Econ 108 FAll 2022
Problem Set 7

This problem set is due at 7pm, Friday, November 18, 2022.

1. Familiarize with the codes in semiconductor.R, glass.R and credit.R posted on canvas (in the modified_code directory under the files tab) before working on this question.

   (a) Continue from Question 1 of the last problem set. Split the sample so that 1/4 of the sample is used for testing, and the remaining 3/4 is used for training.

   (b) Apply a K-nearest neigborhood classifier to classify the emails in the testing sample into spam and not-spam. You should try a few different values of K, for example $K = 5, 10, 30, 50, 100$, etc.

   (c) For each K, calculate the accuracy of the classication in the test data. Accuracy is defined as the fraction of times when the data is classified correctly. If the classifications are coded as factors with levels 0 and 1, you might find it useful to convert the factors into integers. A R command that does this is: as.integer(levels(x))[x].

   (d) Which K generates the most accurate out of sample prediction?

2. Also continue from Question 1 of the last problem set. Split the sample into half for training and half for testing.

   (a) Draw an in-sample ROC curve and an out-of-sample ROC curve using a LASSO logit regression.

   (b) Suppose the average cost of misclassifying a normal email as spam is $100, while the average cost of misclassifying a spam email as normal is $30. Derive the optimal classification rule as function of the predicted probability of an email being spam.

   (c) Using the LASSO regression, apply the above classification rule to the test sample. Calculate the resulting specificity and sensitivity, and plot the corresponding point on the ROC curve.

   (d) Redo the previous two exercises assuming now that the average cost of misclassifying a normal email as spam is $30, while the average cost of misclassifying a spam email as normal is $100.