Data Analysis: Assignment 2: Deadline: 2024-02-05: 13:30

Exercise 1: Principal Component Analysis

The data set `pendigits.txt` contains data on pen-based handwritten digits. The data were collected from 44 writers, each of whom wrote 250 examples of the digits 0,1,2,...,9 in a random order. The digits were written inside boxes of 500×500 pixels on a pressure sensitive tablet. Unknown to the writers, the first 10 digits were ignored as writers became familiar with the input device.

The raw data on each of the $n = 10992$ characters consisted of a sequence $(x_t, y_t) : t = 1, \ldots, T$ of tablet coordinates of the pen at fixed time intervals of 100 milliseconds, where (x_t, y_t) were integers in the range $0 - 500$. The data were then normalised to make them invariant to translation and scale distortions. The new coordinates had maximum range between 0 and 100. Then 8 regularly spaced measurements (x_t, y_t) were chosen. This gave a total of 16 input variables. Columns 1-16 denote the variables, column 17 is the class code, 0 - 9. These are the only columns of interest.

1. Compute the variance of the 16 variables and show that they are very similar.
2. Carry out a PCA using the covariance matrix.
3. How many PCs explain 80% resp. 90% of the total variation in the data?
4. Display the first three PCs using pairwise scatterplots.
5. Carry out a PCA using the correlation matrix. Is there any substantial difference?pendigits
6. Draw the scree plots, for PCA using covariance and for correlation. How many PCs would you use based on this?
7. Is there ill-conditioning in the data matrix? Base your answer on the PCA.

Exercise 2: Mantel Randomisation

Consider the data for ozone measurements from thirty two locations in the Los Angeles area, found in the file `ozone.csv` in the course data directory. Perform a Mantel test to see whether the differences between ozone measurements are smaller for stations that are closer together.

Exercise 3: Clustering

The data in `primate.scapulae.txt` (and `primate.scapulae.xls`) contain indices and angles that are related to scapular shape (shoulder bones of primates), but not to functional meaning. There are 8 variables in the data set. The first five (AD.BD, AD.CD, EA.CD, Dx.CD, SH.ACR) are indices and the last three (EAD, β, γ) are angles. Of the 105 measurements on each variable, 16 were taken on *Hylobates* scapulae, 15 on *Pongo* scapulae, 20 on *Pan* scapulae. 14 on *Gorilla* scapulae, and 40 on *Homo* scapulae. The angle γ was not available for *Homo*.

1. Apply agglomerative and divisive hierarchical methods for clustering the variables using all 5 indices and the 2 angles available for all items. Which linkage methods give outliers?

2. Find the five-cluster solutions for these methods. Construct confusion tables and compute the misclassification rate. Which method gives the lowest rate? Which gives the highest rate?

Exercise 4: Doctor Visits Data

Consider the `DoctorVisits` data in the **AER** package. Use a Poisson regression for the number of visits. Is the Poisson model satisfactory? If not, where are the problems and what can be done about them? (**Note** Please note the limitations of diagnostics for count data. For example, if we have $X_1, \ldots, X_n$ i.i.d. Bernoulli($\frac{1}{2}$), then each observation will be either 0 or 1, so even if we have the 'correct' model and the 'correct' estimate $\widehat{p} = \frac{1}{2}$, the 'error sum of squares' will still be $\sum_{j=1}^{n}(X_j - \frac{1}{2})^2 = \frac{n}{4}$ which is substantial. Hence, for the negative binomial model, a large residual sum of squares does not necessarily imply that the model is bad.)

Exercise 5: Colour-Stimuli

In an experiment designed to study the perceptions of colour in human vision (Ekman 1954), 14 colours differing only in their hue (i.e. wavelengths from $434\mu m$ to 674 μm) were projected two at a time onto a screen in an all-pairs design to 31 subjects who rated each of the possible $m = 91$ pairs on a five point scale from 0 (no similarity) to 4 (identical).The ratings for each pair of colours was averaged over all subjects and the results divided by 4 to bring the similarity ratings into an interval $[0, 1]$. These mean ratings were then collected into a 14×14 table which was treated as a correlation matrix. They are found in the file `color-stimuli.rda` on the course page. Carry out a scaling of the data and show that the solution is a 'colour circle' ranging from violet ($434m\mu$) to blue ($472m\mu$) to green ($504m\mu$) to yellow ($584m\mu$) to red ($674m\mu$). Compare with a non-metric scaling solution.

Exercise 6: Wisconsin Breast Cancer Data

1. The Wisconsin breast cancer data is found in the file `wdbc.rda` on the course page. Use a random forest on this data set. Repeat the analysis 100 times using different random seeds to start each replication. For each repetition, find the OOB misclassification rate and draw the boxplot for OOB misclassification rates. Repeat this for different values of m (number of variables selected as candidates for splitting) and B (number of bootstrap trees in the forest). What can you say about the effect of m and B on the OOB misclassification rate?

2. Run a batch-SOM analysis on the Wisconsin Breast Cancer data. How well does the SOM method cluster the tumours into the classes corresponding to the variable 'class'?