Final project STK4030-f12 - Modern Data Analysis

This is the problem set for the project part of the finals in STK4030-f12. 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.

The deadline for turning in the reports is

Monday December 3'rd at 5 pm.

Three copies marked with your candidate number shall be placed in Anders Rygh Swensen'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 Anders Rygh Swensen

Problem 1

Consider the situation where the training set consists of N responses/targets and the inputs/covariates can be arranged in a $N \times (p+1)$, matrix, $\mathbf{X}$, where the first column of $\mathbf{X}$ consists of 1's. The targets or responses are collected in the vector $\mathbf{y}$. Assume that all the responses have the same variance, σ_ε^2.

a) In the situation where $p = 1$ explain why adding a constant c to all the targets implies that all the predicted/ fitted values are similarly shifted when ordinary least squares (OLS) is used.

b) Will the result in part a) hold for general p? Explain why.

c) Show that this invariance will not hold if the predicted values are $\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ with $\lambda > 0$. Remark that it will be sufficient to write out the details for the case $p = 0$.

The ridge regression coefficients are defined as a solution of the following minimization problem:

$$\widehat{\beta}^{ridge} = \text{argmin}_\beta\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\} \qquad (1)$$

where λ is the penalty. Note that the coefficient β_0 is not part of the sum of penalties.

d) Show that the minimization problem can alternatively be written as

$$\widehat{\beta}^c = \text{argmin}_{\beta^c}\{\sum_{i=1}^{N}[y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c]^2 + \lambda\sum_{j=1}^{p}(\beta_j^c)^2\}, \quad (2)$$

where $\bar{x}_j = \sum_{i=1}^{N} x_{ij}/N$ and show how $\widehat{\beta}^c$ can be expressed by $\widehat{\beta}^{ridge}$, i.e. give the correspondence between β^c and the original β in (1).

e) Explain how the result in part d) can be used to separate ridge regression into two parts: first the coefficient β_0^c is estimated as $\bar{y} = \sum_{i=1}^{N} y_i/N$, then the remaining coefficients are estimated by ridge regression without intercept from the centered values of the inputs, $x_{ij} - \bar{x}_j$

We will therefore in the rest of the problem assume that the inputs have been centered, and denote the $N \times p$ matrix consisting of these values as $\mathbf{X}_c$. Then the ridge regression estimates can be expressed as $\widehat{\beta}^{ridge} = (\mathbf{X}_c^T\mathbf{X}_c + \lambda\mathbf{I})^{-1}\mathbf{X}_c^T\mathbf{y}$.

f) In section 7.6 in the textbook *The Elements of Statistical Learning* (ESL) the effective degrees of freedom is defined as $df(\widehat{\mathbf{y}}) = \sum_{i=1}^{N} \text{Cov}(\widehat{y}_i, y_i)/\sigma_\varepsilon^2$. Show that this is the same as the expression in formula (3.50) in section 3.4.1, i.e $\sum_{j=1}^{p} d_j^2/(d_j^2 + \lambda)$ where $d_1, \ldots, d_p$ are the singular values of $\mathbf{X}_c$.

Problem 2

This problem is an analysis of the data set `no2` which is available on the course web-page. You can read it into R by the following command:

```
no2<-read.table("http://www.uio.no/studier/emner/matnat/math/STK4030/h12/
undervisningsmateriale/no2.data",header=T,row.names=NULL)
```

This data set consists of 50 hourly observations (this is a subset of a much larger data set collected in the period from October 2001 to August 2003) of NO_2 concentration at a road in Oslo with corresponding measurements of the number of cars and meteorological variables.

Some information on the data:

- 1 response variable
  - logNO2: the (natural) logarithm of the NO_2 concentration
- 7 predictors
  - logCars: the (natural) logarithm of the number of cars

- temp: temperature 2 m above ground (degree C)
- tempDiff: temperature difference between 25 m and 2 m above ground (degree C)
- windSpeed: wind speed (m/s)
- windDir: wind direction (degrees between 0 and 360)
- hour: time of day (hour)
- dayNo: day number (counted from Oct. 1, 2001 - e.g., Oct.1 2001 = 1, Oct. 2 2001 = 2)

a) Estimate a linear regression model with logNO2 as response, and with all 7 predictors, by (ordinary) least squares. You may use the `lm` function in R. Report the estimated coefficients. Report also the estimated variance of the noise (= variance of error term).

In the remaining part of this exercise, you should first estimate the same model by ridge regression with optimal tuning parameter found by cross-validation as detailed in Step 1) and Step 2) and answer the questions b)-e). Then you should find the optimal tuning parameter using AIC and answer question f).

Step 1) First, standardize each predictor to have standard deviation 1 by dividing each predictor by its sample standard deviation, i.e. x_{ij}/s_j, where $s_j = \sqrt{1/n \sum_{i=1}^{i=n}(x_{ij} - \bar{x}_j)^2}$ and n is the number of observations. Then, for a given λ, find the ridge regression coefficient. Then transform the estimated regression coefficients back to original scale. You may use the R function `lm.ridge` from the MASS library for this.

Step 2) Consider the following 17 candidate values for λ: $\{10^5, 10^4, 10^3, 500, 100, 50, 10, 5, 1, 0.5, 0.1, 0.05, 0.01, 0.005, 10^{-3}, 10^{-4}, 10^{-5}\}$. Perform 10-fold cross-validation to find the optimal value of λ among these candidate values. You may use the `cv.k` function and other parts of the computer code from Extra exercise 3.4, see the course web page `http://www.uio.no/studier/emner/matnat/math/STK4030/h12/`.

b) Explain why it is reasonable to standardize the predictors to have the same standard deviation.

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

d) Report the estimated regression coefficients for the optimal value of λ. Compare them with the (ordinary) least squares estimates from task a). Is the result as what you could expect?

e) Compute the effective degrees of freedom (= effective number of parameters) for each value of λ. Here, you can ignore the the contributions from from the intercept and the error variance. Plot the effective degrees of freedom against the logarithm with base 10 of the λ values.

f) Assume that the noise is Gaussian with constant variance. Compute the value of AIC (Akaike's Information Criterion) for each value of λ, where the error variance is assumed known and set equal to the estimate variance of the noise from task a). Plot the AIC values against the logarithm with base 10 of the λ values. What are the optimal λ value according to this model selection criterion?

Problem 3

This is essentially Exercise 6.10 on page 217 in the textbook ESL. Consider N samples generated from the model $Y = f(x) + \varepsilon$, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. The inputs $x_1, \ldots, x_N$ are considered fixed, and $\varepsilon_1, \ldots, \varepsilon_N$ are independent. The regression function $f(x)$ is estimated using a linear smoother $\mathbf{S}_\lambda$ with smoothing parameter λ, so the fitted values are $\widehat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$ where $\mathbf{y} = (y_1 \ldots, y_N)'$ contains the observed responses. Consider the in-sample prediction error $E_{Y^0}[\frac{1}{N}\sum_{i=1}^{N}(Y_i^0 - \widehat{f}_\lambda(x_i))^2]$, defined formula (7.18) in ESL and specified to squared error loss.

a) Find the expectation of the training error $\overline{err}$ defined in formula (7.17) in ESL when the loss function is squared error.

b) Show that for squared error loss will the random variable C_λ defined as $\overline{err} + \frac{2\sigma^2}{N} trace(\mathbf{S}_\lambda)$ have the same expectation as the in-sample prediction error.

Problem 4

This is an elaboration on exercise 5.5 in the textbook ESL analyzing the `phoneme` data, which can be found on the textbook web-page and read in by the following R-command

```
phoneme <- read.table("http://www-stat.stanford.edu/~tibs/ElemStatLearn/
datasets/phoneme.data",header=T,sep=",")[-1]
```

The data consist of 4509 pronunciations of the phonemes "sh" "iy" "dcl" "aa" or "ao", together with log periodograms at 256 frequencies. Here the phoneme type is the response, which therefore has five categories, and the 256 log-periodograms are inputs or covariates. These features are correlated, so filtering them using splines is a possibility.

Draw a random sample consisting of 3000 observations that will be used as a training set, and the remaining 1509 as a test set.

a) Plot the natural cubic splines in the interval $[0, 1]$ in the basis using the truncated power representation introduced in section 5.2.1 in ESL when there are 4 knots at $0.2, 0.4, 0.6$ and 0.8.

b) Explain how the natural cubic splines can be used to filter the inputs.

c) Consider the following three choices of number, M, and localization of the knots $\xi_1, \ldots, \xi_M$

   - $M = 3$, knots at the $0.25, 0.50, 0.75$ percentiles of the frequencies, i.e. at the 64'th, 128'th and 192'th frequency.
   - $M = 3$, knots at the $0.1, 0.50, 0.9$ percentiles of the frequencies,
   - $M = 4$, knots at the $0.2, 0.4, 0.6, 0.8$ percentiles of the frequencies.

   Use these three alternatives for filtering and classify the data in the training set using quadratic discriminant analysis. Report the test error rates.

d) Explain how five-fold cross-validation based on the data in the training set can be used to choose among the possible alternatives. Report the estimates of the prediction errors in this case.

e) Discuss the results from part c) and d). What is your conclusion?

[R hints: For implementing the quadratic discriminant analysis you can use the procedure `qda` from the MASS library.]