University of Bristol
EFIMM0095 (Data Science for Economics)
Professor Vincent Han

Coursework 3

Due: 6 May at 11:00am

- This coursework is a team project. The output will be evaluated as a team, that is, all team members will receive the same score. Therefore, cooperation among team members is important. If there is a free-rider problem, try to figure out how to resolve it. Please show work ethic to your team members.
- All the questions should be answered using Python. Report all the codes used in answering the questions.
- Please submit one pdf file per group with the names of all the members. Please indicate the group number in the file name.
- All questions are equally weighted. This assignment is one of three assignments that consist of 50% of your grade.

1. Consider the data set *credit.csv* which contains information about individual's credit scores and other characteristics. Using this dataset, we want to understand which characteristics are important in predicting average credit card debt (*balance*). Specifically, we want to consider the following (nonlinear) regression model:

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{J} \gamma_{jk} x_{ij} x_{ik} + e_i$$

where y_i is the credit score and $\{x_{ij}\}_{j=1}^{J}$ are the continuous characteristics (*standardized*) and characteristic dummies. We want to make sure creating relevant dummy variables and dropping missing observation in *credit.csv* before proceeding.

   (a) What are the number of observations (n) and the number of predictors (p) in this regression? Make an argument why lasso may be the better procedure in this context, compared to OLS.

   (b) Conduct the estimation of the model using lasso. For the first regression, set $\lambda = 0.5$. Report and discuss the result.

   (c) Calculate the training MSE given the estimation results.

   (d) This time, calculate $CV_{(k)}$ defined in (5.3) in the textbook (p. 181) using a 5-fold CV.

   (e) Compare the answer in (c) and (d) and explain the discrepancy.

(f) We want to choose the optimal λ using a 5-fold CV as in (d) and do subsequent analyses. We do *not* want to use a built-in function to choose the optimal λ. Instead, proceed with the following:

i. Create a grid for λ with 100 grid points. The range should include zero as a starting value. The range should be determined at your discretion. For example, try to run several lassos with, say, $\lambda = 10, 100, 1000$, and see how sensitively the result changes. You want to pick a good range that yields different lasso estimates, i.e., different number of variables selected. Given the grid, calculate CV_n for each λ and draw a picture with λ on the x-axis and CV_n on the y-axis. You may want to use a loop in your code to implement this. Determine the optimal choice of λ.

ii. Conclude by reporting the final estimation results with the optimal λ.

iii. Report the coefficient plot as in the left panel of Figure 6.6 in the text book.

iv. Predict the balance of a married 70-year-old Asian female whose income is 100, limit is 6000, rating is 500, has 3 cards, has 12 years of education, and is not a student.

2. Consider using tree-based methods for the same data set.

(a) We want to fit a random tree with maximum depth of 3. Visualize this tree and interpret the result.

(b) To evaluate the predictive performance of this tree, use 5-fold CV to calculate the test MSE.

(c) Now we want to use a random forest with the maximum depth of 3. In doing so, we want to vary the number of trees in the forest and compare the performance. Calculate the test MSEs with the number of trees in $\{1, 5, 10, 50, 100, 200\}$. Then, calculate the test MSEs without specifying the maximum depth. Plot all the results in one graph. Discuss the findings.

(d) Using the random forest that yields the lowest test MSE in (c), predict the balance of the person in part 1(f)-v.

(e) For this random forest, plot a graph that shows the importance of variables (i.e., similar to Figure 8.9 in the text book) and discuss the findings. Also, compare the findings with what you found in (a) as well as in part 1(f)-iii.