IFI 8420

Assignment 4

Group Submission

LASSO, Tree Regression, and Cross Validation

Part 1 (100 points)

You will work with College data for the assignment.

Description of College data set:

Statistics for a large number of US Colleges from the 1995 issue of US News and World Report.

A data frame with 777 observations on the following 18 variables.

- Private A factor with levels No and Yes indicating private or public university
- Apps Number of applications received
- Accept Number of applications accepted
- Enroll Number of new students enrolled
- Top10perc Pct. new students from top 10% of H.S. class
- Top25perc Pct. new students from top 25% of H.S. class
- F.Undergrad Number of fulltime undergraduates
- P.Undergrad Number of parttime undergraduates
- Outstate Out-of-state tuition
- Room.Board Room and board costs
- Books Estimated book costs
- Personal Estimated personal spending
- PhD Pct. of faculty with Ph.D.'s
- Terminal Pct. of faculty with terminal degree
- S.F.Ratio Student/faculty ratio
- perc.alumni Pct. alumni who donate
- Expend Instructional expenditure per student
- Grad.Rate Graduation rate

We will predict the number of applications received Apps using all other variables in the College data set and apply **LASSO** and **Tree regression** models and compare their performance (test MSE).

Part 1 **(50 points)**

LASSO

Predict the number of applications received Apps using all other variables in the College data set using **LASSO** model for variable selection:

a. Split the data set randomly into training and test data set. (5 points)

b. Fit Lasso model using on the training data set. (5 points)

c. Perform cross-validation on the training data set to choose the best lambda. (5 points)

d. Estimate the predicted values using the best lambda obtained in part (c) on the test data and compute test MSE. (10 points)

e. Compare the Lasso predicted test MSE (with the best lambda) with the null model (lambda=infinity) test MSE and least square regression model (lambda=0) test MSE. What do you conclude? (10 points)

f. Now construct the Lasso model for the entire data set and obtain the Lasso coefficients using the best lambda obtained in part (c) and report the number of non-zero coefficient estimates. (7 points)

g. Now use the Lasso predictors obtained in part (f) to fit the Linear Regression Model and report the summary of the linear model. (8 points)

Note: Your grade will be based on accuracy of the code. You do not have to provide any written explanation of the Part 1 (a) – (g) in the code.

Part 2 **(50 points)**

Regression Tree

Predict the number of applications received Apps using all other variables in the College data set based on a Regression Tree:

Perform the following tasks: **Use the training and test data set that you created in Part 1(a).**

a. Fit a Regression Tree (max depth =3) to the training data, with Apps as the response and the all other variables as predictors. Create a plot of the tree. Note how many terminal nodes the tree has. (4 points)

b. Print: Training accuracy and Test Accuracy, and Test MSE, (6 points)

c. Now to find the optimal depth that will improve performance use cost complexity pruning to prune the decision tree. (5 points)

d. Based on part (c), produce a plot to observe total impurity versus effective alpha for training set. What do you observe? (5 points)

e. Construct plots to display number of nodes versus alpha and tree depth versus alpha. What do you observe? (5 points)

f. Construct plot to observe changes of training and test accuracy with respect to alpha. What do you observe? (5 points)

g. Find which alpha corresponds to the highest test score? Find the depth of the tree corresponding to the best alpha and create the tree using that depth. Print the decision tree. Compute the mean squared test error corresponding to the tree with best alpha. What do you observe? (5 points)

h. Compare the above test error rates in part (g) with the one obtained using LASSO regression (test MSE) in Part 1(d). Which model will you select for this decision problem and why? Provide a detail explanation. (15 points)

Note: Provide your explanations in the python at end of each part (comment them appropriately). Your grade will be based on the execution of the code.

Deliverables:

1. Submit Python code for both parts of the assignment in one single code script. (Team Submission)

2. Python code should provide comments on sections of the assignment the code is intended for. For example mark the code (using comment symbol #) for which parts 1 and 2 the code is for marking each subparts a – h clearly.

3. Submit Python scripts electronically in iCollege.

4. **The Python code should clearly indicate Names of the Contributors at the beginning of the code. If a team members' name is not present as one of the contributors, he/she will earn zero for the assignment.**