MATH 3333 3.0 - Winter 2019-20
Assignment 2

(Due Date: February 11, 2020)

Question 1: Please download the bike sharing data set from the following website: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. There are two data sets and we will use the day.csv. Import the data into R and perform the following exploratory analysis.
a) Use the LASSO package to analyze the data with a collection of predictors and provide the summary of the output of the command "lars".

b) Based on the output of "lars", please provide the sequence of candidate models. For example, the first model is $\{X_5\}$, the second model is $\{X_5, X_3\}$ and the third model is $\{X_5, X_3, X_{10}\}$, etc.

c) Use the cross validation method, select the best value for the fraction s based on the plot of cross validation error againt the fraction s. The fraction s measures the ratio of the L_1 norm of the penalized estimate over the L_1 norm of the regular penalized estimate.

d) Use the optimum s you select, perform the penalized regression and output the optimum model and the estimated coefficients.

Question 2: In the following marketing set, we have 9 years with the sales in 10 million euro and the advertising expenditure in million euro.

Year	Sales	Advertisement
1	65	23
2	76	26
3	85	30
4	106	34
5	119	43
6	129	48
7	142	52
8	144	57
9	151	58

a) In assignment one, we have computed the least square regresson for this data set and we have the estimated linear model. Now two additional data points arrived. They are Year 10, Sales 154, and Advertisement 61; Year 11, Sales 157, and Advertisement 63. Please use the online algorithm to update the linear model twice. First use the new observation of Year 10 to perform the sequential learning and update the model. Then use the new observation of Year 11 and update the model again. Provide all of the work of your calculation. Please perform the computation by hand with the help of a calculator.

b) Based on the original 9 observation data and perform a ridge regression. Program it with R. Output the ridge regression results at a few different values of λ.

c) In your ridge regression, when λ increaes, what do you observe from the values of the estimated coefficients. Does any of the estimated coefficients shrink to zero like the L_1 LASSO regression?

Question 3:

a. In this question, we will investigate the problem of mulitple testing. Consider the hypothesis testing of $H_0 : \mu = 0$, vs $H_a : \mu \neq 0$. Under the null hypothesis, the Z test statistic is a standard normal random variable. We reject the null hypothesis when $|Z|$ is greater than 1.96 at the significance level of 0.05. Write a R program to simulate 1000 Z test statistic from standard normal $N(0, 1)$. (The command to use is ztests=rnorm(1000,0,1).)

b. If we perform hypothesis testing using the significance level of 0.05. Among the 1000 test statistics you generated, how many of them are rejected?

c. As the Zs are generated from the null hypothesis, we consider these rejections are all false positive discoveries. Please use a short paragraph to summarize the problem we are facing when we perform multiple testings.

Question 4: Consider a data set with the response vector $Y = (y_1, \ldots, y_n)^T$ and the data matrix

$$\begin{pmatrix} 1 & x_{11} & x_{12} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

We model the relationship between X and Y using the linear regression model: $y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon \sim N(0, \sigma^2)$. Let the parameter vector be denoted as $\theta = (\theta_0, \theta_1, \theta_2)^T$. We wish to minimize the sum of weighted squared residuals: $SSE = \sum_{i=1}^{n} w_i(y_i - (\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}))^2$. Derive the formula for the solution of θ which minimizes the weighted sum of squared errors.

Question 5: Analyze the German data set from the site:
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

a) Perform the logistic regression on the dataset. You can some of the predictors in the model. Please use 900 observations as the training set and use your model to predict the default status of the remaining 100 loans. What is the cutoff value of the probability do you use for your analysis? How many default ones are predicted to be non-default ones (number of false negative)? How many non-default ones are predicted to be default ones (number of false positive). Then you need to improve your model by adding more predictors or adding some higher order terms or interaction terms. Please demonstrate that your new model has lesser errors than your first model in the 100 testing cases.

b) Please investigate how the false negative number and false positive number change with respect to the different cutoff value of probability.

Question 6:

In logistic regression, we assume $Y = (Y_1, \ldots, Y_n)^T$ are a collection of n binary observations. For each Y_i, we observe $X_i = (X_{i1}, \ldots, X_{ip})^T$ predictors. We assume

$$\log \frac{p_i}{1 - pi} = X_i^T \theta,$$

where $\theta = (\theta_1, \ldots, \theta_p)^T$ is the vector of regression coefficients.

a) Formulate the overall loglikelihood of the dataset $l(Y)$.

b) Derive the first derivative $\partial l(Y) / \partial \theta_2$.