AI5213/EC4213: Machine Learning and Deep Learning **Student ID:**
Homework 2 (Linear Regression, Classification) **Name:**

Due: Sunday, Sep 24, 11:59PM

This homework comprises a set of conceptual problems and one coding exercise. Some problems are trivial, while others will require a lot of thought. Start this homework early!

> Guideline for those new to data analysis using Python:
> We recommend you to review the Lab section within each chapter (e.g., p. 116 of Ch. 3 or https://islp.readthedocs.io/en/latest/labs/Ch03-linreg-lab.html) prior to tackling the programming tasks. Additionally, find datasets and Jupyter notebooks at https://github.com/intro-stat-learning/ISLP_labs/ and https://islp.readthedocs.io/en/latest/.
>
> Visit https://www.statlearning.com/forum — a dedicated forum created by and for the ISL community. Whether you have a question or encounter issues with ISLP labs, this platform is your go-to resource for assistance and collaborative discussions.

Deliverables:

1. Submit a PDF of your homework, with an appendix listing all your code, to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. Please start each question on a new page. If there are graphs, include those graphs in the correct sections. Do not put them in an appendix. We need each solution to be self-contained on pages of its own.

   • 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, among others, have tools to let you sign a PDF file.) We want to make extra clear the consequences of cheating.

   "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

   • 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.)

2. Submit all the code needed to reproduce your results to the Gradescope assignment entitled "HW2 Code". You must submit your code twice: once in your PDF write-up (above) so the readers can easily read it, and again in compilable/interpretable form so the readers can easily run it. Do NOT include any data files we provided. Please include a short file named README listing your name, student ID, and instructions on how to reproduce your results. Please take care that your code doesn't take up inordinate amounts of time or memory.

For staff use only

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
/ 4	/ 6	/ 8	/ 12	/ 14	/ 6	/ 6	/ 10	/ 14	/ 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."

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 whatsoever from anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "Junho explained to me what is asked in Q2-a")

(b) Did you give any help whatsoever 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 Q2")

(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.3, Exercise #1 [4 pts]

Describe the null hypothesis to which the p-values given in Table 1 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Table 1: For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold (sales) on TV, radio, and newspaper advertising budgets. (Recall that the sales variable is in thousands of units, and the three predictor variables are in thousands of dollars.)

Q2. Solve ISLP Ch.3, Exercise #3 [6 pts]

Suppose we have a data set with five predictors, $X_1 =$ GPA, $X_2 =$ IQ, $X_3 =$ Level (1 for College and 0 for High School, $X_4 =$ Interaction between GPA and IQ, and $X_5 =$ Interaction between GPA and Level. 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, and why? [2 pts]

(i) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

(ii) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

(iii) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

(iv) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

(b) Predict the salary of a college graduate with IQ of 105 and a GPA of 3.9. [2 pts]

(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. [2 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.3, Exercise #10 [12 pts]

This question should be answered using the Carseats data set.

[Note] Your code for all of the programming exercises including this one should be submitted to the corresponding Programming submission slot on Gradescope.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US. [1.5 pts]

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative! [1.5 pts]

(c) Write out the model in equation form, being careful to handle the qualitative variables properly. [1.5 pts]

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$? [1.5 pts]

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. [1.5 pts]

(f) How well do the models in (a) and (e) fit the data? [1.5 pts]

(g) Using the model from (e), obtain 95% confidence intervals for the coefficients. [1.5 pts]

(h) Is there evidence of outliers or high leverage observations in the model from (e)? [1.5 pts]

Q5. Solve ISLP Ch.3, Exercise #14 [14 pts]

This problem focuses on the *collinearity* problem.

(a) Perform the following commands in Python:

```
rng = np.random.default_rng(10)
x1 = rng.uniform(0, 1, size=100)
x2 = 0.5 * x1 + rng.normal(size=100) / 10
y = 2 + 2 * x1 + 0.3 * x2 + rng.normal(size=100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients? [2 pts]

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables. [2 pts]

(c) Using this data, fit a least squares regression to predict y using x1 and x2. 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 $H_0 : \beta_2 = 0$? [2 pts]

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$? [2 pts]

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$? [2 pts]

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer. [2 pts]

(g) Suppose we obtain one additional observation, which was unfortunately mismeasured. We use the function np.concatenate() to add this additional observation to each of x1, x2, and y.

```
x1 = np.concatenate([x1, [0.1]])
x2 = np.concatenate([x2, [0.8]])
y = np.concatenate([y, [6]])
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers. [2 pts]

Q6. 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]

Q7. Solve ISLP Ch.4, Exercise #7 [6 pts]

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. *You will need to use Bayes' theorem.*

Q8. Solve ISLP Ch.4, Exercise #12 [10 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

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

Your friend fits a logistic regression model to the same data using the *softmax* formulation in (4.13), and find that

$$\widehat{Pr}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\alpha}_{\text{orange0}} + \hat{\alpha}_{\text{orange1}} x)}{\exp(\hat{\alpha}_{\text{orange0}} + \hat{\alpha}_{\text{orange1}} x) + \exp(\hat{\alpha}_{\text{apple0}} + \hat{\alpha}_{\text{apple1}} x)}.$$

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

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

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

(d) 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]

(e) Finally, suppose that you apply both models from (d) 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]

Q9. Solve ISLP Ch.4, Exercise #13 [14 pts]

This question should be answered using the Weekly data set, which is part of the ISLP package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns? [1.5 pts]

(b) Use the full data to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones? [1.5 pts]

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression. [1.5 pts]

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010). [1.5 pts]

(e) Repeat (d) using LDA. [1.5 pts]

(f) Repeat (d) using QDA. [1.5 pts]

(g) Repeat (d) using KNN with $K = 1$. [1.5 pts]

(h) Repeat (d) using naive Bayes. [1.5 pts]

(i) Which of these methods appears to provide the best results on this data? [2 pts]

Q10. Exploratory Data Analysis with NYC Taxi Dataset [20 pts]

Please complete the exercises in the following Google Colab notebook: https://bit.ly/mldl23f-hw2-nyc-taxi and submit your .ipynb file.