Stat 485 HW4 **Spring 2024** **Due February 14, 2024**

1. Read Sections 7.10.3 and 7.11 in the ESL textbook by yourself.

2. Consider the simulation study described in Section 7.10.2 in the ESL textbook. There are $N = 50$ samples with 25 labeled as "1" and 25 labeled as "2", denoted by Y as the response. There are $p = 5000$ covariates $X_1, \ldots, X_{5000}$ simulated i.i.d. from standard normal distribution, which are also independent of Y.

(1) Do 50 times simulations as follows (called *Wrong Procedure*): 1°) Find the top 100 predictors $X_{(1)}, \ldots, X_{(100)}$ out of $X_1, \ldots, X_{5000}$ in terms of absolute sample correlation with Y; 2°) Use 5-fold cross-validation to estimate the error rate of 1-nearest neighbor classifier with the selected 100 predictors (*Hint:* You may use `R` function `knn` in package `class`). Find the average cross-validation error rate.

(2) Do 50 times simulations as follows (called *Correct Procedure*): 1°) Divide the $N = 50$ samples into $K = 5$ cross-validation folds equally and randomly; 2°) For each fold $k = 1, \ldots, K$, find the top 100 predictors $X_{(1)}, \ldots, X_{(100)}$ out of $X_1, \ldots, X_{5000}$ in terms of absolute sample correlation with Y, using all the samples except those in fold k; then employ 1-nearest neighbor classifier with all samples except those in fold k (training data) to predict the responses in fold k (testing data), using $X_{(1)}, \ldots, X_{(100)}$ only and recording the testing error rate; 3°) Find the average cross-validation error rate.

(3) The expected error rate of any classifier is 50%. Are your average cross-validation error rates obtained in (1) and (2) close to 50%? If not, why?