Final project STK4030/9030 - Modern Data analysis - Fall 2013

Available Tuesday November 26th.

Handed in: Tuesday Desember 10th. at 14.00

This is the problem set for the project part of the finals in STK4030 fall 2013. The reports shall be individually written. You may discuss the solutions with your fellow students, but the intention is that the final formulations shall be done individually.

Three copies marked with your candidate number shall be placed in Geir Storvik's post box at room B700 at the seventh floor in N. H. Abel's house. Handwritten reports are acceptable. Enclose the parts of the computer outputs which are necessary for answering the questions. The other parts can be collected in appendices. When you refer to material in these, be careful to indicate explicitly where.

Magne Aldrin and Geir Storvik

Problem 1

In this exercise we will discuss classification and some extensions of what we have discussed in the lectures. We will consider the general setting where we have K possible classes, where π_k is the probability for an observation to belong to class k and, within class k, the observation x follow a distribution $f_k(x)$.

As a motivation we will start by looking at a specific example. Assume $K = 2$, $\pi_k = 0.5, f_k = N(\mu_k, \sigma_k)$ where $\mu_1 = 0, \mu_2 = 3$ and $\sigma_1 = 1, \sigma_2 = 0.5$.

(a) Calculate the optimal decision for all values of x in this case. In particular, plot $\Pr(\hat{g} = 1|x)$ and $\Pr(\hat{g} = 2|x)$ as a function of x in the interval $[-4, 8]$. Discuss the results.

We will however extend the prediction space to include a "doubt" category which can be used if there is considerable doubt in the classification. In particular, assume

$$L(g, \hat{g}) = \begin{cases} 0 & \text{if } \hat{g} = g \text{ (correction decision)} \\ 1 & \text{if } \hat{g} \neq g \text{ and } \hat{g} \in \{1, ..., K\} \text{ (wrong decision)} \\ d & \text{if } \hat{g} = \mathcal{D} \text{ (being in doubt)} \end{cases}$$

(b) Discuss this loss function. What are reasonable values for d?

(c) Find the optimal decision in this case.

(*d*) Consider the specific example with $K = 2$ above and assume now $d = 0.2$.

Find the optimal decisions for all possible x.

Make a plot showing the optimal decisions in this case. Discuss the results.

Another extension of classification is to take into account outliers.

(*e*) Assume now that $X = x$ is observed and we want to test whether this observation fits the model. Define this as a hypothesis testing problem and construct a test for testing whether the observed value is an outlier.

(*f*) Consider again the $K = 2$ example above. Specify now the different conclusions you obtain for different values of x. Discuss the results.

Problem 2

In this exercise, you will need three data sets (`pm10.train1, pm10.train2, pm10.test`) about traffic related air pollution. These three data sets consist of 50, 100 and 350 observations, respectively, all with hourly observations of PM_{10} concentration at a road in Oslo with corresponding measurements of the number of cars and meteorological variables. PM_{10} means particulate matter with particles up to 10 micrometers in size. The data sets are all random sub sets of a much larger data set collected in the period from October 2001 to August 2003.

The data sets are available available on the course web-page, and you can read them into R by the following commands:

```
pm10.train1<-read.table(
"http://www.uio.no/studier/emner/matnat/math/STK4030/h13/data/pm10train1.dat",
header=T)

pm10.train2<-read.table(
"http://www.uio.no/studier/emner/matnat/math/STK4030/h13/data/pm10train2.dat",
header=T)

pm10.test<-read.table(
"http://www.uio.no/studier/emner/matnat/math/STK4030/h13/data/pm10test.dat",
header=T)
```

Some information on the data:

- 1 response variable
  - log.PM10: the (natural) logarithm of the PM_{10} concentration
- 7 predictors
  - log.Cars: the (natural) logarithm of the number of cars

- temp.2: temperature 2 m above ground (degree C)
- temp.25min2: temperature difference between 25 m and 2 m above ground (degree C)
- wind.speed: wind speed (m/s)
- wind.direction: wind direction (degrees between 0 and 360)
- hour: time of day (hour)
- day.no: day number (counted from Oct. 1, 2001 - e.g., Oct.1 2001 = 1, Oct. 2 2001 = 2)

First, use the smallest data set `pm10.train1` as a training data set and estimate linear regression models in tasks (a), (b) and (c) below.

(a) Estimate a linear regression model with log.PM10 as response, and with all 7 predictors, by (ordinary) least squares (OLS). You may use the `lm` function in R. Report the estimated coefficients.

(b) Estimate the same linear regression model by ridge regression, where optimal tuning parameter are chosen by 10-fold cross validation. Consider the following 17 candidate values for the tuning parameter λ:
{1e5,1e4,1e3,500,100,50,10,5,1,0.5,0.1,0.05,0.01,0.005,1e-3,1e-4,1e-5}.
You may use the R function `lm.ridge` from the MASS library and the `cv.k` function and other parts of the computer code from exercises in the course.

Plot the cross-validated root mean squared error against the logarithm with base 10 of the λ values (you can use the R function `log10`). What is the optimal value of λ?

Report the estimated coefficients.

(c) Estimate the same linear regression model by lasso, where optimal tuning parameter are chosen by 10-fold cross validation. You may use the `lars` function from the `lars` library in R. Consider the following candidate values for the standardized tuning parameter s:
{0,0.05,0.10,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50,0.55,0.60,0.65,0.70,0.75,0.80,0.85,0.90,0.95,1}.

Plot the cross-validated root mean squared error against s. What is the optimal value of s?

Report the estimated coefficients.

(d) Explain why it is reasonable to standardize the predictors to have the same standard deviation in an internal step of the computations when you perform ridge regression or lasso.

(e) Give a few comments on the difference between the estimated coefficients from OLS, ridge regression and lasso, in light of what you general can expect when using these methods.

(f) Use the three estimated models to predict the response values from the predictors in the test set `pm10.test`. Calculate the root mean squared errors (RMSE) for each method and comment on the results.

(g) Now, read in data set `pm10.train2` and put it together with data set `pm10.train1` into a combined training data set with 150 observations. Re-estimate the linear regression models by OLS, ridge regression and lasso. Select tuning parameters by 10-fold cross validation among the same candidate values as before. Report the estimated coefficients and the optimal tuning parameters. Comment on the differences from what you got when you had a training data set with 50 observations.

(h) Estimate a (generalized) additive model on the same data set, with the same response and the same predictors. You can use the `gam` function from the `mgcv` package in R. Plot the estimated curves and interpret the effect of each predictor.

(i) Finally, use the four estimated models from (g) and (h) to predict the response values in the test set `pm10.test`. Calculate the root mean squared errors (RMSE) for each method and comment on the results.

Problem 3

In this exercise you are supposed to analyze a data set for a large industry supplier, HATCO. In total 11 input variables are measured, attributes identified in past studies as the most influential:

X_1	Delivery speed	Numerical
X_2	Price level	Numerical
X_3	Price flexibility	Numerical
X_4	Manufacturer's image	Numerical
X_5	Overall service	Numerical
X_6	Saleforce's image	Numerical
X_7	Product quality	Numerical
X_8	Size of firm	Categorical, 2 groups
X_9	Specification buying	Categorical, 2 groups
X_{10}	Structure of procurement	Categorical, 2 groups
X_{11}	Type of industry	Categorical, 2 groups

These data were obtained by purchasing managers of firms buying from HATCO, rating HATCO on each attribute.

As output we will use evaluations of the respondent's satisfaction with HATCO, measured at different levels. These are

Y_1	Satisfaction level	Numerical
Y_2	Type of buying situation	Categorical, 3 groups

The aim of the exercise is to built up models for predicting the output variables based on the input variables. This is to be done separately on each of the two output variables. You are to use what you have learned in the course in order to do so.

You shall write a report containing

- A description of the methods that you have used.
- The results you have obtained.
- A short summary of your most important findings (not more than 250 words).
- The code that you have used.

The data set is available from the course webpage, called `HATCO.txt` where the columns are $x_1, ..., x_{11}, y_1, y_2$ and can be read into R with the command

```
x = read.table("HATCO_SET.TXT",header=T,sep=";",
   colClasses=c(rep("numeric",7),rep("factor",4),"numeric","factor"))
```

Good luck!