Midterm

Graded

Student

ARNAV KRISHNAKUMAR MARDA

Total Points

92 / 100 pts

Question 1

Short Answers **17** / 17 pts

1.1 **Data collection & Bias. (a)** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Partially correct

1.2 **Data collection & Bias. (b)** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Partially correct

1.3 **Data collection & Bias. (c)** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Partially correct

1.4 **Data collection & Bias. (d)** **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect

– 1 pt Partially correct

1.5 **KNN. (a)** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt answer is not no

– 1 pt wrong or missing explanation for the answer no

– 1 pt wrong or missing answer for how to improve the model

1.6 **KNN. (b)** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt answer is not no

– 2 pts wrong explanation

1.7 **KNN. (c)** **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts Incorrect

Question 2

Linear Regression **22** / 23 pts

2.1 **(a)** **6** / 6 pts

✔ **– 0 pts** Correct

– 6 pts Unanswered

– 1 pt No intercept

– 1 pt Wrong variable(s) or missing indicator variable(s)

– 1 pt Missing single-variable term(s)

– 2 pts No interaction terms at all

– 1 pt Missing interaction term(s)

– 1 pt Wrong interaction term (should be multiplication)

– 1 pt One indicator variable for each of the three conditions, i.e., having x_2', x_2'', x_2'''.

– 1 pt No mentioning of what each variable captures and if it is binary or real valued

– 1 pt Incorrect coefficient(s)

– 1 pt Redundant term(s)

– 1 pt 1 extra term

– 2 pts 2 extra terms

– 3 pts 3 extra terms

– 4 pts 4 extra terms

– 5 pts 5 extra terms

– 6 pts 6 extra terms

2.2 **(b)** **11** / 12 pts

– **0 pts** Correct

– **1 pt** Intercept coefficient: not specific or partially correct

– **2 pts** Intercept coefficient: incorrect or missing interpretation. (e.g. only stating it's the intercept is not seen as an interpretation.)

– **2 pts** non-interaction coefficient: wrong interpretation for 1 coefficients

– **4 pts** non-interaction coefficient: wrong interpretation for 2 coefficients

– **6 pts** non-interaction coefficient: wrong interpretation for 3 coefficients

– **2 pts** Interaction coefficient: wrong interpretation for 1 coefficient

– **4 pts** Interaction coefficient: wrong interpretation for 2 coefficients

✔ – **1 pt** Special: Interpretation otherwise correct, but did not specify **average shoot length** in interpretation of intercept and/or non-interaction coefficient(s)

– **1 pt** Special: The non-interaction coefficient for x_1 (sunlight) measures the change in mean shoot length for a 1 unit increase in sunlight **for the control group** (or whichever nitrogen group was chosen to have $x_2 = 0$). Only applies to 2 indicator variables.

– **0 pts** Marker: used 3 indicator variables for part a

– **12 pts** Unanswered or completely incorrect

2.3 **(c)** **5** / 5 pts

✔ – **0 pts** Correct

– **0 pts** Marker: using the "three binary variables and 8 betas" standard

– **0 pts** Marker: using the "three binary variables and 6 betas" standard

– **1 pt** wrong answer for 1 coefficient

– **2 pts** wrong answer for 2 coefficients

– **3 pts** wrong answer for 3 coefficients

– **4 pts** wrong answer for 4 coefficients

– **5 pts** wrong answer for 5 coefficients

Question 3

Model Selection & Bias-Variance Trade-off **15** / 16 pts

3.1 **(a)** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts No / Incorrect / Insufficient explanation. (explanations only referring to distributions without discussing what this means fall in this category)

– 1 pt Partially correct explanation but no reference to model complexity or overfitting / underfitting or bias / variance or regularization term dominates in optimization

– 1 pt Incorrect k for 1 model

– 2 pts Incorrect k for 2 models

– 3 pts incorrect k for 3 models

– 4 pts Interprets distribution as distribution of coefficients instead of distribution of residuals

– 4 pts No answer/ wrong k values and wrong explanation

3.2 **(b)** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts No / Incorrect / Insufficient Explanation (Answers only referring to how the residuals are distributed with no interpretation of what this means, fall here)

– 1 pt Incorrect model with highest variance

– 1 pt Incorrect model with highest bias

– 4 pts Incorrectly interpretation of distribution as distribution of coefficients ($\beta_1, \ldots, \beta_k$) instead of distribution of residuals

– 1 pt Partially correct explanation but missing clear reference to model complexity / expressivity or underfitting / overfitting etc.

– 4 pts No answer

3.3 **(c)** **4** / 4 pts

✔ **– 0 pts** Correct

– 1 pt Correct choice but insufficient / partially incorrect explanation

– 2 pts Correct choice but no / incorrect explanation

– 2 pts Incorrect choice

– 2 pts No mention of issues with other models

– 1 pt Incorrect issues with other models

– 4 pts No answer

3.4 (d) 3 / 4 pts

– 0 pts Correct. As long as "$\beta_i = 0$ for all $i > 2$" is mentioned.

– 1 pt saying $\beta_1 = 0$ or $\beta_2 = 0$

✔ – 1 pt Only saying $\beta_i, i > 2$'s are "small" or "approximately 0" instead of saying that they are exactly 0. Note that a big differentiating factor of L1/L2 is the fact that L1 makes values exactly 0 (feature selection).

– 3 pts Recognizes that regularization causes coefficients to be minimized, but no other remarks

– 2 pts Mentions that L1 regularization causes most coefficients will be zero, but no specifics or incorrect specifics

– 1 pt Mentions that higher order terms will be pushed to zero, but no or incorrect specifics

– 3 pts Recognizes that the ideal k is 2 (quadratic), but no other remarks

– 4 pts Incorrect or no answer

Question 4

Logistic Regression & Decision Boundary **22** / 26 pts

4.1 **(a)** **4** / 4 pts

✔ **– 0 pts** Correct

– 0.5 pts Minor mistakes

– 2 pts Modeling is partly incorrect. For example, missing one term.

– 0.5 pts Answer missing the $\ln$ term (i.e. only $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ is provided.)

– 3.5 pts Answers only include $\ln(\frac{P(Y=1)}{P(Y=0)})$ without any modeling (i.e. missing $\beta_0 + \beta_1 X_1$).

– 4 pts Incorrect or unanswered.

4.2 **(b)** **8** / 8 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect interpretation for β_0

– 2 pts Incorrect interpretation for β_1

– 2 pts Incorrect interpretation for β_2

– 2 pts Incorrect interpretation for β_3

– 8 pts Unanswered or incorrect.

– 2 pts No mention of hours practiced.

4.3 **(c)** **4** / 4 pts

✔ **– 0 pts** Correct.

– 4 pts Unanswered or completely wrong.

– 3 pts the correct expression is not provided

– 1 pt other minor mistakes

4.4 **(d)** **0** / 4 pts

✔ **+ 0 pts** Unanswered or incorrect.

+ 2 pts We can only say that **we're not confident** or **cannot conclude**. Note that failing to reject the null hypothesis does not mean that the null hypothesis is definitely true.

+ 2 pts Pointed out that 0 is in the interval.

4.5 (e) **6** / 6 pts

✔ **– 0 pts** Correct

– 0.5 pts Minor mistakes on the coefficients, but formulations are correct.

– 1 pt Incomplete or partly incorrect decision boundary, e.g. missing "=0"

– 2 pts Incorrect or missing decision boundary

– 2 pts Incorrect or unanswered result for smoker

– 2 pts Incorrect or unanswered result for non-smoker

– 6 pts Unanswered or incorrect

– 1 pt Miscalculated result for smoker

– 1 pt Miscalculated result for non-smoker

Question 5

Classification Metrics **16** / 18 pts

5.1 (a) **8** / 8 pts

✔ **– 0 pts** Correct

– 2 pts F1 wrong

– 2 pts Accuracy wrong

– 2 pts Recall wrong

– 2 pts Precision wrong

+ 2 pts Partial credits for reasonable attempts

– 8 pts Unanswered

– 1 pt F1 miscalculated

– 1 pt Accuracy miscalculated

– 1 pt Precision miscalculated

5.2 (b) **2** / 4 pts

– 0 pts Correct

– 2 pts Did not or wrongly explain the discrepancy between accuracy and f1 score.

✔ **– 2 pts** Use the wrong metric for evaluation.

– 1 pt didn't explicitly mention imbalanced data

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

✔ **– 0 pts** Correct

– 3 pts Wrong answer

5.4 (d) **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts wrong

– 1.5 pts mentioned changing the threshold, but didn't explicitly say to increase it

Data Science (CS 148)
February 14, 2024

Midterm Exam

Duration: 90 minutes

Write your name and UID: Arnav Marda (405772661)

Note 1: Please only write in the corresponding box for each question.

Note 2: If you need scratch paper or more space for a question, use back of the last page.

Note 3: If you find a question difficult, move on with the rest of the questions and come back to it in the end!

1 Short Answers (17 points)

Data collection & Bias. What kind of biases can the following data collection methods introduce? *Note: For each item, specify 'one' type of bias that is most relevant/prominent.*

(a) **(2 points)** Collecting data by asking the following question: "This is an easy concept, don't you understand?"

Response bias

(b) **(2 points)** Collecting data by asking a question from all your (relatively large pool of) Facebook friends.

Convenience bias

(c) **(2 points)** Collecting data about addiction by sending a non-anonymous survey to a random sample of people, including addicts.

Non-response bias

(d) **(2 points)** Posting an advertisement and collecting data from people who reach out to you.

Voluntary bias

KNN. For all questions below, please provide a short justification along with the answer:

(a) **(3 points)** If the scale of the predictors is very different, do you expect a KNN model to perform well? Explain briefly. If your answer is no, what do you do to improve the model's performance?

No. ~~In order~~ If the scale of the predictors is very different, the model will be inaccurate as you won't be able to pick the k neighbours accurately. To improve the model, scale all the predictors including binary predictors (maybe to values in [0,1]).

(b) **(3 points)** If your data has many predictors, do you expect a KNN model to work well? Explain briefly.

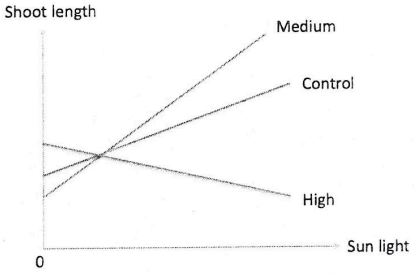
No, this is because the extra predictors would make it harder to accurately find the k-nearest nighbours and it would also increas variance.

(c) **(3 Points)** How do you expect the value of K in a KNN model to impact the variance of your model?

I would expect variance to decrease as the value of k increases in the model.

2 Linear Regression (23 points)

Suppose we measured the shoot length (in cm) of some plant species based on the amount of sunlight and a nutrient treatment with three levels of nitrogen (Control, Medium, High nitrogen concentration). The following plot shows the shoot length vs the amount of sunlight for different nutrient treatments.



$X_2 = 1$ implies medium treatment,
$X_3 = 1$ implies high,
and $X_2 = 0, X_3 = 0$ implies control.

$X_3 \rightarrow$ High treatment (binary)

Figure 1: Shoot length vs sunlight for groups with different levels of nitrogen treatment.

(a) **(6 points)** How do you model shoot length (Y) based on sunlight (X_1) *and* treatment (X_2), using *one interpretable* linear regression model with minimum number of predictors? Write the formulation of your linear regression model in 1 line. *Note: mention what each variable captures and if it is binary or real valued.*

~~$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$~~ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$

~~$X_1 \rightarrow$ Real valued variable for sunlight~~ $X_1 \rightarrow$ Sunlight (real valued)
$X_2 \rightarrow$ medium treatment (binary)

$X_2 \rightarrow$ Real valued (categorical) variable for treatment where 0 = control, 1 = medium, 2 = high. We do not need multiple predictors since they have an ordinal relationship.

(b) **(12 points)** Write the interpretation of each of the coefficients in your model.

~~If $X_2 = 0$, i.e. control treatment~~

Control treatment:

$Y = \beta_0 + \beta_1 X_1$

$\beta_0 \rightarrow$ base shoot length for control treatment

$\beta_1 \rightarrow$ amt. of change in shoot length for 1 unit change in sun light for control treatment

Medium treatment:

$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$

$\beta_0 + \beta_2 \rightarrow$ base shoot length for medium treatment.

$\beta_1 + \beta_4 \rightarrow$ amt. of change in shoot length for 1 unit change in sun light for medium treatment.

High treatment:

$Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$

$\beta_0 + \beta_3 \rightarrow$ base shoot length for high treatment.

$\beta_1 + \beta_5 \rightarrow$ amt. of change in shoot length for 1 unit change in sunlight for high treatment.

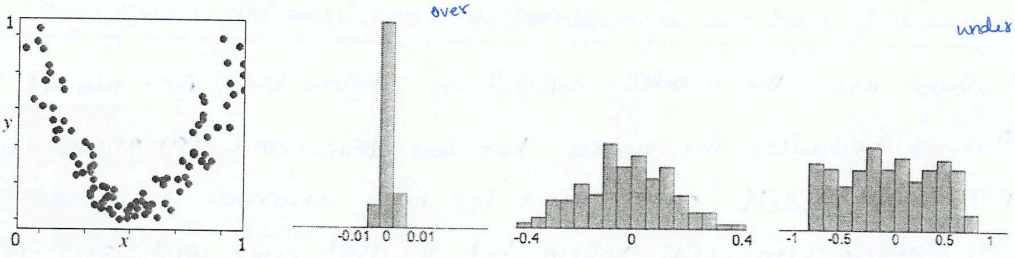
(c) **(5 points)** Write the sign of the coefficients of your model based on Figure 1.

$\beta_0 \rightarrow$ positive $\quad \beta_2 \rightarrow$ negative $\quad \beta_4 \rightarrow$ positive

$\beta_1 \rightarrow$ positive $\quad \beta_3 \rightarrow$ positive $\quad \beta_5 \rightarrow$ negative

3 Model Selection & Bias-Variance Trade-off (16 points)

We use a linear regression with higher order terms to model the data in Fig. 2a. Formally, the model is: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$, and we use a MSE loss to traing the model.



(a) Data (b) Model 1 (c) Model 2 (d) Model 3

Figure 2: (a) Data, (b), (c), (d) Distribution of residuals for three models with different values of k.

(a) **(4 points)** Fig. 2b, 2c, 2d show the distribution of residuals for 3 different models with different values of k. Which value of k is used in each model (Fig. 2b, or 2c, or 2d)? *Hint: You can either chose a number for k or specify a range.*

For (2b), k would be greater than or equal to 3.

For (2c), k would be about 2.

For (2d), k would be 1.

(b) **(4 points)** Which model (Fig. 2b, or 2c, or 2d) has the largest variance and which one has the largest bias? Explain briefly.

Figure (2b) would have the largest variance due to the overfitting of the model. The residuals are all centred around 0 so the model has low test error due to the overfitting of the model on test data leading to greater variance.

Figure (2d) has the largest bias. This can be seen from the plot of the residuals as there is a uniform distribution around 0. This implies that there was underfitting of data and thus ~~the most~~ highly inaccurate results.

(c) **(4 points)** Which model would you use and what is the issue with those that you don't chose?

I would use the model dipicted in Figure (2c). i.e. model 2. This is because the model has low bias and variance. I would not use Model1 since ~~the~~ it has high variance so as we train, the training error will reduce, but the test error will still remain high due to overfitting. I would not choose Model 3 since it has high bias so the training and test error will not reduce that much with additional training leading to an inaccurate model due to underfitting.

(d) **(4 points)** Assume we use $k = 10$ and use L_1 regularization, to train the model with the following loss function: $L_{reg}(\beta) = L_{MSE}(\beta) + \lambda R(\beta)$ where $R = \sum_{i=1}^{k} |\beta_i|$. After model selection to decide about best value of λ, what do you expect the value of the coefficients (i.e., β_is) be? *Hint: You can either chose a number for each β_i or specify a range.*

I would expect $\beta_i \approx 0$ for $i = 3, 4, \ldots, 10$. I would expect β_0, β_1 to also be low about $|\beta_0|, |\beta_1| \in [0, 2]$. I would expect β_2 to also be in such a range about ~~$|\beta_2| \in [0, 5]$~~. $|\beta_2| \in [0, 3]$.

4 Logistic Regression & Decision Boundary (26)

We want to use Logistic regression to model the probability for an individual to finish a 10K Marathon ($Y = 1$), based on whether the individual is smoking ($X_1 = 1$) or not smoking ($X_1 = 0$) and the number of hours they practiced (X_2).

(a) **(4 points)** Write the logistic regression formulation to model log (use ln) odds of finishing the marathon, based on whether the individual is smoking or not. Include an interaction term between smoking and hours of practice.

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

(b) **(8 points)** Interpret the coefficients of your model. *Note: write your answer based on* e^{β_i}.

If the individual is smoking,

$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2$

$\beta_0 + \beta_1 \rightarrow$ base log odds for $P(Y=1)$ i.e. when $X_2 = 0$. base $P(Y=1)$ would be $e^{\beta_0 + \beta_1}$.

$(\beta_2 + \beta_3) \rightarrow$ a 1 unit increase in X_2 would increase the log odds by $\beta_2 + \beta_3$ and would increase the odds by a factor of $e^{\beta_2 + \beta_3}$. If X_2 was decreased, we would have a similar decrease in odds.

If the individual does not smoke,

$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_2 X_2$

$\beta_0 \rightarrow$ base log odds for $P(Y=1)$ i.e when $X_2 = 0$. Base odds for $P(Y=1)$ would be e^{β_0}

$\beta_2 \rightarrow$ a 1 unit increase in X_2 would increase the log odds by β_2 and would increase the odds by a factor of e^{β_2}. If X_2 was decrease, we would have a similar decrease in odds.

(c) **(4 points)** Estimate the odds ratio of finishing the marathon comparing smokers to non-smokers using this model. Write in 1 sentence how do you interpret the odd ratios?

The odds ratio would be $e^{\beta_1 + \beta_3 X_2}$. Thus, individuals who smoke are $e^{\beta_1 + \beta_3 X_2}$ times more likely to finish the race (note that β_3 can be negative). We compute this ratio by holding the value of X_2 constant.

(d) **(4 points)** If the 95% confidence interval for β_1 is $[-2.8, 0.2]$, interpret the effect of smoking on the probability of finishing the race.

Thus, smoking most likely reduces the odds for an individual to finish the race. This is because $\beta_1 \in [-2.8, 0.2]$ with 95% confidence. It is possible for it to increase the odds if ~~$\beta_1 \in [$~~ $\beta_1 \in (0, 0.2]$.

(e) **(6 points)** If $\beta_0 = -4, \beta_1 = -8, \beta_2 = 0.2, \beta_3 = 0.1$, write the formulation for the decision boundary. How many hours a smoker and a non-smoker individual need to practice to finish the race?

Decision boundary: ~~-4~~ $-4 - 8X_1 + 0.2X_2 + 0.1X_1X_2 = 0$.

For smokers: $-4 - 8 + 0.2X_2 + 0.1X_2 = 0 \Rightarrow 0.3X_2 = 12 \Rightarrow X_2 = 40$.
Smokers would have to practice $\geq$ 40 hrs.

For non-smokers: $-4 + 0.2X_2 = 0 \Rightarrow X_2 = \frac{4}{0.2} \Rightarrow 20$

Non-smokers would have to practice $\geq$ 20 hrs.

5 Classification Metrics (18 points)

We train a binary classifier to predict if a patient has cancer ($Y = 1$). We use the output probability of the classifier $P(Y = 1|X)$ to predict $Y = 1$ for a patient X, if $P(Y = 1|X) \geq 0.5$, and we predict $Y = 0$ otherwise. The classifier has following confusion matrix.

	Predicted $\hat{Y} = 1$	Predicted $\hat{Y} = 0$
Actual $Y = 1$	4	96
Actual $Y = 0$	4	9896

(a) **(8 points)** Calculate the **Accuracy, Precision, Recall** and **F1 Score** and of the classifier.

Precision $= \frac{4}{4+96} = 0.04$.

Recall $= \frac{4}{4+4} = 0.5$

Accuracy $= \frac{9900}{10000} = 0.99$

F1-score $= \frac{2 \times 0.5 \times 0.04}{0.5 + 0.04} = \frac{0.04}{0.54} = \frac{4}{54} = \frac{2}{27}$

(b) **(4 points)** Explain the discrepancy between accuracy and F1 score. Based on the above confusion matrix, which metric is better to evaluate the performance of the classifier?

There is vast discrepancy between accuracy and F1-score because there are extremely few cases for the $Y=1$ scenario. This skews the precision and recall which skews the F1-score. Based on the confusion matrix, accuracy is the better metric.

(c) **(3 points)** How do you use the output probabilities of the above classifier to predict $Y = 1$ for a cancer screening test, where it is most important to identify *all* the possible cancer patients?

In this case, you would ~~say that~~ predict $Y=1$ for a patient X if $P(Y=1|X) \geq 0.4$.

(d) **(3 points)** How do you use the output probabilities of the above classifier to predict $Y = 1$ only if the patient has cancer (i.e., to minimize the number of patients who are flagged by mistake)?

In this case, you would predict $Y=1$ for patient X if $P(Y=1|X) \geq 0.6$.