Statistical Learning and Analysis
Instructor: Prof. Varun Rai
Assignment 3

Question 1

Let's talk about k-fold validation.

a. Explain how k-fold cross-validation is implemented.
b. What are the advantages and disadvantages of k-fold cross validation relative to
    i. The validation set approach; and
    ii. LOOCV?

Question 2

In Assignment 2, we used logistic regression to predict the probability of high-quality red wine. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

a. Fit a logistic regression model that uses all predictors (except quality) to predict final_quality.
b. Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
    i. Split the sample set into a training set and a validation set.
    ii. Fit a multiple logistic regression model using only the training observations.
    iii. Obtain a prediction of final_quality for each wine in the validation set by computing the posterior probability of high-quality for that wine, and classifying the wine to the high-quality category if the posterior probability is greater than 0.5.
    iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

Question 3

We continue to consider the use of the same logistic regression model in the question 2.

a. Write a function, boot_fn(), that takes as input the redwine data set as well as an index of the observations, and that outputs the coefficient estimates for each predictor in the multiple logistic regression model.
b. Use the boot() function together with your boot_fn() function to estimate the standard errors of the logistic regression coefficients for each predictor.
c. Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function. One way to get these values in python is use of statsmodel library.