chiehjul
50341634

Statistical Data Mining I
Homework 2

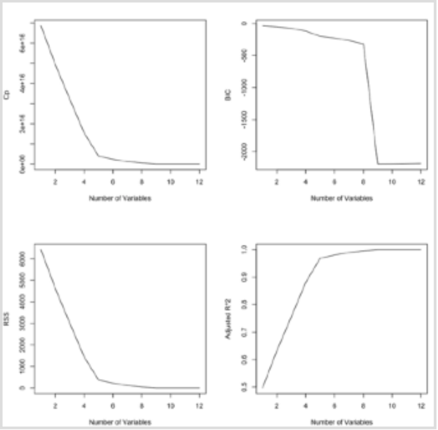
1) (10 points) Consider the cereal dataset in UBlearns. Suppose that you are getting this data in order to build a predictive model for nutritional rating.

a) Divide the data into test and training. Fit a linear model and report the MSE.
Ans:
According to the professor's suggestion, I randomly separated the cereal dataset with 20% of the test dataset, and 80% of the training dataset. Also, I dropped off the variables "name", "mfr" and "type" due to non-numeric value, hoping that it would present the better results of the linear regression model.

After processing training and testing data, I used the "rating" be the predicted data to fit the linear regression model, then I got the MSE of training data is 7.441635e-14 and of testing data is 1.154347e-13.

I also tried to calculate the error rate of training/test data, but I found that the predictive values are almost the same as the expected when I set the predictive values with the round digit to 5. Hence, it is still better to use MSE to calculate the details error.

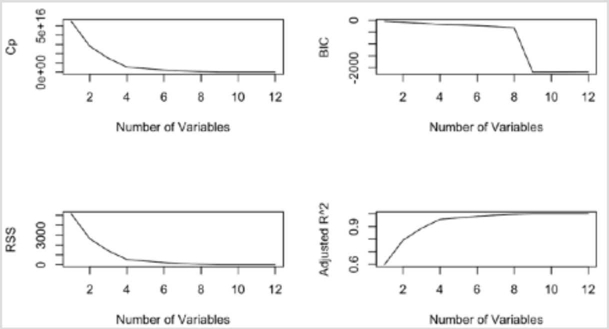
b) With the data in (a) perform **<u>either</u>** forward or backwards subset selection.
Ans:
I choose backward subset selection to apply my dataset, because forward subset selection may not be optimal, as claimed by the lecture.



According to the BIC and RSS plot with the subset size, I speculated that the best size of the subset is might between 5 and 9, but the best model in CP is 10 variables with {calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, weight}, and in BIC is 9 variable with {calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins}.

In this case, I considered that we need to perform exhaustive subset selection to determine the most suitable model.

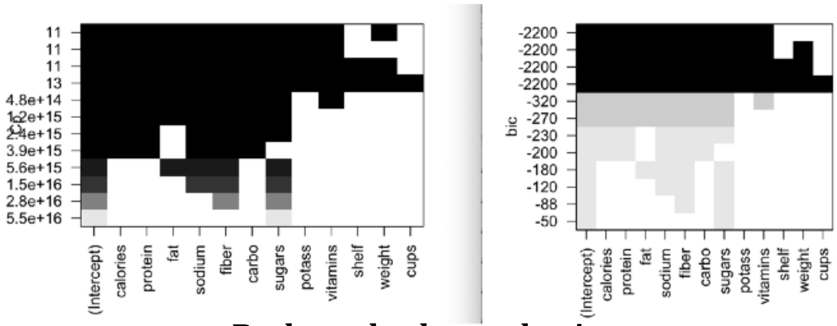
c) With the data in (a) perform exhaustive subset selection.
Ans:
For the result of exhaustive subset selection, there is no obvious difference with the result of backwards subset selection. Also, the best size of subsets in CP and BIC, be the same as backward subset selection, is 10 and 9. In my view, the best



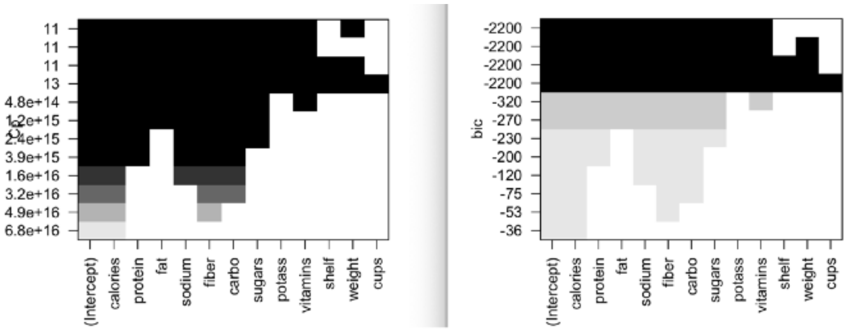
fit of the subset size is 4 for RSS in exhaustive subset selection, it is not enough improvement in performance to justify a more complex model.

d) Draw some conclusions through comparisons between models (a-c). Reflect on the comparative predictive accuracy, and model interpretation. Which model would you say is the "best one" based on your results?

Ans:



Backward subset selection



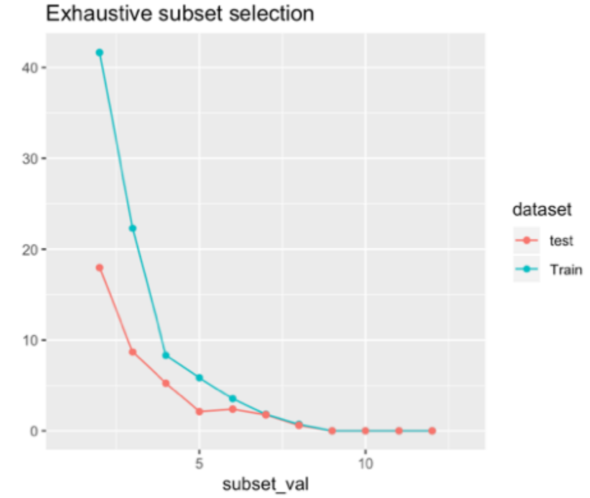
Exhaustive subset selection

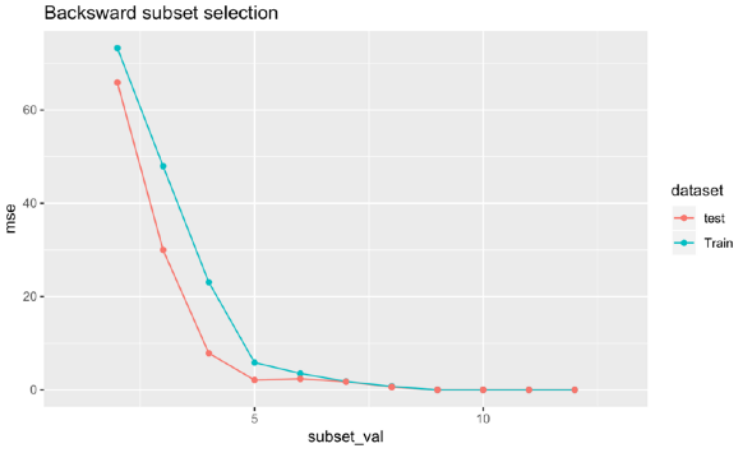
For models b and c, it shows the same selection of variables in both CP and BIC, apart from the subsets with the fewer variable. When the number of variables in the subset is four, the selected variables of the two models will be different, so the best subset is when the variable is five, which is {calories, protein, sodium, fiber, carbo}.

After selecting appropriate subset, I put the the subset into linear model and predict again, it presented quite higher MSE, which is 5.844245 in train data and 2.101434 in test data.

Thus, instead of presenting one subset, I plot the relationship with different subsets selection and their MSE with training/test data. And it is clear to find that if the number of variables in the subset is less than 9, the MSE will become worse.

As a result, I believe that this dataset is not





suitable for finding the best group by using the lab method, it may due to the fewer data in this case. Yet overall I think the exhaustive subset selection can present the best performance with 9 variables in a subset, this data set is not very large. Although the difference between MSE from 9 to 12(all) is not huge, there are still relatively few variables.

2) (10 points) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the *zipcode* data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k. The *zipcode* data is available in the ElemStatLearn package – or the website for the text. Note that you do not have to divide the data into test and training because it is done for you.

Ans:

First of all, I selected 2's and 3's zipcode training/test data and utilized them to the linear regression model. And I found that the prediction result would present not only 2 and 3 but also other numbers.

I separated all prediction into 2 and 3 with the criteria means, because we only have two outputs, then got the training/test error rate which is 0.005039597 and 0.03846154.

And for the KNN classification:

(Training/test error rate)

a. k = 1,
0 /0.02472527

b. k = 3,
0.005039597/0.03021978

c. k = 5
,0.005759539/0.03021978

d. k = 7
,0.006479482/0.03296703

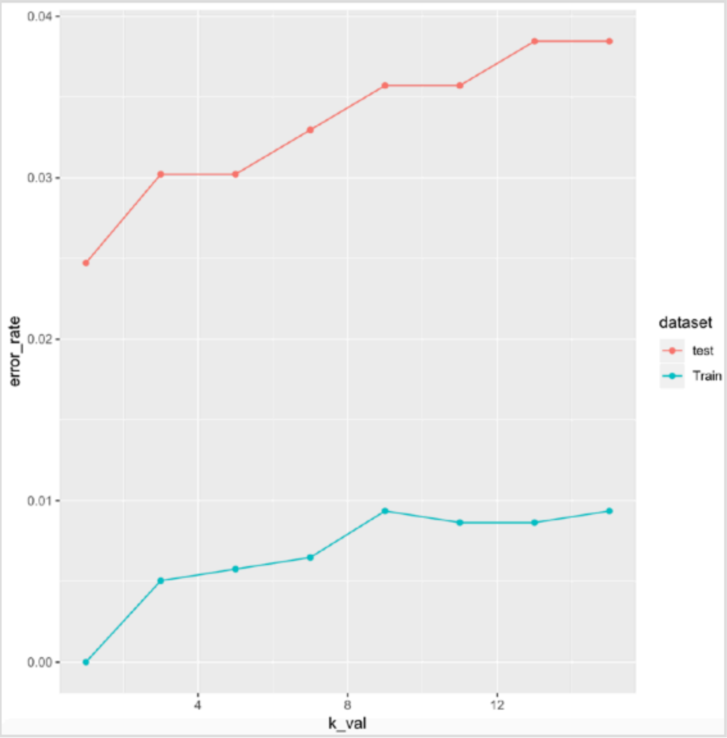
e. k = 9
,0.009359251/0.03571429

f. k = 11 ,
0.008639309/0.03571429

g. k = 13 ,
0.008639309/0.03846154

h. k = 15 ,
0.009359251/0.03846154

i. k = 9
,0.009359251/0.03571429



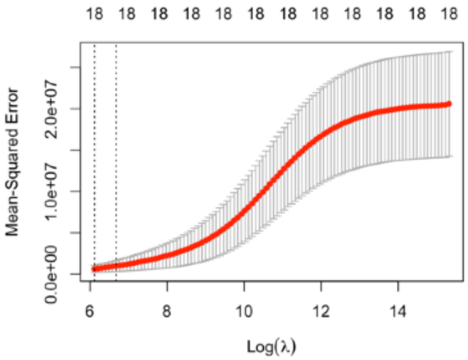
Hence, for zipcode dataset, k-nearest neighbor classification with k = 1 and 3 have higher performance than linear regression model.

3) (10 points) In this exercise, we will predict the number of applications received using the other variables in the College data set in the ISLR package.
*** be sure to look closely at this data, you may want to consider the multi-scale nature of the problem, and perhaps use a transformation on some of the variables.***

(a) Split the data set into a training set and a test set. Fit a linear model using least squares on the training set, and report the test error obtained.
Ans:
I split the college dataset into 30% of the test dataset, and 70% of the training dataset randomly after omitting data with NULL value. Then, I set the variable "Apps" to the output result for fitting the linear regression model. The test error rate (# of error result/ total # of test data) is 1 and the MSE of test data is 1194330.

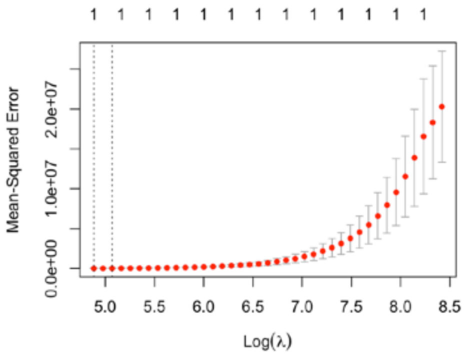
I also tried to transfer the variable "Private" into 1(Yes) and 0(No), yet the MSE of test data became 1261423, which is higher than the original data, so I took the modified data off.

(b) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.
Ans:
Before fitting a ridge regression model, I transformed the type of training dataset from dataframe to matrix, and also changed the type of each data from string to numeric, which means the value in variable "Private" will become 1 or 2.
According to the cross-validation, the best λ (minimum) is 453.2772, and the MSE of test data change from 6912862 (without λ) to 195620.3, which is quite lower than the test error with the linear regression model.



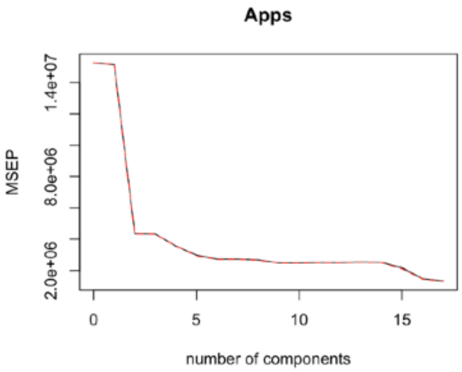
(c) Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.
Ans:
The MSE of test data is 2110629 if I didn't set λ, then I put the best lamda 132.1327 from cross validation and the MSE became 16047.41. The performance is more better than in linear and ridge regression model.

For the non-zero coefficient, my data present most zero coefficient.



(d) Fit a PCR model on the training set. Report the test error obtained, along with justification for the choice of “k”.

Ans:

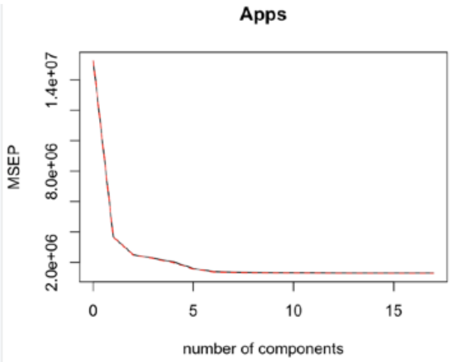


For PCR model, I used the former dataframe type to utilized the model, and in the Validation Plot, there is a huge gap within 1 to 5, and the best performance number of component above 15; however, the change of improvement is much lower than the difference between 0 and 5. So I tried all the number of components in the gap (0-5) only and got the lowest MSE 1631099 with 5 components.

(e) Fit a PLS model on the training set. Report the test error obtained, along with justification for the choice of “k”.

Ans:

For PLS, it is clear that 5 or 6 is the best number of components, because the line with number of component greater than 7 tends to be stable. In this case, the 5 components will present the lowest MSE 1142412.



(f) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

Ans: