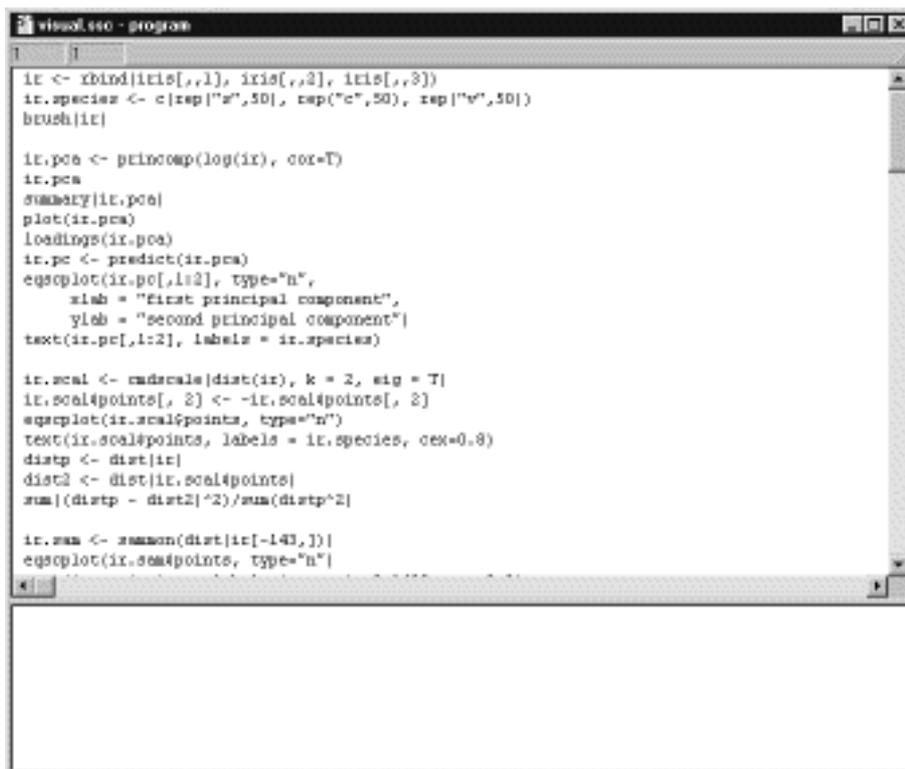
We suggest you work in a script window, although you can equally well work in the commands window. Script windows can be opened from **New** or **Open** on the **File** menu, and should be associated with file extension `.ssc` and so open in **S-PLUS** by double-clicking on them.

We have provided five scripts, one for each session, called

```
visual.ssc    trees.ssc
neural.ssc    kNN.ssc        assess.ssc
```

These are in the `scripts` folder of the `SDM` library. You should be able to load these into **S-PLUS**, launching it if necessary, by double-clicking on the icon in **Explorer**. (Please save any changes to a copy on your own local drive.)

These scripts contain the **S** commands in the course notes together with a few extra lines, for example to set screen layouts. They also contain commands to make the figures not described in the notes, notably for the visualization of the virus data, and some further examples for the neural networks and performance assessment sessions.



```
1 | 1
ir <- rbind(iris[,1], iris[,2], iris[,3])
ir.species <- c(rep("s",50), rep("c",50), rep("v",50))
brush(ir)

ir.pca <- princomp(log(ir), cor=T)
ir.pca
summary(ir.pca)
plot(ir.pca)
loadings(ir.pca)
ir.pc <- predict(ir.pca)
eqsplot(ir.pc[,1:2], type="n",
        xlab = "first principal component",
        ylab = "second principal component")
text(ir.pc[,1:2], labels = ir.species)

ir.scal <- cmdscale(dist(ir), k = 2, sig = T)
ir.scal$points[, 2] <- -ir.scal$points[, 2]
eqsplot(ir.scal$points, type="n")
text(ir.scal$points, labels = ir.species, cex=0.8)
distp <- dist(ir)
dist2 <- dist(ir.scal$points)
sum((distp - dist2)^2)/sum(distp^2)

ir.som <- somon(dist(ir[-143,]))
eqsplot(ir.som$points, type="n")
```

Figure 1: A script window.

Using a script window

A script window is divided into two panels (Figure 1). **S** commands can be typed into the top window and edited there. Groups of commands can be selected (in the usual ways

in **Windows**, perhaps easiest by dragging across the text with the left mouse button depressed), and submitted by pressing the function key **F10** or by the leftmost button on the lower toolbar when the window has focus (marked to represent the 'play' key on a VCR). If text output is produced, this will appear in the bottom part of the subwindow.

It is the help features that mark a scripts window as different from a commands window. Select a function name by double-clicking on it. Then help on that function is available by pressing the function key **F1**, and the right-click menu has items **Show Dialog...** and **Expand Inplace** to pop-up a dialog box for the arguments of the function and to paste in the function body.

More than one script window can be open at once. To avoid cluttering the screen script windows can be hidden (and unhidden) from the **Windows** file menu. The **Hide** item hides the window which has focus, whereas the **Unhide...** provides a list of windows from which to select.

Exploratory Projection Pursuit

For projection pursuit we hope to use the program **GGobi** (www.ggobi.org), but it was not released in final form when these notes were written.

If that is not available, I will demo using the program **XGobi** produced at Bellcore, which runs under **X11** and has been ported to **Windows** by **BDR**.

Further details at the time of the course.

Further Exercises

Visualization

1. There is a dataset called FT in library SDM. This is the *Financial Times* rankings of ‘universities’¹ on 16 features and an overall rating. An exercise in visualization.
2. Dataset crabs can be examined within S-PLUS as well as in GGobi.

Neural Networks

1. Data frame biopsy contains data on 699 biopsies of breast tumours, which have been classified as benign or malignant. The nine variables on each biopsy are a rating (1 to 10) by the coordinating physician; ratings on one variable are missing for some biopsies. (On-line help is available on the dataset, which is in library MASS.)

Fit logistic regressions and neural networks to find a rule to classify tumours based solely on the biopsy variables.

Near-neighbour Methods

1. Apply near-neighbour methods to the biopsy data frame.

Tree-based Methods

1. Examine in more detail the trees produced to predict the type of forensic glass in data frame fg1 by both tree and by rpart. In particular, investigate the effect of large the original tree is grown on the size and performance of pruned trees.
2. Apply classification trees to the biopsy data frame.

Assessing Performance

1. The script assess.ssc contains code to produce calibration plots for various neural network fits to the two-class synthetic data example. Try these out.
2. Assess the performance of your fits to the biopsy prediction problem.

¹including some sub-units such as colleges of and even schools of the University of London.