Assignment 5

February 24, 2022

Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

d. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is well-organized, clear, and concise (3 points)

- 1. It's easy enough to identify different sections of the homework assignment (e.g., if there are different sections of an assignment, they're clearly distinguishable by section headers or the like)

- 2. It's clear which code/markdown blocks correspond to which questions.

- 3. There aren't excessively long outputs of extraneous information (e.g., no printouts of entire data frames without good reason)

e. This assignment is **due at 11:59pm on Monday, March 7th**. Good luck!

f. There are two parts in this assignment - Part 1 and Part 2. You have the **option to choose any one part** out of the two and submit the assignment. Each part is worth 35 points. Note that both parts are relevant for the final exam. However, for the purpose of this assignment you need to do any one part of your choice. If you do both parts, only the first part will be graded.

Submissions will be graded with a maximum of **40 points** – 35 points for code & answers, 5 points for anonymity and proper formatting.

Part 1

The datasets *ENB2012_Train.csv* and *ENB2012_Test.csv* provide data on energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters.

1. X1 Relative Compactness
2. X2 Surface Area
3. X3 Wall Area
4. X4 Roof Area
5. X5 Overall Height
6. X6 Orientation
7. X7 Glazing Area
8. X8 Glazing Area Distribution
9. y1 Heating Load

For this part, if the three criteria (AIC / BIC / Adjusted R-squared) suggest different numbers of predictors, consider the best model as per the BIC criterion.

Q1a

Suppose that we want to implement the best subset selection algorithm to find the first order predictors (X1-X8) that can predict heating load (y1). How many models for E(y1) are possible, if the model includes (i) one variable, (ii) three variables, and (iii) eight variables? Show your steps without running any code.

Note: The notation E(y1) means the expected value of y1 or the mean of y1

(3 points for answer)

Q1b

Implement the best subset selection algorithm to find the "best" first-order predictors of heating load (y1). Print out the model summary.
Use *ENB2012_Train.csv* and consider only the first-order terms.

(4 points for code)

Q1c

Should R-squared be used to select from among a set of models with different numbers of predictors? Justify your answer.

(1 point for answer, 2 points for justification)

Q1d

Calculate the RMSE of the model found in (b). Compare it with the RMSE of the model using all first-order predictors. You will find that the two RMSEs are similar. Seems like the best subset model didn't help improve prediction. However, did it help improve inference? Justify your answer.

Hint: VIF!

(2 points for code, 2 points for answer)

Q1e

Let us consider adding all the 2-factor interactions of the predictors in the model. Please answer the following questions without running code.
i) How many predictors do we have in total?
ii) Assume best subset selection is used. How many models are fitted in total?
iii) Assume forward selection is used. How many models are fitted in total?
iv) Assume backward selection is used. How many models are fitted in total?
v) How many models will be developed in the iteration that contains exactly 10 predictors in each model – for best subsets, fwd/bwd regression?
vi) What approach would you choose for variable selection (amongst best subset, forward selection, backward selection)?

(8 points for answer)

Q1f

Use forward selection to find the "best" first-order predictors and 2-factor interactions of the predictors of y1 (Heating Load). Print out the model summary.

(5 points for code)

Q1g

Calculate the RMSE of the model found in (f). Compare it with:

i) the RMSE of model you found in (b) and

ii) the RMSE of the model using all first-order and 2-fatcor interaction terms and discuss about your finding.

(2 points for code, 2 points for answer)

Q1h

Assume that we found another dataset of 32 variabels on the same set of 768 buildings (542 for training) that we would want to add into our model. We want find the "best" model of all 40 predictors and their 2-factor interaction terms. Would you choose forward or backward selection? Justify your answer.

(1 point for answer, 3 points for justification)

Part 2

See https://exoplanetarchive.ipac.caltech.edu (for context/source). We are using the Composite Planetary Systems dataset

Q2a

Say we're interested in modeling the radius of exoplanets in kilometers, which is named as `pl_rade` in the data. Note that the variable `pl_rade` captures the radius of each plant as a proportion of Earth's radius, which is approximately 6,378.1370 km.

Develop a linear regression model to predict `pl_rade` using all the variables in *train_CompositePlanetarySystems.csv* except `pl_name`, `disc_facility` and `disc_locale`. Find the RMSE (Root mean squared error) of the model on *test1_CompositePlanetarySystems.csv* and *test2_CompositePlanetarySystems.csv*.

(4 points for code)

Q2b

Develop a ridge regression model to predict `pl_rade` using all the variables in *train_CompositePlanetarySystems.csv* except `pl_name`, `disc_facility` and `disc_locale`. What is the optimal value of the tuning parameter λ?

Hint: You may use the following grid of lambda values to find the optimal λ: `alphas = 10**np.linspace(2,0.5,200)*0.5`

Remember to standardize data before fitting the ridge regression model

(5 points for code)

Q2c

Use the optimal value of λ found in the previous question to develop a ridge regression model. What is the RMSE of the model on *test1_CompositePlanetarySystems.csv* and *test2_CompositePlanetarySystems.csv*?

(5 points for code)

Q2d

Note that ridge regression has a much lower RMSE on test datasets as compared to Ordinary least squares (OLS) regression. Shrinking the coefficients has improved the model fit. Appreciate it. Which are the top two predictors for which the coefficients have shrunk the most?

To answer this question, find the ridge regression estimates for $\lambda = 10^{-10}$. Treat these estimates as OLS estimates and find the predictors for which these estimates have shrunk the most in the model developed in 2c.

(4 points for code, 1 point for answer)

Q2e

Why do you think the coefficients of the two variables indentified in the previous question shrunk the most?

Hint: VIF!

(4 points for justification - including any code used)

Q2f

Develop a lasso model to predict `pl_rade` using all the variables in *train_CompositePlanetarySystems.csv* except `pl_name`, `disc_facility` and `disc_locale`. What is the optimal value of the tuning parameter λ?

Hint: You may use the following grid of lambda values to find the optimal λ: `alphas = 10**np.linspace(0,-2.5,200)*0.5`

(4 points for code)

Q2g

Use the optimal value of λ found in the previous question to develop a lasso model. What is the RMSE of the model on *test1_CompositePlanetarySystems.csv* and *test2_CompositePlanetarySystems.csv*?

(5 points for code)

Q2h

Note that lasso has a much lower RMSE on test datasets as compared to Ordinary least squares (OLS) regression. Shrinking the coefficients has improved the model fit. Appreciate it. Which variables have been eliminated by lasso?

To answer this question, find the predictors whose coefficients are 0 in the lasso model.

(2 points for code, 1 point for answer)