STA4241 Homework 1, Fall 2021

Please turn in your own work, though you may discuss these problems with your classmates, professor, and TA. The assignment is due on Wednesday, September 15th at midnight. Note that there are data sets used in both the second and third problems, and the R code to read in the data is available on the course website.

(1) Show that the least squares solution for β_0 and β_1 in simple linear regression is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

Note that you do not need to work with second derivatives to prove that your solution is in fact a minimum of the least squares objective function.

(2) Here we will compare the Bayes classifier and the K-nearest neighbors approach to classification. Suppose we observe two predictors, X_1 and X_2, along with an outcome Y, which is binary. Suppose also that we know the true conditional distribution of $Y|X$, which is given by

$$P(Y = 1|X_1, X_2) = \Phi(0.5X_1 - 0.4X_2)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Note that you can calculate the CDF of a normal random variable using the pnorm function in R.

(a) Treat the first 500 subjects in the data as your training data, and the remaining 500 as your test data.

(i) If we use the Bayes classifier using the *known* probability above, do we expect the error rates to be different between the training and test data sets?

(ii) Use the Bayes classifier to find the error rate for the training data and for the test data.

(iii) Use K-nearest neighbors with $k = 3$ to classify the outcomes in the test data set. What is the test error rate now?

(iv) Given the test error rate in the previous part, and the error rates found in part (i), do you think that $k = 3$ is the best choice of k? Why or why not?

(v) Create a plot that shows the test error rate as a function of k. What do you think is the best choice of k?

(vi) Do you think that KNN does a good job at approximating the Bayes classifier in this data set? Why or why not?

(vii) Now suppose that we have 20 additional covariates which are not predictive of Y. These are saved in the data file as xrandom. use the KNN method with $k = 40$, but use the original two predictors as well as the additional 20 predictors. What is the test set error rate?

(viii) What does the previous part tell you about KNN?

(3) Load the stock market data used in the textbook into R.

(a) Fit a linear regression model that aims to predict Today (daily return in the stock market) using lags 1-5, Year, and Volume.

(i) How did you include Year into the model and why?

(ii) Run an F-test to determine if the covariates are predictive of the outcome. Your answer should include a test statistic, p-value, and decision regarding a hypothesis test.

(iii) Fit a model that includes lag 1 in the model with a three degree of freedom polynomial. Run a test to see if this model fits the data better than the previous model, which only included lag 1 linearly.

(b) Now we will use the same predictors, but the outcome is the binary variable Direction, which indicates whether the stock market went up or down that day.

(i) Randomly select half of the data to be your training set, and half to be your test set. Use KNN to predict the outcome on the test data and find the smallest test set error you can achieve.

(ii) Does the above result tell you anything about how predictive the covariates are of the outcome?

(4) Suppose that we are interested in making predictions for a continuous outcome based on a set of covariates $\boldsymbol{X}$.

(i) I want to create a confidence interval for the average value of the outcome among subjects with $\boldsymbol{X} = \boldsymbol{x}_0$ as well as for a randomly chosen individual with $\boldsymbol{X} = \boldsymbol{x}_0$. Are these two confidence intervals the same? If yes, explain why. If no, explain why not and how they differ.

(ii) Do the widths of the two confidence intervals described in part (i) go to zero as $n \to \infty$? Explain why or why not.