EAS506LEC000 : STATISTICAL DATA MINING

HW ASSIGNMENT – 4

SUBMITTED BY : DITHYA SRIDHARAN

CLASS NUMBER : 48

QUESTION 1

PROBLEM STATEMENT

For the prostate data of Chapter 3, carry out a best-subset linear regression analysis, as in Table 3.3 (third column from the left). Compute the AIC, BIC, five-and tenfold cross-validation, and bootstrap .632 estimates of prediction error.

1. DATA SUMMARY:

These data come from a study that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. It is data frame with 97 rows and 9 columns.

2. MALLOWS CP AND BIC PREDICTION ERROR USING BEST SUBSET:

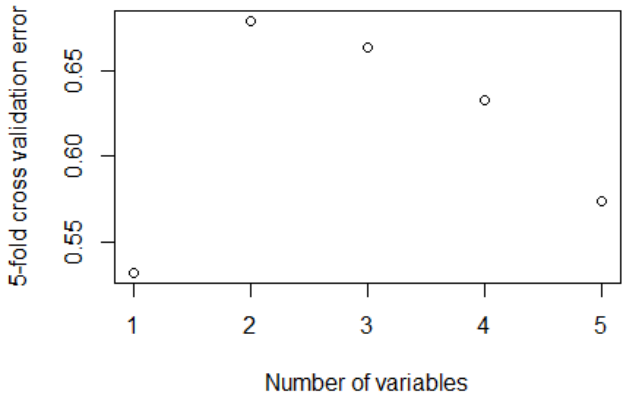
The regsubsets() function (part of the leaps library) performs selection. Here we use the regsubsets function but specify the `method="exhaustive" option. The summary() command outputs the best set of variables for each model size. Error values obtained for Cp is : 4.886363 2.723052 1.481855 1.562523 3.006168 4.701049 6.369559 8.183109. 3rd error is the least. Therefore, selection based on Cp gives 3 variables. Error values obtained for BIC is : 4.937344 5.051128 5.932996 8.150890 11.867325 15.864883 19.830740 23.961148. 1st gives least error. Therefore, selection based on BIC gives 1 variable.

3.CHECKING IF MODEL OBTAINED IS A GOOD FIT

The model obtained using prediction errors of Mallows Cp and BIC were fit. Test MSE for the two models and gave test MSE equaling 0.5048945 and 0.504 respectively. Therefore, good fit.

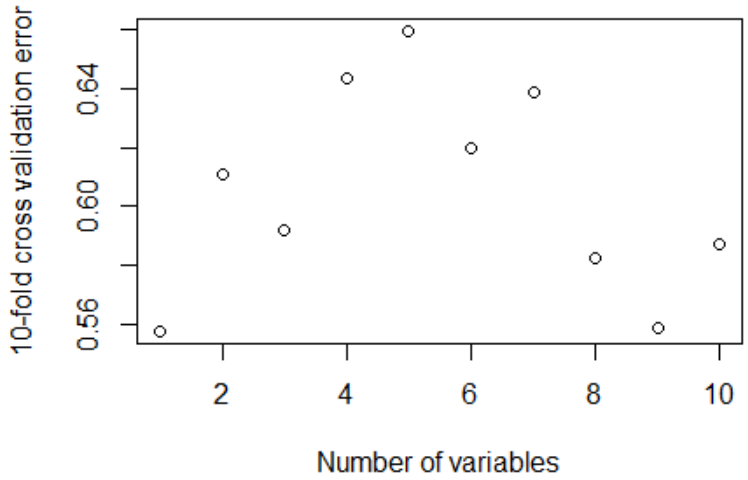
4. FIVE-FOLD CROSS VALIDATION

A model on the training set was fit. We have considered for k=5. It is found that the cross validation errors are as follows : 0.5320913 0.6787811 0.6630620 0.6327106 0.5737362. It can be seen that, 1st gives least error. Therefore, selection based on 5-fold cross validation gives 1 variable as best fit.



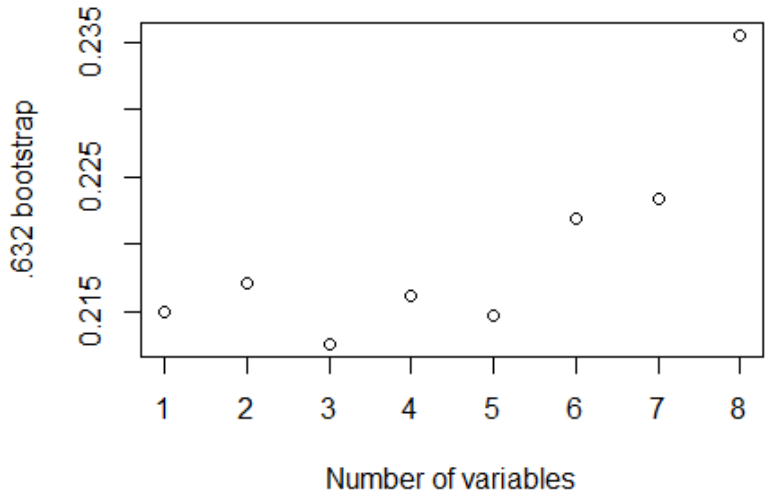
4. TEN-FOLD CROSS VALIDATION

A model on the training set was fit. We have considered for k=10. It is found that the cross v alidation errors are as follows : 0.5577201 0.6110327 0.5920459 0.6435376 0.6595717 0.620 0632 0.6387941 0.5826988 0.5590168 0.5872642.It can be seen that, 1st gives least error. Th erefore, selection based on 10-fold cross validation gives 1 variable as best fit.



4. .632 BOOTSTRAP ESTIMATE

A model on the training set was fit. We have considered for .632 bootstrap estimate. It is found that the errors are as follows : `0.2149120 0.2170678 0.2126162 0.2161217 0.2146683 0.2218999 0.2233176 0.2354857.` It can be seen that, 3rd gives least error. Therefore, selection based on .632 bootstrap gives 3 variables as best fit.



QUESTION 2

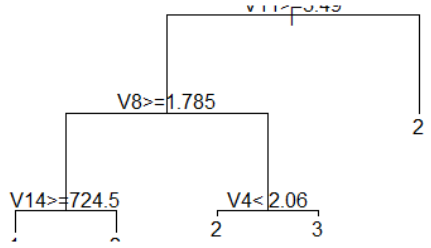
PROBLEM STATEMENT

Access the wine data from the UCI machine learning repository(https://archive.ics.uci.edu/ml/datasets/wine). These data are the results of a chemical analysis of 178 wines grown over the decade 1970-1979 in the same region of Italy, but derived from three different cultivars (Barolo, Grignolino, Barbera). The Babera wines were predominately from a period thatwas much later than that of the Barolo and Grignolino wines. The analysis determined the quantities MalicAcid, Ash, AlcAsh, Mg, Phenols, Proa, Color, Hue, OD, and Proline. There are 50 Barolo wines, 71 Grignolino wines, and 48 Barbera wines. Construct the appropriate-size classification tree for this dataset. How many training and testing samples fall into each node? Describe the resulting tree and your approach.

1. DATA SUMMARY

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The data set has 13 variables and 178 instances.

2. GROWING A TREE MODEL AND CHECKING MISCLASSIFICATION ERROR



A decision tree for the wine set data was plotted. A model was fit and the prediction error obt -ained. The misclassification error for the model was 0.0833. The model was found to be 91. 667% accurate. Thus, it is a good fit.

2. PRUNING AND CHECKING MISCLASSIFICATION ERROR

The tree was pruned and misclassification error was obtained. It was found that the misclassi-fication error was found to be the same for tree model and pruned model.

Misclassification error was 0.0833 making the model 91.667% accurate(good fit).

3. NODE INFORMATION

Summary(wine_model) produces a detailed summary of the tree and it's nodes giving infor-mation on the splis, no of observations, predicted class etc.

Node	Total no of training/test observations	Predicted Class
1	142	2
2	94	1
3	48	2
4	55	1
5	39	3
8	47	1
9	8	2
10	2	2
11	37	3

4. CONCLUSION

- The decision tree approach gives a good model with 91% accuracy for the given data set
- We get similar results when you try for other parameters as well(in this case, V1).
- This suggests that this model is superior to other models for this data set.

QUESTION 3

1.PROBLEM STATEMENT

Apply bagging, boosting, and random forests to a data set of your
choice (not one used in the committee machines labs). Fit the models on a training set, and evaluate them on a test set. How accurate are these results compared to more simplistic (non-ensemble) methods (e.g., logistic regression, kNN, etc)? What are some advantages (and disadvantages) do committee machines have related to the data set that you selected?

2.DATA SUMMARY

Statistics for a large number of US Colleges from the 1995 issue of US News and World Report. A data frame with 777 observations and 18 variables.

3.SIMPLISTIC METHODS :

a) LINEAR REGRESSION

Linear regression and least squares model was applied on the data set and the Test MSE and Adjusted R2 was obtained to test model accuracy. The model was fitted with respect to Apps(no of applications received) variable of the data set. The adjust R2 was 0.8926812 making the model 89% accurate.

b)LOGISTIC REGRESSION

Logistic regression model was applied on the data set and the Test MSE and Adjusted R2 was obtained to test model accuracy. The model was fitted with respect to Apps(no of applications received) variable of the data set. The adjust R2 was 0.8717949 making the model 87% accurate.The test MSE observed was 0.1417323.

4.RANDOM FORESTS :

Random Forest model was applied on the data set and the misclassification error and Adjusted R2 was obtained to test model accuracy. The model was fitted with respect to Apps(no of applications received) variable of the data set. The adjust R2 was 0.9135211 making the model 91% accurate. The misclassification error observed was 0.5.

5.BAGGING :

Bagging was applied on the data set and the misclassification error and Adjusted R2 was obtained to test model accuracy. The model was fitted with respect to Apps(no of applications received) variable of the data set. The adjust R2 was 0.902788 making the model 90% accurate. The misclassification error observed was 0.7.

5.BOOSTING :

Boosting was applied on the data set and the misclassification error and Adjusted R2 was obtained to test model accuracy. The model was fitted with respect to Apps(no of applications received) variable of the data set. The adjust R2 was 0.8944636 making the model 89% accurate. The misclassification error observed was 0.8935897.

6.CONCLUSION

- Random forest gave the lowest classification error and a model accuracy of 91% making it the best fit model among above obtained for this dataset.

7.ADVANTAGE AND DISADVANTAGE OF COMMITTEE MACHINES

- **Disadvantage** : Simplistic models are computationally cheaper and faster than committee machines for this data set.
- **Advantage** : Can be applied recursively(eg boosting) and error could be made arbitrarily small for this data set.