Multivariate Statistical Analysis: Assignment 1

Submit your answers in p.d.f. or html format by Monday 2nd December 2024, 13:00 to me at `noble@mimuw.edu.pl`

When presenting your answers: imagine that you are carrying out statistical consultancy work for a client who presents you with these data sets. Present your conclusions clearly and motivate them appropriately.

Exercise 1

The file `bodyfat2.xlsx` contains measurements of the percentage of bodyfat for 252 men. The Y-variable is the bodyfat percentage; there are 13 explanatory variables (X-variables). *Do not include 'density' as an explanatory variable; the* 13 *variables to the right of 'bodyfat' are the explanatory variables.* One way to load an `.xlsx` file is by:

```
> www =
"https://www.mimuw.edu.pl/~noble/courses/MultivariateStatistics/data/bodyfat2.xlsx"
> library(rio)
> bodyfat = import(www)
```

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

2. Perform regression analysis on this data, using the principal component approach, partial least squares and the other methods. 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? And which method gives the best results.

3. Now give special attention to the LASSO method applied to the bodyfat data set. Indicate the LASSO path and decide on a suitable model. Justify your choice.

Exercise 2

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.

Exercise 3

Consider the data in the file `primate.scapulae.xslx` in the course data directory; the object is to carry out discriminant function analysis. Do not use the *gamma* variable; use the other 10 variables for classification. Carry out five linear discriminant analyses (one for each primate species), where each analysis is of the form 'one class versus the rest' (i.e. for each of the 5 analyses, you pose the question 'is it in class i or is it in one of the other classes?'). 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 (i.e. where, from your 5 analyses, you do not get a clear answer for which class the observation belongs to).

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.

Exercise 4: Boston Housing Data

Least Angle Regression and LASSO can be carried out using routines in the **lars** package and also the **glmnet** package. The Boston housing data can be found in the file
`boston_corrected.txt` in the course data directory. Information may be found here:

`https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`

There are 506 observations on census tracts in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. For the response variable, use the *logarithm* of `MEDV`, the median value of owner-occupied houses in thousands of dollars. There are 13 input variables (plus information on location of each observation). The 13 explanatory variables are: `CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT`. The url above gives information on these variables. Compute the OLS estimates.

- **Note 1** The first few lines are text and should be removed from the file.
- **Note 2** Not all the instantiations are complete. Remove the incomplete instantiations before performing a regression analysis.

1. Perform regression analysis. Is OLS regression effective? Or do the penalised regression techiques give a better answer? Decide on the best regression technique (from those dealt with in the course) and analyse the data according to this method.
2. Try the Regression Tree approach and compare results, both in terms of accuracy and speed.