STAT 542/CS 598 Fall 2018

Coding Assignment 2

Due Thursday, Sept. 27, 11:30 p.m. (PT)

The assignment is related to the Boston Housing data. The original data is from the `R` library "`mlbench`", which has 506 observations on 19 variables.

crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B-0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
medv	median value of owner-occupied homes in USD 1000's
cmedv	corrected median value of owner-occupied homes in USD 1000's
town	name of town
tract	census tract
lon	longitude of census tract
lat	latitude of census tract

First, we apply some suggested transformations on the data, then remove three variables `medv`, `town`, and `tract`, and use `cmedv` as the response variable `Y`.

Consider following 10 procedures:

- Full: run a linear regression model using all features,
- AIC.F and AIC.B: Forward/backward selection with AIC,
- BIC.F and BIC.B: Forward/backward selection with BIC,
- R.min and R.1se: Ridge regression using lambda.min or lambda.1se,
- L.min and L.1se: Lasso using lambda.min or lambda.1se,
- L.Refit: Refit the model selected by Lasso using lambda.1se.

1. Load `BostonHousing1.Rdata`, which has 16 variables including the response variable `Y`.

   a) Repeat the following simulation 50 times. In each iteration, randomly split the data into two parts, 75% for training and 25% for testing. fit the model based on the training data and obtain a prediction on the test data, record the mean squared prediction error (MSPE) on the test set, the selected model-size or effect dimension (for Ridge), and the computation time for each procedure.

      Exclude intercept in computing model-size or effect dimension.

   b) Summarize your results on MSPE and model size graphically, e.g., using boxplot or stripchart.

2. Load `BostonHousing2.Rdata`, which has 135 variables including the response variable `Y`. In addition to the original 15 predictors, the data contains their quadratic and all pairwise interaction terms.

   Repeat (a-b) above for only five methods: R.min, R.1se, L.min L.1se, and L.Refit.

3. Load `BostonHousing3.Rdata`, which has 635 variables including the response variable `Y`. In addition to `BostonHousing2.Rdata`, the data contains 500 noise features.

   Repeat (a-b) above for only five methods: R.min, R.1se, L.min L.1se, and L.Refit.

(Continue on the next page ⟶)

What you need to submit?
A PDF file (maximum two-page) and the R/Python code that produces the PDF file. Your code will be run in a directory that has the three data files.

- The PDF file should contain three sets of figures (one for each data set) which provide graphical summary of MSPE and model size.
- The PDF file should contain the computation time for each procedure for each data set. Students can display the computation time graphically or just provide the numbers.
- Students are allowed to use R/Python code to generate Markdown file in PDF. Since the file size is restricted to be two pages, suggest to hide your code and only display the results.
- Name your R/Python file starting with

  `Assignment_1_xxxx_netID..`,

  where "xxxx" is the last 4-dig of your University ID.

  For example, the submission for Max Y. Chen with UID 672757127 and netID mychen12 would be named

  `Assignment_1_7127_mychen12_MaxChen.R`.

  You can add whatever characters after your netID.
- Name the PDF file similarly, starting with

  `AssignmentOutput_1_xxxx_netID...pdf`,

  where "xxxx" is the last 4-dig of your University ID.
- Set seed at the beginning of your code to be the last 4-dig of your University ID.