Homework 3

Big Data Analytics

Due time: December 13, 2018, 11:59 PM

Instructions:
1. Please turn in your homework in a pdf format generated by R Markdown.
2. Discussion are encouraged, but you need to work on the problems by yourself.

1 (50 points) Problem 1 (Problem 8, Chapter 5)

We will now perform cross-validation on a simulated data set.

(a) Generate a simulated data set as follows:

```
> set.seed(1)
> x=rnorm(100)
> y=x-2*x^2+rnorm(100)
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

(b) Create a scatterplot of X against Y. Comment on what you find.

(c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:

(i) $Y = \beta_0 + \beta_1 X + \epsilon$

(ii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

(iii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

(iv) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

Note you may find it helpful to use the *data.frame()* function to create a single data set containing both X and Y.

(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

(e) Which of the models in (c) had the smallest LOOCV error? IS this what you expected? Explain your answer.

(f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation?

2 (50 points) Problem 2 (Problem 8, Chapter 6)

In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) Use the *rnorm()* function to generate a predictor X of length $n = 100$, as well as a noise vector ϵ of length $n = 100$.

(b) Generate a response vector Y of length $n = 100$ according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

where $\beta_0, \beta_1, \beta_2$, and β_3 are constant of your choice.

(c) Use the *regsubsets()* function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \cdots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2? Shoe some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the *data.frame()* function to create a single data set containing both X and Y.

(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?

(e) Now fit a lasso model to the simulated data, again using the predictors $X, X^2, \cdots, X^{10}$. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

(f) Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \epsilon,$$

and perform best subset selection and the lasso. Discuss the results obtained.