Spring 2022

Quantitative Analysis
Chung-Ming Kuan
Jui-Chung Yang

Final: Computer Exam

This is a 2-hour, computer-based exam. You can refer to any "hard copy" or any materials stored in your computer, while access to the Internet and discussing with others are strictly prohibited. Good luck!

1. (10 points)
   In your answer sheet, there are two simulated data $(x_i)_{i=1}^{100}$ and $(y_i)_{i=1}^{100}$. Let $\bar{x} = (1/100)\sum_i x_i$ and $\bar{y} = (1/100)\sum_i y_i$. The estimator we are interested in is:

$$\hat{z} = \frac{\bar{x}^5}{(\bar{x}+\bar{y})^2}.$$

   Suppose that the true data generating process for x_i and y_i are **unknown** to us, and that the sample $(x_i, y_i)_{i=1}^{100}$ in the Answer Sheet is the only available data we have. Please bootstrap the given sample for $10,000$ times to estimate the variance of the estimator $\hat{z}$. For each bootstrap, set `set.seed(b)`, where `b` is the indicator of the times of boostrap.

2. (15 points)
   Consider the following simulation settings. The data generating process (DGP) is

$$\text{DGP}: y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \;\; \beta_0 = 2, \; \beta_1 = \beta_2 = 1,$$

   where all x_{ij} and ϵ_i are i.i.d. $\mathcal{N}(0,1)$, and the sample size is $n = 55$. We estimate a LASSO model with 50 regressors, including the "signal" variables in the DGP and other "noise" variables (generated as i.i.d. $\mathcal{N}(0,1)$). Replicate this simulation for 100 times with `set.seed(r)`, where `r` is the indicator of the times of the simulation.

   (a) What is the percentage for x_1 being picked by LASSO (i.e. $\hat{\beta}_1 \neq 0$)? For each simulation, consider $\lambda \in [0, 2]$ with 100 partitions, with the optimal λ determined by 10-fold CV.

   (b) What is the percentage for the first signal variable x_1 being picked by backward stepwise selection?

   (c) What are the bias, variance, and MSE (calculated by $bias^2 + variance$) for $\hat{\beta}_1$ calculated by OLS, LASSO, and backward stepwise selection?

3. (15 points)
Please load the data set `Wage` from the package `ISLR`. Suppose that we want to fit the variable of interest `wage` on a single predictor `age`.

   (a) Use the natural cubic spline with 3 knots located at `age` = 20, 40 and 60, and boundary knots located 10 and 70. Please fit the model and calculate this model's residual sum of squares.

   (b) Use the local regression with span = 50%. Please fit the model and calculate this model's residual sum of squares.

   (c) Suppose that

$$\texttt{wage} = \beta_0 + f_1(\texttt{age}) + f_2(\texttt{year}) + u,$$

   where f_1 is a step function with five intervals of equal length, while f_2 is a natural cubic spline with 1 interior knot located at the median of `year`. Please fit this generalized additive model and calculate its residual sum of squares.

4. (15 points)
Please load the `heart` data set from the package `kmed`. Let `class` be our variable of interest and set it to 0 if it's 0 and 1 otherwise (that is, `class` only has 2 values, 0 and 1), and all the other 13 variables in the data set be our predictors. Some data processing has been done in your answer sheet.

   (a) Grow a classification tree with the gini index as the splitting criterion and then determine the optimal node for this tree using 10-fold CV with `set.seed(1)`.

   (b) Plot the OOB errors of random forest with m being 2, 4, 10 and 13, respectively, based on the `heart` data with `set.seed(1)`.

5. (15 points)
Please load the `Boston` data set from the package `MASS`. Let `medv` be our response and all the other 13 variables be our predictors.

   (a) Please use the LASSO regression with 10-fold CV with `set.seed(1)` to choose the optimal λ^* from $\lambda \in [0, 1]$ with 100 partitions. What is the value of λ^*?

   (b) Treat 50% of the data set as the training set with `set.seed(1)` and the remaining as the testing set. Fit a linear regression with the variables chosen by the LASSO with λ^* and evaluate its test MSE with the testing set.

   (c) Please fit a neural network with 4 hidden layers (5, 4, 3 and 2 neurons in each layer) with the training set in part (b) and evaluate its test MSE with the testing set with `set.seed(1)`.