STA 4320 – Takehome Exam

Instructor: He Jiang

Name: _____
Student Number: _____

This is a takehome project/exam. You must complete this assignment entirely on your own, without any discussions with other people. Your solution must be written entirely by yourself.

You do not have to submit this instruction page.

Please compile your coding into a single PDF file.

Please submit your compiled (PDF file) report to the corresponding assignment on Gradescope.

For the current assignment, your report should be no longer than the following amount of pages: 4

1. In this question, we consider fitting a LASSO regression on the Credit dataset, from the ISLR2 package. For consistency of the results, please use set.seed(1) for this assignment.

   Grading Method

   The grading of this assignment will be based on:

   Correctness.

   Tasks for this Assignment

   (a) (2 points) Load the relevant packages (ISLR2 for the data and glmnet for lasso and cross validation) and load the Credit dataset. Check that there are no na terms in the Credit dataset.

   (b) (1 point) Using Balance as the response variable, build the data matrix $\mathbf{X}$, including all data excluding the response and excluding the intercept column, and build the response vector $\mathbf{y}$, consisting of Balance values.

   (c) (1 point) Split the data into a training set and a testing set of exactly equal size. The entire dataset consists of $n = 400$ rows, so the training set and the testing set should each consist of 200 rows. (Hint: we can use train = sample(1:n, n/2, replace = FALSE)). Importantly, please use set.seed(1) for this assignment.

   (d) (3 points) On the training set, fit a LASSO regression, using the following λ as the initial grid: 10^seq(10, -2, length = 100). Then, on the same training set, conduct a 10-fold cross validation, produce a plot of Mean Squared Error vs $\log \lambda$.

   (e) (1 point) Using the best λ from cross validation, compute the test error on the testing set (of size 200).

   (f) (1 point) Using the best λ from cross validation, fit a LASSO regression on the entire dataset (of size $n = 400$). Report the resulting coefficients in a vector format, rounded to 4 decimal places. (Hint: there should be 12 coefficients, including the intercept).

   (g) (1 point) Correctly label the question on Gradescope. Also, report is in required format, and within the required page limit (see the beginning of this assignment). For this assignment, please label the entire solution to Question 1 on Gradescope.