CS 4342 Assignment #1

Total points: 70

Conceptual and Theoretical Questions

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. $\sigma^2 = \mathrm{Var}(\epsilon)$, is extremely high.

Points: 5

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* and *p*.

(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 20 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 ten other variables.

(c) We are interested in predicting the % 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 2012. 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.

Points: 5

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, 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 five curves. Make sure to label each one.

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

Points: 5

4. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) Describe three real-life applications in which cluster analysis might be useful.

Points: 5

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Points: 5

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Points: 5

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	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

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

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

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

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

Points: 5

Applied Questions

1. This exercise involves the Auto data set. 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.

Hints:

- **Range:** The **range of a set** of data is the difference between the highest and lowest values in the **set**. To find the **range**, first order the data from least to greatest. Then subtract the smallest value from the largest value in the **set**.
- You can use NumPy package. It provides functions for the mean, min, and max of arrays.
- Load Data: You can use Pandas package. You might use functions in Pandas like read_csv.
- Scatter plots: check this https://jakevdp.github.io/PythonDataScienceHandbook/04.02-simple-scatter-plots.html

Points: 15

2. This exercise involves the Boston housing data set.

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

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

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

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

(g) Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Hints:

- You can find the description of Boston data set in the following link: https://rdrr.io/cran/ISLR2/man/Boston.html
- **Median** is defined as the value that is in the physically/positionally middle of a sorted **array**.
- You might use NumPy package

Points: 15