Homework 1

● Graded

Student

ARNAV KRISHNAKUMAR MARDA

Total Points

86 / 94 pts

Question 1

Data & Bias **12** / 12 pts

1.1 **(a)** **6** / 6 pts

✔ **– 0 pts** Correct

– 3 pts Only one bias is provided with full explanation, or only the names of the bias are provided without any explanation.

– 3 pts Explanation of bias is provided, but specific kinds of bias (voluntary, response, etc.) are not given.

– 5 pts Only one bias is provided without any explanation.

– 6 pts No answers

1.2 **(b)** **6** / 6 pts

✔ **– 0 pts** Correct

– 3 pts Part 1 (why biased against women)

– 3 pts Part 2 (why bias against women is not eliminated; should mention how other non-gender features could still be used by the AI to infer gender)

– 6 pts No answer

Question 2

Linear Regression **38** / 41 pts

2.1 **(no title)** **5** / 6 pts

- **– 0 pts** Correct
- **– 1 pt** (a): β_0 is wrong
- **– 1 pt** (a): β_1 is wrong
- **– 0.5 pts** (a): Calculation error only for β_0 (correct formula)
- **– 0.5 pts** (a): Calculation error only for β_1 (correct formula)
- **– 0.5 pts** (b): wrong R^2 value
- ✔ **– 1 pt** (b): claim that R^2 is enough to judge if the estimated regression line fits
- **– 1 pt** (b) No explanation
- **– 1 pt** (c): wrong conclusion or no explanation over conclusion
- **– 0.5 pts** (c) correct conclusion but wrong explanation
- **– 1 pt** (c): no plot or plot is wrong
- **– 6 pts** No answer

2.2 **(no title)** **4** / 4 pts

- ✔ **– 0 pts** Correct
- **– 1 pt** claim without explanation in correlation between heart disease and wine consumption.
- **– 2 pts** wrong conclusion in correlation between heart disease and wine consumption.
- **– 1 pt** claim without explanation in concluding relation (need at least one reason)
- **– 2 pts** wrong conclusion concluding relation/no conclusion regarding causation
- **– 4 pts** No answer

2.3 **(no title)** **6** / 6 pts

- ✔ **– 0 pts** Correct
- **– 2 pts** (a): wrong β_1 or β_0 value for each model: each worth 0.5 point.
- **– 2 pts** (b): wrong R^2 value
- **– 2 pts** (b): no explanation
- **– 1 pt** (b): incomplete explanation. An explanation is considered complete if it mentions **either** of the following: (1) **both** the relationships between consumption and income, and between experience and income, **or** (2) the relationship between all three variables.
- **– 6 pts** No answer
- **– 3.5 pts** Linear regression was asked not linear SVM

2.4 **(no title)** **15** / 15 pts

✔ **– 0 pts** Correct

– 2 pts (a): wrong β_1 or β_0 value (each 1 point)

– 1 pt (a): wrong β_0 value

– 3 pts (b): computation (1 point), Interpretation (1 point), weak relation (1 point)

– 1 pt (b): wrong value of R^2

– 3 pts (c): correct conclusion with appropriate wordings (reject the null hypothesis, significant relation, etc.)

– 1 pt (c): missing 1 component (rejecting null, or significant relation)

– 2 pts (d): incorrect interpretation (1 point), meaningfully different from 0 (1 point)

– 4 pts (e) spot the contradiction (1 point), correctly explain the reason (2 point), reasonable suggestion (1 point)

– 2 pts (e) incorrect explanation of the reason behind contradiction

– 1 pt (e) reasonable suggestion to mitigate this issue

– 15 pts No answer

– 1 pt The confidence interval is 2*standard error

– 1 pt D) Confidence interval

2.5 **(no title)** **8** / 10 pts

– 0 pts Correct

– 1 pt Wrong R^2 value

– 2 pts Wrong Interval

– 2 pts (a): Correct R^2 value (1 point), explain the meaning of R^2 (1 point). If the student argues that R^2 alone is not enough to make the conclusion, they should also get full score.

– 4 pts (b): correct interval (2 point), interpretation (2 point)

– 4 pts (c) correct conclusion (2 point), explanation no the decision (2 point)

– 10 pts No answer

✔ **– 2 pts** Wrong Interval

– 2 pts (c) no conclusion or wrong conclusion

– 2 pts (c) no explanation

– 2 pts (b) no interpretation

Question 3

Interpretation of Coefficients in Linear Regression **15** / 20 pts

3.1 **(a)** **5** / 5 pts

✔ **– 0 pts** Correct choice and reasoning that the feature does not follow ordinal relationships.

– 2.5 pts Incorrect choice but provided detailed reasoning that partly makes sense.

– 5 pts Incorrect answer of left blank.

– 1.5 pts No explanation

3.2 **(b)** **3** / 5 pts

– 0 pts Correct. Also correct if not using X_2^C (or one of the indicator variable).

– 0.5 pts Missing one term but otherwise correct. For example, missing the intercept term or the term for X_1.

– 1 pt Missing two terms.

✔ **– 2 pts** Missing several terms but the listed terms are correct. For example, missing all interaction terms or missing all terms related to X_1.

– 3 pts Missing several terms and the listed terms are only partly correct.

– 5 pts Incorrect or left blank.

3.3 (c) **7** / 10 pts

– 0 pts Correct. If the interpretations are correct/sensible, full credit can be awarded.

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Partly incorrect interpretation on the intercept terms. (β_0)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1.5 pts Missing/incorrect interpretation on the intercept terms. (β_0)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 0.5 pts Partly incorrect interpretation on the expected change of sales per unit of the fish's weight. ($\beta_1 X_1$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Missing/incorrect interpretation on the expected change of sales per unit of the fish's weight. ($\beta_1 X_1$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Partly incorrect interpretation on the base sale for fish type A term. ($\beta_2 X_2^A$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Missing/incorrect interpretation on the base sale for fish type A term. ($\beta_2 X_2^A$)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 2 pts Partly incorrect interpretation on the interaction terms (e.g. the expected change of sales per unit of the fish's weight with respect to each fish specie: $X_1 X_2^A$).

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

✔ **– 3 pts** Missing/incorrect interpretation on the interaction terms (e.g. the expected change of sales per unit of the fish's weight with respect to each fish specie: $X_1 X_2^A$).

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Missing details of explanation (e.g. only stating β_1 is the coefficient for variable X_1, etc)

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 3 pts Explanation should contain phrases like "compared to type C (or whichever specie that is missing in the equation"

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 10 pts Question unanswered.

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

– 1 pt Missing one interaction term

Note that the β_i mentioned here corresponds to the β_i in the homework solution on BruinLearn)

Question 4

Bias, Variance and Regularization **10** / 10 pts

4.1 **(a)** **5** / 5 pts

✔ **– 0 pts** Correctly identifies which model has highest variance and which has highest bias.

Correctly plots the training error and test error curves. In particular:

(Model 1) model that underfits - (1) it should have the highest training error (2) its test error should be higher than that of model 2.

(Model 2) good model - (1) train error should not touch 0, but both train error and test error should decrease throughout training and be small at the end (2) its training error should be the second highest (3) its test error should be the lowest

(Model 3) model that overfits - (1) train error should be 0 at the end, and making it the lowest among the three models (2) test error should be higher than that of model 2. (3) **for the test error, either a decreasing-increasing pattern or a monotonically decreasing pattern should be considered correct.**

– 2 pts Plotted accuracy instead of error. But the trend is correct if we translate that into plot of error.

– 1 pt Train error != 0 (nor is it close enough) for model that overfits (model 3)

– 1 pt Train error = 0 for good model (model 2)

– 2 pts The relative scale of the training error among the three models is incorrect. It should be model 1 > model 2 > model 3.

– 2 pts The relative scale of the test error among the three models is incorrect. It should be model 1 > model 2, and model 3 > model 2

– 2 pts Train error increases through training

– 2 pts Otherwise incorrect graph shape

– 4 pts No graphs drawn but correctly describes graph behavior

– 5 pts Incorrect/no graphs drawn

4.2 **(b)** **5** / 5 pts

✔ **– 0 pts** Correct.

L1 Regularization. for a)

Encourages coefficients set to 0, hence horizontal line.

– 5 pts Incorrectly identifies which model uses L1 regularization.

– 3 pts No explanation for why L1 regularization.

– 2 pts Insufficient explanation for L1 regularization. Mention to horizontal line but not to coefficients = 0.

✔ **– 0 pts** Correct

– 3 pts wrong answer

Question 5

Logistic Regression **11** / 11 pts

5.1 **(a)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt wrong odds/odds value not mentioned
Correct answer: 1.

– 1 pt wrong probability

5.2 **(b)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt wrong X1 interpretation

– 1 pt Wrong x2 interpretation

– 1 pt Partially correct/Numerical value not mentioned

5.3 **(c)** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt Wrong interpretation on β_0

– 1 pt Wrong interpretation on β_1

– 0 pts Correct

– 1 pt Wrong interpretation either for odds/probability

– 0 pts Correct

– 0 pts Correct

5.4 **(d)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt Wrong equation of boundary/missing equation

– 2 pts Incorrect/Missing answer

5.5 **(e)** **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt Did not mention that coefficients are not unique when with or without multicollinearity

– 1 pt Unreasonable reasons for potential difficulties.

Question assigned to the following page: 1.1

CS M148 HW 1

Arnav Marda

February 11, 2024

1. Data & Bias

(a) (6 points) Your friend working at UCLA dining hall has been given the task of determining how students feel about this year's menus. Your friend wants to complete the task by scraping Reddit for key words related to UCLA food and then run them through a model that can do sentiment analysis. The given model can determine if the text contains positive, negative, or neutral sentiments. Does your friend's data collection method exhibit any selection bias? Explain each kind of bias you give in the context of this situation. (Refer to Week 1 Lecture 2, slide 39 for a list).

Solution: The biases in the above example include:

1. Undercoverage bias - Since only people who are active on Reddit are included, the collection method excludes all the students who are not active on Reddit, or who do not post / did not post on Reddit about UCLA food.

2. Convenience bias - There is evidence of a convenience bias since data is collected from a convenient source such as Reddit which is not representative of the entire population.

3. Response bias - People with extreme / polarizing opinions are more likely to post their opinions on Reddit. Thus, the sample is littered with data that is skewed to the extremes which is not an accurate representation of the entire population of students.

4. Overcoverage bias - Since it is possible that a single student posted multiple times about UCLA food, these students would be counted multiple times in the survey which may lead to inaccurate results.

(b) (6 points) Long since 2018, companies have started to explore the potential of AI as the recruiter for their hiring process, but AI recruiters were faced with many problems at that time and the idea was eventually scrapped due to bias issues, especially against women.

Question assigned to the following page: 1.2

i. Explain why the tool was discriminating against women?

Solution: Since the companies' workforces consisted of a majority of men and received more male applicants, the AI that was trained on the resumes and information of such applicants was biased towards men. Due to the bias in the data provided, the AI taught itself to prefer male candidates and reject women candidates leading to the discrimination. Thus, this was due to a bias in the data provided to the AI model. This could be classified as a case of undercoverage bias as women were improperly represented in the sample data.

ii. The developer decides to drop the gender in their data. Would this eliminate the bias? Why?

Solution: This would not be enough to eliminate the bias since it is possible that the AI would latch onto keywords in the resume e.g. "Women's chess club" or "Men's water polo", etc. This would lead to similar discrimination based on gender. If such keywords were also dropped from the resume, then it is still possible that the AI uses other keywords to discriminate against a particular part of the population (maybe not women).

Question assigned to the following page: 2.1

2. Linear Regression: goodness of fit & Interpretation

(a) (6 points) US population was around 9 million in 1820, 40 million in 1870, 92 million in 1910, 151 million in 1950, and 281 million in 2000.

i. The closed-form solution of linear regression with an MSE loss is $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$. Use the formula to fit the above data. What will the population be like in 2010 under this model?

Solution: We can start by calculating $\overline{x}$ and $\overline{y}$.

$$\overline{x} = \frac{1820 + 1870 + 1910 + 1950 + 2000}{5} = 1910$$
$$\overline{y} = \frac{9 + 40 + 92 + 151 + 281}{5} = 114.6$$

Using this, we can find $\hat{\beta}_1$ and $\hat{\beta}_0$,

$$\hat{\beta}_1 = 1.491$$
$$\hat{\beta}_0 = 114.6 - 1.491 \times 1910 = -2732.678$$

I omit the calculations above for simplicity. Thus, we have that

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}$$

For $\hat{x} = 2010$,

$$\hat{y} = -2732.678 + 1.491 \times 2010 = 263.672$$

The population in 2010 is predicted to be 263.672 million.

ii. What is R2 for your model? Based on the value of R2 can we say whether the estimated regression line fits the data well?

Solution: Using the formula for R^2 given in lecture,

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\overline{y} - y_i)^2} = 0.932$$

Since, the R^2 value is high, the estimated regression line fits the data well.

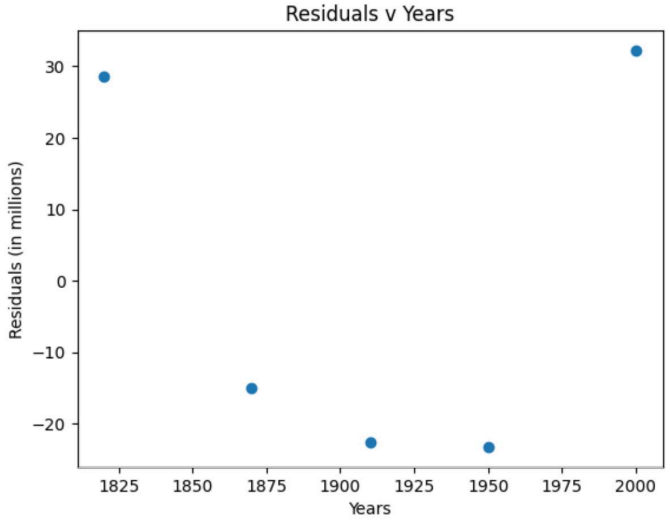
iii. Plot the residuals versus year. Do you think this is a good model? Why?

Solution: To plot the residuals v year, we need to first calculate the residuals. The residual for any given year x_i can be calculated using:

$$y_{res} = y_i - \hat{y}_i$$

Questions assigned to the following page: 2.1 and 2.2

Below I present the graph of residuals v year plotted using Python (matplotlib). Note that calculations were done by hand but due to simplicity, graph was plotted using matplotlib.



Using the graph above, I do not believe that this is a good model. This is due to the large discrepancies as shown with the residuals. This could be because of the small data set. Since the R^2 value is high, it is clear that the linear model fits the data well so the one of the reasons for the high residuals could be the small data set.

(b) (4 points) The following plot shows how the number of deaths due to heart disease varies with wine consumption, in different countries. Is there a strong correlation between heart disease and wine consumption? Can we conclude that drinking more wine will reduce the risk of heart disease? Explain your reasoning. (Note: The figure in the question has not been shown here).

Solution: The R^2 value reported of 0.71 This suggests a good (not exceptional) fit and a strong correlation between heart disease and wine consumption. Moreover, if we evaluate the model itself, it suggests a negative correlation between heart disease and wine consumption i.e. higher wine consumption leads to lower heart disease. The conclusion in the question stands based on the model itself. This can be interpreted from the negative slope provided in the regression model. However, there could be many other factors affecting heart disease such as genetic conditions. Furthermore, it is also possible that people with predisposed heart conditions would drink less wine which would skew the data. This is just one example of factors that could lead the data askew.

(c) (6 points) [You can use Python] The Income Data contains data from 14 individ-

Question assigned to the following page: 2.3

uals. The first column shows the average income per year (Income). the second column shows the average spending per year (Consumption), and the third column shows the number of years of working experience (Experience),

i. Report β_0, β_1 for two linear classifiers that model: (i) consumption based on income, and (ii) income based on working experience.

> **Solution:** The parameters for the 2 linear classifiers are:
>
> (i) $\beta_0 = 0.620$, $\beta_1 = 4.271$
>
> (ii) $\beta_0 = 35.40$, $\beta_1 = 7.58$

ii. Report R^2 for the above classifiers and explain the relationships between consumption, working experience, and income. Analyze the potential reason behind this.

> **Solution:** The R^2 values are as shown below:
>
> (i) $R^2 = 0.582$. This suggests a moderate fit for the model. The coefficients calculated above suggest a positive relationship between consumption and income. For every unit increase in Income, Consumption increases by about 4.271. Moreover, the intercept shows that even small values of Income lead to positive Consumption. This is intuitive since greater income is likely related to greater disposable income and ergo greater Consumption. Thus, the model makes good intuitive sense.
>
> (ii) $R^2 = 0.602$. Again, this suggests a moderate fit for the model. The slope calculated above suggests a positive relationship between Income and Experience. For every year increase in Experience, Income increases by 7.58 units. And, the intercept shows that small values of Experience also lead to at least 35.4 units of Income. This makes sense since greater experience implies that you have been part of the workforce longer i.e. you have likely received pay raises and/or promotions that would lead to an increased income. Moreover, the intercept also makes sense since you get a base positive income even with 0 experience.
>
> The above suggests a chain of relations between the variables in the sense that greater experience leads to greater income which in turn leads to greater consumption.

(d) (15 points) You can use Python]. The Experiment dataset containing a thousand (x, y) data points, from a scientific experiment.

i. Fit a linear model to the data and compute β_0, β_1.

Question assigned to the following page: 2.4

Solution: $\beta_0 = 4.529, \beta_1 = 0.099$

ii. Is there a strong linear relation between x and y? Explain your reasoning.

Solution: I do not think that there is a strong linear relation between x and y. This is justified with a low correlation coefficient of 0.459. Moreover, the R^2 value for the linear model is also low ≈ 0.211.

iii. Conduct the test $H_0 : \beta_1 = 0$(reject the null hypothesis if the p-value for β_1 is less than 0.05). Analyze your result.

Solution: The p-value for β_1 is about $2.85 \times 10^{-53} < 0.05$. Thus, we reject the null hypothesis. This implies that there is a relationship between x and y.

iv. Calculate a 95% confidence interval for β_1, using $\beta_1 \pm 2 \times SE(\beta_1)$, and interpret your interval. Suppose that if $\beta_1 \geq 1$, then we consider it to be meaningfully different from 0, in our research. Does the 95% confidence interval suggest that β_1 is meaningfully different from 0?

Solution: We have that $SE(\beta_1) = 0.006$. Thus, we can calculate a 95% confidence interval as follows:

$$95\%Interval : (0.099 - 2 \times 0.006, 0.099 + 2 \times 0.006) = (0.087, 0.111)$$

From the 95% confidence interval, we can conclude that β_1 is not meaningfully different from 0.

v. Summarize the contradiction you've observed in parts (c) and (d). What is causing the contradiction, and what would you recommend we should always do while analyzing data?

Solution: Part (c) suggests that there exists a relationship between x and y since we rejected the null hypothesis. Part (d) implies that there is no meaningful relationship between x and y since β_1 is not sufficiently different from 0. This contradicts part (c).

One of the reasons of the contradiction could be that the residuals are not independent of x i.e. the residuals increase as x increases. Another reason could be the noise present in the data which leads to irreducible errors.

It is important to always look for and clean noise in the data before modelling it. Moreover, evaluate the magnitude and practical significance of the relationship between x and y, in addition to statistical significance. Even

Questions assigned to the following page: 2.4 and 2.5

if the null hypothesis is rejected, a small effect size may not have practical implications. Furthermore, instead of relying solely on hypothesis testing or confidence intervals, interpret results in a comprehensive manner, considering both statistical evidence and practical implications.

(e) (10 points) [You can use Python] The Volcano dataset contains 21 consecutive volcanic eruptions. Use a linear model to predict the time until the next eruption (next), given the duration of the last eruption (duration).

i. Is the linear model a good model? Analyze your result using R^2.

Solution: After fitting the model, we got a R^2-value of 0.749. This implies that the model was a good fit and so a good model.

ii. If the duration of the last eruption was 5 minutes, obtain a 95% prediction interval for the time until the next eruption occurs, and interpret your prediction interval.

Solution: The 95% prediction interval for the time until the next eruption occurs is $(74.412, 85.515)$. This implies that if the duration of the last eruption was 5 minutes, the model can predict with 95% certainty that the next eruption will occur between 74.412 and 85.515 minutes (or units).

iii. If you need to leave in 50 minutes, can you determine if you can see the eruption based on the data? Explain your reasoning.

Solution: It is possible to determine if you can see the eruption based on the data provided that you know the duration of the last eruption. If you know the duration of the last eruption, you can use the model to predict how long it will take until the next eruption (at least with 95% confidence). Using this, you can deduce if it will happen in the next 50 minutes. For example, we can use the prediction from (b). If the last eruption was 5 minutes long, then with 95% confidence we can say that, you will not see the eruption if you have to leave in 50 minutes. Of course this also depends on how long it has been since the last eruption. Using the same example, if it has been anything more that 24.412 minutes since the last eruption (5 minutes long), then you would be able to see the next eruption even if you have to leave in 50 minutes. Thus, this would require you to have a little more data than just the data provided in the dataset.

Questions assigned to the following page: 3.1 and 3.2

3. Interpretation of Coefficients in Linear Regression

Suppose that we want to model the market sales of fish in a fish market on the weight of three different species of fish. Moreover, we are expecting a linear growth-response over a given range of weight with the sales. Hence, we want to model the outcome Y (sales) as a linear function of the weight X_1 and the fish specie X_2. There is **no** ordinal relationship between the fish species.

(a) (5 points) As X_2 is a categorical feature, we need to first convert it through encoding. Which of the following encoding will be more preferable? Explain your reasoning.

(1) Create one variable $X_2 = \{1, 2, 3\}$. Specifically, let $X_2 = 1$ if fish species is A, $X_2 = 2$ if fish species is B, and $X_2 = 3$ if fish species is C.

(2) Create three indicator variables X_A^2, X_B^2 and X_C^2. Specifically, let $X_A^2 = 1$ if fish species is A and 0 otherwise. X_B^2 and X_C^2 are encoded similarly.

Solution: I think the encoding described in option 2 i.e. a one-hot encoding mechanism would be more preferable. This is because option 1 assumes an ordinal relationship between the categories however it is specified at the top of the question that there is no such ordinal relationship. On the other hand, option 2 assumes that each category is independent. This is more suitable for the question at hand since it assumes no natural order among the categories.

(b) (5 points) Based on the encoding you chose, how do you model the weight of the fish on the sales of different fish species? Hint. Use $\beta_1, \beta_1, \ldots$ to denote the coefficients and write the model in the form of $Y = \beta X + \ldots + \epsilon$.

Solution: Consider the following coefficients:

- β_0 - The constant term or the intercept value for the model.
- β_1 - The coefficient of weight(X_1).
- $\beta_2^A, \beta_2^B, \beta_2^C$ are the coefficients of the indicator variables X_A^2, X_B^2 and X_C^2 respectively.
- ϵ - The error term in the model.

Thus, we get the final model as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2^A X_A^2 + \beta_2^B X_B^2 + \beta_2^C X_C^2 + \epsilon$$

(c) (10 points) How do you interpret each coefficients in your model? Your answer should include interaction terms (for example, $\beta X_i X_j$). Hint. When doing interpretation, try to discuss by cases. For example, when the fish species is A/B/C.

Question assigned to the following page: 3.3

Solution: For the model given above, we can also show it as follows:

$$Y = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2^A + \epsilon; \text{For species A} \\ \beta_0 + \beta_1 X_1 + \beta_2^B + \epsilon; \text{For species B} \\ \beta_0 + \beta_1 X_1 + \beta_2^C + \epsilon; \text{For species C} \end{cases}$$

Thus, consider the following cases:

A. When the fish species is A, the average sales for a fish will be $\beta_0 + \beta_2^A$. Moreover, for every unit increase in weight, the sales will increase by β_1. Note that this is common for all the cases and so will not be mentioned hereafter. Also, $\beta_2^A - \beta_2^B$ is the difference in sales between species A and species B. And, $\beta_2^A - \beta_2^C$ is the difference in sales between species A and species C.

B. When the fish species is B, the average sales for a fish will be $\beta_0 + \beta_2^B$. Also, $\beta_2^B - \beta_2^A$ is the difference in sales between species A and species B. And, $\beta_2^B - \beta_2^C$ is the difference in sales between species B and species C.

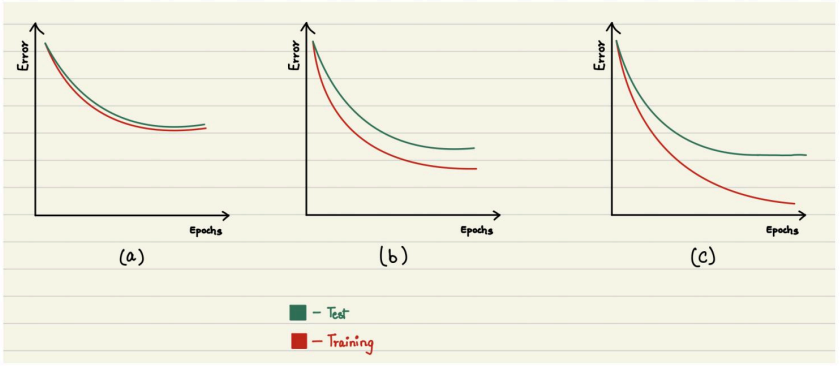
C. When the fish species is C, the average sales for a fish will be $\beta_0 + \beta_2^C$. Also, $\beta_2^C - \beta_2^A$ is the difference in sales between species A and species C. And, $\beta_2^C - \beta_2^B$ is the difference in sales between species B and species C.

Questions assigned to the following page: 4.1 and 4.2

4. Bias, Variance and Regularization

(a) (5 points) The figures below show the decision boundary of three different classifiers. In which of the figures below does the classifier have a larger bias and in which figure does the classifier has a larger variance? Draw out an approximate graph where you demonstrate how the training and test error for each classifier will change over time during the course of training the model, and explain how do you expect each classifier to perform (accuracy) on the test set.
Note: You don't need to calculate the errors, but you should show how the training and test errors compare for different classifiers.
(Image not shown)

Solution: Let the images be labelled (a), (b) and (c) from left to right.
Then we have that figure (a) has the greatest bias while figure (c) has the greatest variance.
The graphs for how the training and test errors will change over time for each of the classifiers is shown below:



(b) (5 points) One strategy to reduce variance and improve generalization is regularization. In figure below, the blue lines are the logistic regression without regularization and the black lines are logistic regression with L1 or L2 regularization. In which figure L1 regularization is used and why?
(Image not shown)

Solution: Figure (a) corresponds to the L1 regularization since the decision boundary shows that the coefficients tend to be zero. For Figure (b) , both coefficients are close to zero, as a result, it may give you a larger slope in the decision boundary.

Questions assigned to the following page: 5.2, 5.3, and 5.1

5. Logistic Regression

Suppose we fit a multiple logistic regression: $\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

(a) (2 points) Suppose we have $p = 2$, and $\beta_0 = 3, \beta_1 = 5, \beta_2 = -8$. When $X_1 = X_2 = 1$, what are the odds and probability of the event that $Y = 1$?

Solution: Thus, we have

$$\begin{aligned}
\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) &= 3 + 5 - 8 \\
\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) &= 0 \\
\frac{P(Y=1)}{1-P(Y=1)} &= e^0 \\
P(Y=1) \times (1 + e^0) &= e^0 \\
P(Y=1) &= \frac{e^0}{1+e^0} \\
&= 0.5
\end{aligned}$$

(b) (2 points) Suppose we increase the X_1 value by 1, how does it change the log odds and odds of the event that $Y = 1$? What if instead, we decrease the X_2 value by 1?

Solution: The formula for $P(Y = 1)$ can be decomposed into:

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

Increasing the value of X_1 by 1, will increase the log odds by 5 and the odds of the event that $Y = 1$ by a factor of e^5.

Similarly, decreasing the value of X_2 by 1 will increase the log odds by 8 and the odds of the event that $Y = 1$ by a factor of e^8.

(c) (2 points) Explain how increasing or decreasing $\beta_0, \beta_1, \beta_2$ affect our predictions.

Solution: Increasing or decreasing β_0 will shift the intercept of the decision boundary up or down respectively.

Increasing β_1 will increase the effect of X_1 on the log odds of the event $Y = 1$, while decreasing β_1 will have the opposite effect.

Increasing or decreasing β_2 will have a similar effect on X_2 that β_1 will have on X_1.

Questions assigned to the following page: 5.5 and 5.4

(d) (2 points) What is the formulation of the decision boundary? Which points are on the decision boundary?

Solution: The decision boundary is the line where the log odds of the event $Y = 1$ are equal to 0. It is the solution to the equation:

$$\begin{aligned} \beta_0 + \beta_1 X_1 + \beta_2 X_2 &= 0 \\ 3 + 5X_1 - 8X_2 &= 0 \\ 8X_2 - 5X_1 &= 3 \end{aligned}$$

Thus, the points on the decision boundary are all the points that satisfy $8X_2 - 5X_1 = 3$.

If we were to generalize this, we can say that the points on the decision boundary are all points that satisfy:

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0$$

(e) (2 points) Suppose we fit another two logistic regression models: one with only X_1 and the other one with only X_2, and we observe that the coefficients of X_1 and X_2 in the two models are different than those specified in part (a). Explain what is the potential reason and why it could be problematic that the coefficients are different than those specified in part (a).

Solution: It is possible that the coefficients of X_1 and X_2 are different because one of the predictors could be a better or worse fit to model the data. Moreover, it is also possible that models with only one of the predictors lack interaction terms and complexity. The difference could also be due to multicollinearity.

If the coefficients differ significantly, it may indicate that the logistic regression model does not accurately capture the relationship between predictors and the response variable in the specific context described in part (a). This could lead to biased predictions and unreliable inference when applying the model to new data or making decisions based on its results.