Statistical Data Analysis: Take-Home Examination

Submit your answers in p.d.f. format by Friday 9th February, 13:00 to me at `noble@mimuw.edu.pl`

1. The data set `ushighways.txt` in the course data directory consists of the approximate length (in miles) of all 212 U.S. interstate highways (spurs and connectors).

   (a) Try making a histogram. What is your choice of window width and why?

   (b) Compare kernel density estimates for these data using UCV, BCV and SJPI window width estimators. Which method do you recommend for this example?

2. The file `bodyfat2.XLS` contains measurements of the percentage of bodyfat for 252 men. The Y-variable is the bodyfat percentage; there are 13 explanatory variables (X-variables).

   (a) Consider the correlations between the 13 explanatory variables. Are there grounds to suspect ill-conditioning?

   (b) Perform an OLS regression using all 13 explanatory variables. Which variables are significant?

   (c) Perform Forwards Stepwise and Backward Elimination procedures. What are the resulting models?

   (d) Use leave-one-out cross-validation for estimating the mean prediction error as a criterion for model selection. Which subset of variables gives the best model, based on this criterion?

   (e) Apply LASSO to the bodyfat data set. Indicate the LASSO path and decide on a suitable model. Justify your choice.

   (f) Apply LARS to the bodyfat data set and compare your results with the LASSO.

3. Consider the `yarn` data set, which is included in the **pls** package. It is also in the file `PET.txt` in the course data directory. The Y variable is the density (measured in kg/m^3); the 30 explanatory variables are various frequencies.

   (a) Perform a PCR and a PLSR on the `yarn` data. For PCR, how many PCs do you recommend and why? What is the resulting PLSR model? Justify your choice of model in each case.

   (b) Preform a ridge regression for the `yarn` data set, using leave-one-out cross validation to compute the optimal ridge parameter. What is the resulting model?

   (c) Which techniques give the most satisfactory results for this example?

4. 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.

5. 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.

(a) Compute the variance of the 16 variables and show that they are very similar.

(b) Carry out a PCA using the covariance matrix.

(c) How many PCs explain 80% resp. 90% of the total variation in the data?

(d) Display the first three PCs using pairwise scatterplots.

(e) Carry out a PCA using the correlation matrix. Is there any substantial difference?pendigits

(f) Draw the scree plots, for PCA using covariance and for correlation. How many PCs would you use based on this?

(g) Is there ill-conditioning in the data matrix? Base your answer on the PCA.

6. Consider the 'car marks' data set in `carmarks.txt` in the course data directory. The data are averaged marks for 24 cars from a sample of 40 persons. The marks range from 1 (very good) to 6 (very bad). The first two columns contain 'type' and 'model'. The next 8 columns contain the variables: economy, service, non-depreciation of value, price (1 is cheapest), design, sporty car, safety, easy handling. Let the X variables be (price, value stability) and let the Y variables be (economy, service, design, sporty car, safety, easy handling). Perform a canonical correlation analysis on the data and draw suitable conclusions.

7. 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*.

   (a) Apply agglomerative and divisive hierarchical methods for clustering the variables using all 5 indices and the 2 angles available for all items. Construct dendrograms with single-linkage, average-linkage, complete-linkage and Ward-linkage for the methods.

   When an isolated observation appears high enough up in the dendrogram, it becomes a cluster of size one and hence plays the role of an outlier. Which linkage methods give outliers?

(b) 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?

8. Consider the data in the file `primate.scapulae.txt` in the course data directory. Carry out five linear discriminant analyses (one for each primate species), where each analysis is of the form 'one class versus the rest'. Find the spatial zone (known as the *ambiguous region*) that does not correspond to any LDA assignment of a class of primate (out of the five considered).

   Suppose that LDA boundaries are found for the `primate.scapulae` data by carrying out a sequence of $\binom{5}{2} = 10$ LDA problems, each involving a distinct pair of primate species. Find the *ambigous region* that does not correspond to any LDA assignment of a class of primate (out of the five considered). Suppose we classify each primate in the data set by taking a vote based upon these boundaries. Estimate the resulting misclassification rate and compare it with the rate from the multi-class classification procedure.