AI5213/EC4213: Machine Learning and Deep Learning **Student ID:**
Homework 1 (Statistical Learning, Regression, Classification) **Name:**

Due: Sunday, Sep 29, 11:59PM

This homework comprises a mix of conceptual problems and coding exercises. Some problems are straightforward, while others requires deeper thought. Please start early!

> Guideline for those new to data analysis using Python:
> We recommend reviewing the Lab section within each chapter (e.g., p. 40 of Ch. 2 or https://islp.readthedocs.io/en/latest/labs/Ch02-statlearn-lab.html) prior to tackling the programming tasks. Additionally, find datasets and Jupyter notebooks at https://github.com/intro-stat-learning/ISLP_labs/.

Deliverables: Submit a PDF of your homework, with an appendix listing all your code, to the Gradescope assignment entitled "HW1 Write-Up". You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. Make sure each solution is on a new page, and graphs are included in the correct sections. We need each solution to be self-contained on pages of its own.

1. On the first page of your write-up, please sign your signature next to the following statement. (Mac Preview, PDF Expert, and Foxit PDF Reader.) We want to make extra clear the consequences of cheating.

2. On the first page of your write-up, please list students who helped you or whom you helped on the homework. (Note that sending each other code is not allowed.)

For staff use only

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Total
/ 8	/ 8	/ 8	/ 8	/ 6	/ 14	/ 16	/12	/ 20	/ 100

Honor Code

Declare and sign the following statement:

"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

Name and Signature:

We welcome group discussions, but the work you submit should be entirely your own. If you use any information or pictures not from our lectures or readings, make sure to say where they came from. Please note that breaking academic rules can lead to severe penalties.

(a) Did you receive any help from anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "Junho explained to me what is asked in Question 2a")

(b) Did you give any help to anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "I pointed Josh to Ch. 2.3 since he didn't know how to proceed with Question 2")

(c) Did you find or come across code that implements any part of this assignment? If your answer is 'yes', give full details (book & page, URL & location within the page, etc.).

Q1. Solve ISLP Ch.2, Exercise #1 [8 pts]

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. [2 pts]

(b) The number of predictors p is extremely large, and the number of observations n is small. [2 pts]

(c) The relationship between the predictors and response is highly non-linear. [2 pts]

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

Q2. Solve ISLP Ch.2, Exercise #3 [8 pts]

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. [4 pts]

(b) Explain why each of the five curves has the shape displayed in part (a). [4 pts]

Q3. Solve ISLP Ch.3, Exercise #4 [8 pts]

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$. [2 pts]

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer. [2 pts]

(b) Answer (a) using test rather than training RSS. [2 pts]

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify the answer. [2 pts]

(d) Answer (c) using test rather than training RSS. [2 pts]

Q4. Solve ISLP Ch.4, Exercise #6 [6 pts]

Suppose we collect data for a group of students in a statistics class with variable $X_1 =$ hours studied, $X_2 =$ undergrad GPA, $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 0.9$.

(a) Estimate the probability that a student who studies for 30 hours and has an undergrad GPA of 3.6 gets an A in the class. [3 pts]

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class? [3 pts]

Q5. Solve ISLP Ch.4, Exercise #12 [8 pts]

Suppose that you wish to classify an observation $X \in \mathbb{R}$ into apples and oranges. You fit a logistic regression model and find that

$$\hat{\Pr}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Your friend fits a logistic regression model to the same data using the *softmax* formulation, and finds that

$$\hat{\Pr}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x)}{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x) + \exp(\hat{\alpha}_{\text{apple}0} + \hat{\alpha}_{\text{apple}1} x)}.$$

(a) What is the log odds of orange versus apple in your model? What is the log odds of orange versus apple in your friend's model? [2 pts]

(b) Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible. [2 pts]

(c) Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{orange0}} = 1.2$, $\hat{\alpha}_{\text{orange1}} = -2$, $\hat{\alpha}_{\text{apple0}} = 3$, $\hat{\alpha}_{\text{apple1}} = 0.6$. What are the coefficient estimates in your model? [2 pts]

(d) Finally, suppose you apply both models from (c) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer. [2 pts]

Q6. From CMU-10701-20: k-NN Black Box [14 pts] ☕

This is a conceptual problem, which does not require programming.

(a) In a k-NN classification problem, assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances $P_1, P_2, ..P_n$ and a new example Q, and the black box outputs the nearest neighbor of Q, say P_i and its corresponding class label C_i. Is it possible to construct a k-NN classification algorithm (w.r.t the unknown distance metrics) based on this black box alone? If so, how and if not, why not? [7 pts]

(b) If the black box returns the j nearest neighbors (and their corresponding class labels) instead of the single nearest neighbor (assume $j \neq k$), is it possible to construct a k-NN classification algorithm based on the black box? If so how, and if not why not? [7 pts]

Q7. Solve ISLP Ch.2, Exercise #10 [16 pts] ⌨

This exercise involves the Boston housing data set studied in the lab.

[Note] Your code for all of the programming exercises including this one should be attached to the PDF file submitting to the Gradescope.

(a) To begin, load in the Boston data set, which is part of the ISLP library. How many rows are in this data set? How many columns? What do the rows and columns represent? [2 pts]

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings. [2 pts]

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship. [2 pts]

(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. [2 pts]

(e) How many of the suburbs in this data set bound the Charles river? [2 pts]

(f) What is the median pupil-teacher ratio among the towns in this data set? [2 pts]

(g) Which suburb of Boston has lowest median value of owner-occupied 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. [2 pts]

(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. [2 pts]

Q8. Solve ISLP Ch.3, Exercise #15 [12 pts] ⌨

We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions. [3 pts]

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$? [3 pts]

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis. [3 pts]

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

[3 pts]

Q9. Design Your Own Problem and Solution Based on Chapters 2–4 [20 pts]

For this task, you are required to create an original problem that is related to any topic from Chapters 2 through 4 of the course. Your problem should be challenging and detailed enough to be worth 20 points in total. Along with your problem, you must provide a complete solution, including explanations and justifications wherever necessary. The problem should be divided into 5-6 sub-parts, testing a variety of skills such as conceptual understanding, computations, and interpretation.

Your work will be evaluated based on the following criteria:

- Creativity and clarity in formulating the problem [10 pts]
- Accuracy and completeness of the provided solution [10 pts]

Instructions:

- Select a topic from Chapters 2–4 that you are comfortable with or find interesting.
- Break your problem down into 5-6 sub-questions, with varying levels of difficulty or different conceptual focus.
- Make sure the total score of your problem adds up to 20 points.
- Provide detailed solutions for each sub-question, ensuring that your steps are clear and logical.

Suggested Topics:

- The curse of dimensionality
- k-nearest neighbors (k-NN)
- Comparing linear models and k-NN
- The bias-variance trade-off
- Challenges in linear models (e.g., correlated errors, multicollinearity, outliers)
- Generalized Linear Models (e.g., Poisson Regression)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)