SDSC5001 Statistical Machine Learning I
Assignment #1

Deadline: 10 October, Friday @ 11:59 PM

1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.

2. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, and test error, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be four curves. Make sure to label each one.

(b) Explain why each of the four curves has the shape displayed in part (a).

3. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the ideal decision boundary (with the smallest test error) in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

4. Use the Auto data set in the ISLP package for this problem. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor?

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

5. Suppose we have a data set with five predictors, X_1= GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

(a) Which answer is correct? Why?
i. For a fixed value of IQ and GPA, males earn more, on average, than females.
ii. For a fixed value of IQ and GPA, females earn more, on average, than males.
iii. For a fixed value of IQ and GPA, males earn more, on average, than females provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, females earn more, on average, than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

6. This problem focuses on the *multicollinearity* problem. Assume three variables X_1, X_2, and Y have the following relationship:

$$
\begin{aligned}
X_1 &\sim \text{Uniform}[0,1] \\
X_2 &= 0.5X_1 + \epsilon/10 \qquad \text{where } \epsilon \sim N(0,1) \\
Y &= 2 + 2X_1 + 0.3X_2 + e \quad \text{where } e \sim N(0,1)
\end{aligned}
$$

(a) Simulate a data set with 100 observations of the three variables, and then answer the following questions using the simulated data.

(b) What is the correlation between X_1 and X_2? Create a scatter plot displaying the relationship between the two variables.

(c) Fit a least squares regression to predict Y using X_1 and X_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0: \beta_1 = 0$? How about the null hypothesis $H_0: \beta_2 = 0$?

(d) Now fit a least squares regression to predict Y using only X_1. Comment on your results. Can you reject the null hypothesis $H_0: \beta_1 = 0$?

(e) Now fit a least squares regression to predict Y using only X_2. Comment on your results. Can you reject the null hypothesis $H_0: \beta_1 = 0$?

(f) Do the results obtained in (c)-(e) contradict each other? Explain your answer.

(g) Now suppose we obtain one additional observation (0.1, 0.8, 6) (i.e., $x_1 = 0.1, x_2 = 0.8, y = 6$), which was unfortunately mismeasured. Please add this observation to the simulated data set and re-fit the linear models in (c)-(e) using the new data. What effect does this new observation have on each model? In each model, is this observation an outlier (outlying Y observation)? A high-leverage point (outlying X observation)? Or both? Explain your answer.