CSE 4095: Introduction to Data Science and Engineering
Assignment 2
Due: March 5, 2024 at 11:59pm

Instructions:

- You must support your answers to receive credit.
- Answers can be typed or handwritten, and should be readable.
- Submit the assignment in one file via HuskyCT.

1. (12 points) Adapted from ISLP 2.4.2.

   Explain whether each scenario is a *classification* or *regression* problem, and indicate whether we are most interested in *inference* or *prediction*. Finally, provide n, the number of observations, and p, the number of features.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 30 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and twelve other variables.

   (c) We are interested in predicting the percent change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2016. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

2. (16 points) Adapted from ISLP 2.4.7.

   The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

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

   Suppose we with to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

   (a) Compute the Euclidean distance between each observation and the test point,

$$X_1 = X_2 = X_3 = 0.$$

i. What is our prediction with $K = 1$? Why?

ii. What is our prediction with $K = 3$? Why?

(b) Compute the Manhattan distance between each observation and the same test point.

i. What is our prediction with $K = 1$? Why?

ii. What is our prediction with $K = 3$? Why?

3. (17 points) Adapted from ISLP 2.4.10.

Load the `Boston` data set, which is part of the ISLP library. Details can be found at `https://intro-stat-learning.github.io/ISLP/datasets/Boston.html`.

(a) How many rows are in this data set? How many columns? What do the rows represent?

(b) Make pairwise scatterplots of the predictors (columns) in this data set with the per capita crime rate.

(c) Are any of the predictors correlated with per capita crime rate? If so, explain the relationship. (Assume "correlated" to mean a Pearson correlation coefficient of at least ±0.5.)

(d) Provide a histogram for the per capita crime rate.

(e) How many suburbs of Boston have a crime rate larger than 30?

(f) Provide the range of each predictor.

(g) How many of the suburbs in this data set bound the Charles river?

(h) What is the median pupil-teacher ratio among the towns in this data set?

4. (18 points) Adapted from ISLP 3.7.8.

This question involves the use of simple linear regression on the `Auto` data set. Details can be found at `https://intro-stat-learning.github.io/ISLP/datasets/Auto.html`.

(a) Use the `sm.OLS()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summarize()` function to print the results.

Note: For simplicity, drop any records that have missing values for `horsepower` and use all remaining data for training.

i. Explain why we can deduce that there is a significant association between the predictor and the response.

ii. How strong is the relationship between the predictor and the response? Use the R^2 statistic to support your answer.

iii. Is the relationship between the predictor and the response positive or negative? How can you tell?

iv. According to this model, what is the predicted `mpg` for my 2023 Volkswagen Golf, which has a `horsepower` of 241?

v. Why might it be inappropriate to use this model to estimate the `mpg` of my Golf?

(b) Plot the response and the predictor in a new set of axes `ax`. Use the `ax.axline()` method or the `abline()` function defined in the lab to display the least squares regression line.

5. (25 points) Adapted from ISLP 3.7.9.

   This question involves the use of multiple linear regression on the `Auto` data set.

   (a) Use the `sm.OLS()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summarize()` function to print the results.

      i. Which predictors appear to have a statistically significant relationship to the response?
      ii. What does the coefficient for the year variable suggest?
      iii. Produce a residual plot which compares the observed `mpg` values to the predicted ones.
      iv. Compute the MSE for this model.

   (b) Fit a new linear model which only contains the significant variables in part (a) as predictors. Use the `summarize()` function to print the results.

      i. In your new model, do all of the predictors now have a statistically significant relationship to the response?
      ii. Compute the MSE for this new model.
      iii. If we tested both models (a) and (b) on unseen holdout data, which model would you expect to have a smaller MSE? Why?

   (c) Fit a new linear model which uses the variables `year`, `origin`, `weight`, and $\texttt{weight}^2$ as predictors. Use the `summarize()` function to print the results.

      i. Produce a residual plot which compares the observed `mpg` values to the predicted ones.
      ii. The coefficient for $\texttt{weight}^2$ is extremely low. Does this suggest that we should remove this variable from the model? Why or why not?

6. (12 points) Adapted from ISLP 3.7.10.

   This question should be answered using the `Carseats` data set. Details can be found at `https://intro-stat-learning.github.io/ISLP/datasets/Carseats.html`.

   (a) Fit a multiple regression model to predict `Sales` using `Price` and `US`. Use the `summarize()` function to print the results.

   (b) Provide an interpretation of each coefficient in the model.

   (c) Fit a new regression model that includes the interaction of `Price` and `US` as predictors. Use the `summarize()` function to print the results.

   (d) Given the estimated coefficient of the interaction term, describe how the relationship between `Price` and `Sales` is different when `US` is “Yes” or “No.”

   (e) Should this interaction be included in this model? Why or why not?