EPPS6323 Knowledge Mining

Assignment 8

1. Run Lab_LDA01.R in R
2. Review ISLR Chapters 6 and look up answers for the following questions
3. From the three methods (best subset, forward stepwise, and backward stepwise):
   a. Which of the three models with k predictors has the smallest training RSS?
   b. Which of the three models with k predictors has the smallest test RSS?
4. Application exercise:

Generate simulated data, and then use this data to perform best subset selection.

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

   Hint:

   ```
   set.seed(1)
   X = rnorm(100)
   eps = rnorm(100)
   ```

2. 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 4, 9, 2, 1 respectively.

   Plot x and y.

3. Use the `leaps` package:

   ```
   require(leaps)
   ```

4. Use the regsubsets() function from the leaps package to perform best subset selection in order to choose the best model containing the predictors. $x, x^2 \dots x^{10}$.

   Hint:

   ```
   regsubsets(Y~poly(X,10,raw=T), data=data.frame(Y,X), nvmax=10)
   ```

   What is the best model obtained according to Cp, BIC, and adjusted R^2? Show 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 .

5. Repeat 3, using forward stepwise selection and using backwards stepwise selection. How does your answer compare to the results in 3?