ANLY512: Homework 2

Due Feb 24 11:59pm

Problem 1: Use best subset selection to determine best heating values

For bioenergy production, the heating value is a measure of the amount of heat released during combustion. The Higher heating value (HHV) is a particular method for determining the heat released during combustion. The higher HHV the more energy released for a given amount of material. You will use the `biomass` dataset from the `model_data` package. Run a `?biomass` after importing the data to read about the domain. The response variable is HHV and the predictor variables are the percentages of different elements. Do not include the sample and dataset variables in your analysis.

a: Create scatterplots of the response and predictor variables and comment on your findings.

b: Split the dataset into train and test datasets using train-test split.

c: Use `regsubsets()` to perform best subset selection to pick the best model according to C_p, BIC, and adjusted R^2.

d: Repeat this procedure for forward stepwise selection and backward stepwise selection, compare the best models from each selection method.

e: Use the `predict()` function to investigate the test performance in RMSE using your "best model".

Problem 2: Create your own cross validation algorithm to predict the interest rate for loan applications

Lending club gained fame for being one of the first major players in retail lending. The dataset `lending_club` in the `model_data` package includes 9,857 loans that were provided. Interest rates of loans are often a good indicator of the level of risk associated with lending. If you are likely to pay back a loan, then you will likely be charged lower interest than someone who has a higher chance of default. Your goal is to determine the best model for predicting the interest rate charged to borrowers using best, forward, and backward subset selection within a five-fold cross-validation framework.

Prep steps: - drop all rows with missing data in the following columns

a: Create a correlation plot of all the numeric variables in the dataset using the `corrplot` package to create a high quality graph, then comment on your findings

b: Run best, forward, and backward subset selection on the entire dataset comment on the findings

c: Create a five-fold cross-validation algorithm usng for loops to compare the CV mse performance of your best two models

Problem 3: Properties of k-fold cross validation

Suppose we are given a training set with `n` observations and want to conduct `k`-fold cross-validation. Assume always that `n = km` where `m` is an integer.

a: Let `k = 2`. Explain carefully why there are $\frac{1}{2}\binom{n}{m}$ ways to partition the data into 2 folds.

b: Let `k = 3`. Explain carefully why there are $\frac{n!}{3!m!m!m!}$ ways to partition the data into 3 folds.

c: Guess a formula for the number of ways to partition the data into `k` folds for general `k`. Check if your formula gives the correct answer for `k=n` (leave-one-out c.v.).

Problem 4: Using cross-validation to select best advertising budget

In this problem, we use the Advertising data download here. We want to predict Sales from TV, Radio and Newspaper, using multiple regression with all three predictors plus up to one interaction term of these three predictors, e.g. TV * Radio or Radio * Newspaper.

a: Should such an interaction term be included? Which one? Try to answer this question by estimating the residual standard error using 10-fold cross validation for all four possible models.

b: Create a single plot showing the return on investment of each advertsing method where the y-axis is `Sales` and the x-axis is advertsing dollars. There should be three lines, one for each method. The slope is the coefficient from you regression. What is the best advertising method to invest in based on return on investment?

Problem 5: ISLR 6.8 #8(a-d)