Midterm

● Graded

Student

TIYA CHOKHANI

Total Points

95 / 100 pts

Question 1

Q1.1a

2 / 2 pts

✔ **– 0 pts** Correct: A/B/C all fine, ambiguous question

– 1 pt Wrong type of bias (if there is any)

– 1 pt Wrong reasoning

– 2 pts Wrong

– 0.5 pts Didn't mention what kind of coverage bias or just generally said selection bias

– 1 pt Wrong selection, but correct bias mentioned

– 1 pt wrong choice but something close to undercoverage/under-representation or introducing bias mentioned

Question 2

Q1.1b

2 / 2 pts

✔ **– 0 pts** Correct

– 2 pts Wrong choice

– 1 pt Wrong type of bias (if there is any)

– 1 pt Wrong reasoning/ correct reasoning but wrong selection

– 2 pts Wrong

Question 3

Q1.1c

2 / 2 pts

✔ **– 0 pts** Correct

– 2 pts Wrong choice

– 1 pt Wrong reasoning

– 2 pts Wrong

– 0.5 pts looking for either voluntary/non response bias, but valid reasoning or close

Question 4

Q1.1d **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Wrong choice

– 1 pt Wrong type of bias (if there is any)

– 1 pt Wrong reasoning

– 1 pt Wrong choice but correct reasoning

– 2 pts Wrong

Question 5

Q1.1e **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Wrong choice

– 1 pt Wrong type of bias (if there is any)

– 1 pt Wrong reasoning

– 2 pts Wrong

Question 6

Q1.1f **1** / 2 pts

– 0 pts Correct

– 2 pts Wrong choice

✔ **– 1 pt** Wrong type of bias (if there is any)

– 1 pt Wrong reasoning

– 1 pt Wrong choice but correct reasoning

– 2 pts Wrong

Question 7

Q1.1g **0** / 2 pts

– 0 pts Correct

✔ **– 2 pts** Wrong choice

– 1 pt Wrong type of bias (if there is any)

– 1 pt Wrong reasoning

– 2 pts Wrong

Question 8

Q1.2a **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts Incorrectly assume that KNN will perform well

– 1 pt Wrong reasoning for why KNN does not perform well

– 1 pt Wrong suggestions to improve the model

– 3 pts Wrong

Question 9

Q1.2b **3** / 3 pts

✔ **– 0 pts** Correct

– 3 pts Incorrectly assume that KNN will perform well

– 1 pt Wrong reasoning for why KNN will not perform well

– 3 pts Wrong

Question 10

Q1.2c **2** / 2 pts

✔ **– 0 pts** Correct

– 2 pts Wrong trend

Question 11

Q2a **4** / 4 pts

✔ **– 0 pts** Correct

– 1 pt Missing the interaction term or wrong formulation

– 1 pt wrong/missing explanation for x_1

– 1 pt wrong/missing explanation for x_2

– 4 pts Wrong

– 1 pt described X1, X2 type of vals but didn't explain what one or both x1, x2 represents in context to the problem

Question 12

Q2b **8** / 8 pts

✔ **– 0 pts** Correct

– 2 pts wrong explanation for β_0

– 2 pts wrong explanation for β_1

– 2 pts wrong explanation for β_2

– 2 pts wrong explanation for β_3

– 8 pts Wrong

Question 13

Q2c **4** / 4 pts

✔ **– 0 pts** Correct

– 1 pt wrong sign for β_0

– 1 pt wrong sign for β_1

– 1 pt wrong sign for β_2

– 1 pt wrong sign for β_3

– 4 pts Wrong

Question 14

Q2d **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts Wrong confidence interval

– 2 pts Wrong interpretation

– 4 pts Wrong

– 0.5 pts Minor mistake in CI

Question 15

Q2e **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts Only one solution is presented, or one of the solution is wrong.

– 4 pts Wrong

Question 16

Q3a

4 / 4 pts

✔ **− 0 pts** Correct

− 2 pts (i) conclusion not reasonable

− 2 pts (ii) missing / draw the conclusion that the model has enough complexity

− 4 pts Wrong

Question 17

Q3b

4 / 4 pts

✔ **− 0 pts** Correct

− 2 pts (i) fail to identify multicollinearity/collinearity

− 1 pt (ii) wrong reasoning

− 1 pt (iii) wrong solution

− 4 pts Wrong

Question 18

Q3c

4 / 4 pts

✔ **− 0 pts** Correct

− 2 pts (i) Conclude the model is a good fitting or overfitting

− 2 pts (ii) wrong conclusion, should be that bias is too high

− 4 pts missing answer

− 2 pts Doesn't specify on underfit and which way bias leans toward

Question 19

Q3d

2 / 2 pts

✔ **− 0 pts** Correct

− 2 pts Wrong conclusion

− 1 pt Mentions not appropriate, but not overfit. Or mentions both but the answer is only overfit.

Question 20

Q3e

4 / 4 pts

✔ **– 0 pts** Correct

– 2 pts (i) wrong selection for $L2$

– 1 pt (i) wrong reasoning

– 2 pts (ii) wrong selection for $L1$

– 1 pt (ii) wrong reasoning

– 4 pts No answer

Question 21

Q4a

4 / 4 pts

✔ **– 0 pts** Correct

– 1 pt Missing log/ln expression

– 0.5 pts Missing X_1 term

– 0.5 pts Missing X_2 term

– 1 pt Missing X_1X_2 term

– 1 pt Missing constant term

– 0.5 pts Minor mistake

– 4 pts Wrong

Question 22

Q4b

8 / 8 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect/missing interpretation for e^{β_0}

– 2 pts Incorrect/missing interpretation for e^{β_1}

– 2 pts Incorrect/missing interpretation for e^{β_2}

– 2 pts Incorrect/missing interpretation for e^{β_3}

– 1 pt Minor mistake

– 8 pts Wrong

Question 23

Q4c **4** / 4 pts

✔ **– 0 pts** Correct

– 4 pts wrong

– 1 pt math mistake

– 2 pts keep study time fixed

Question 24

Q4d **6** / 6 pts

✔ **– 0 pts** Correct

– 2 pts decision boundary

– 2 pts no prerequisite

– 2 pts prerequisite

– 6 pts wrong

– 1 pt minor mistake $\geq$

– 1 pt not solving for x

– 1 pt minor mistake for calculating hours

Question 25

Q5a **4** / 4 pts

✔ **– 0 pts** Correct

– 1 pt (i) wrong conclusion

– 2 pts (ii) wrong reasoning

– 1 pt (iii) wrong selection

Question 26

Q5b **2** / 4 pts

– 0 pts Correct

– 1 pt (i) wrong conclusion

– 2 pts (ii) wrong reasoning

✔ **– 1 pt** (iii) wrong solution

✔ **– 1 pt** ii) says something relevant to AUC, accuracy or TN/FN, but missing that it attributes back to wrong threshold

– 0.5 pts small mistake in existing response reasoning

Question 27

Q5c

3 / 3 pts

✔ **– 0 pts** Correct

– 3 pts Incorrect

– 1 pt Mentions tuning the threshold but not that we need to lower the threshold or something related to FN

Question 28

Q5d

3 / 3 pts

✔ **– 0 pts** Correct

– 3 pts Incorrect

– 1 pt Mentions tuning threshold, but not that specifically we want to increase the threshold or something correct about reducing FP

Data Science (CS 148)
February 12, 2025

Midterm Exam

Duration: 90 minutes

Write your name and UID: Tiya Chokhani 305933966

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!

Note 4: The final grades will be curved if needed.

1 Short Answers (22 points)

Data collection & Bias. For each of the following parts (a) to (g), choose one of the following options (A or B or C) in the answer box, and briefly *explain the reason* for your answer. If you answer A, indicate the *type of the bias* in the reason box.

A: Introduces bias in the data
B: Amplifies the existing bias in the data (if there is any)
C: Does not introduce or amplify bias in the data

(a) **(2 points)** Dropping examples with missing values from a dataset.

Answer:	Reason:
C	Dropping examples with missing values is okay as long as the values missing dont indicate some trend or some reason they should be random & few compared to sample size

(b) **(2 points)** Labeling new examples using the predictions of an existing model.

Answer:	Reason:
B	Can create a cycle if inital model is biased or has biased as it make biased predictions & then learn from those ~~& make~~ data

(c) **(2 points)** Emailing a questionnaire to a large subset of individuals selected via probability sampling and collecting data from people who respond to your email.

Answer:	Reason:
A	Non response bias if a certain group dont reply

(d) **(2 points)** Predicting the outcome of a company's internal election by posting a non-anonymous survey to the employees.

Answer:	Reason:
A	~~Response bias if employees don't answer they~~ Non response bias if employees dont respond due to it being non-anonymous

(e) **(2 points)** Collecting data by observing people's behavior.

Answer:	Reason:
C	Simply observing people's behaviours has no bias as long as the ~~sample~~ population is well represented in the sample observed.

(f) **(2 points)** Collecting photos of students' ID card at UCLA to train a face recognition model.

Answer:	Reason:
A	convinience bias as student population is centered around

(g) **(2 points)** Collecting CVs of software engineers in Google to train a model to identify good candidates for software engineering position at Google.

Answer:	Reason:
B	If previous hiring has been biased & prejedied model will use that to generate new decsions which will lead to a cycle & the model will end up biased such as discriminating against women bc very few womens CV's were entered.

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 because it would make it hard to ~~[illegible]~~ select the true nearest neighbours as ~~or~~ some predictors would be disproportionately favoured. We need to normalize/scale the predictors to improve this

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

No as too many predictors makes it difficult to select the k nearest neighbours because the distances would become very similar we wouldn't know which data points were true neighbours.

(c) **(2 Points)** How do you expect the value of K in a KNN model to impact the variance of your model?

As k increases variance reduces as the curve smoothens out.

2 Linear Regression (24 points)

Suppose we measured the life expectancy of a group of individuals based on (i) the amount of exercise (X_1) (in minutes) per week and (ii) if they smoke or not (X_2). The effect of exercise on life expectancy depends on if the individual smokes or not. ($X_1 X_2$)

(a) **(4 points)** Model life expectancy (Y) based on exercise (X_1) *and* smoking (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.*

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

where Y is life expectancy, X_1 is the exercise (mins/week)

X_2 is the whether they smoke or not.

Y & X_1 are real valued, X_2 is binary (0 nosmoking, 1 smoking)

(b) **(8 points)** Write the interpretation of each of the coefficients (β_i) in your model.

when nonsmokers, $x_2 = 0$

$y = \beta_0 + \beta_1 x_1$

β_0 is the base avg life expectancy when amt of excersize is 0 & a person doesn't smoke.

β_1 is the avg change in life expectancy per unit increase in excersize for a non smoker.

when smoker $x_2 = 1$

$y = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_1$

$\beta_0 + \beta_2$ is the base avg life expectancy when excersize is 0 & a person smokes.

β_2 is the change in base avg life expectancy when excersize is 0 between nonsmokers & smokers

$\beta_1 + \beta_3$ is the avg change in life expectancy per unit change in excersize for smokers

β_3 is the avg change in the amt life expectancy changes per unit change in excersize for smokers vs nonsmokers

(c) **(4 points)** Life expectancy is larger than zero if an individual does not exercise and does not smoke, it increases with more exercise and increases if the person does not smoke, but exercise increases the life expectancy less if the individual smokes. What are the signs (+ or -) of each coefficient (β_i) in your model?

β_0 (+)

β_0 → positive

β_1 → positive

β_2 → negative

β_3 → negative

(d) **(4 points)** (i) If the standard error of β_1 is σ, what is the 95% confidence interval for β_1? (ii) If the 95% confidence interval for β_1 contains zero, interpret the effect of exercise on life expectancy.

95% confidence interval : $[\mu_{\beta_1} - 2\sigma, \mu_{\beta_1} + 2\sigma]$

if the confidence interval contains 0 we cannot conclude that there is a relationship between life expectancy & amt excersize (min/week)

There isn't enough statistical evidence to reject H_0

(e) **(4 points)** Mention *two* ways that you can improve the data to reduce the standard error of β_1 (assume we cannot reduce the noise in the data).

if we want to reduce standard error we can:

(1) Increase n, so increase the number of samples of β_1

(2) Increase Var so increase the variance of data point

3 Model Selection & Bias-Variance Trade-off (17 points)

We fit a multiple linear regression model to a dataset with multiple predictors. The range of y-values in the data is $[-10, 10]$. Answer each of the following questions *independently*, i.e., later questions are not follow ups on the previous ones.

(a) **(4 points)** If on the *test set*, Mean-Squared-Error (MSE) $= 0.5$ and $R^2 = 0.9$, (i) What can we conclude? (ii) Does the model have enough complexity to model the data?

$R^2 = 0.9$ means that 90% of the variation in Y is explained by our predictors, which is pretty high but the MSE is 0.5 which we cannot interpret without having something to compare against & a high R^2 doesn't necessarily indicate a linear relationship.

(b) **(4 points)** Every time we fit the same linear regression model to the data, some of the coefficients change. (i) What does this indicate? (ii) Why is this a problem? (iii) How can we fix this issue?

This indicates colinearity which means 2 or more of the predictors are highly correlated this is a problem when it comes to interpretation as we're unable to tell how each predictor truly influences our response variable as the coeff keep changing but this wouldn't hurt computation. we can fix this by understanding more about the predictors & deciding which one is more important & dropping the other

(c) **(4 points)** If the distribution of the residuals on training data is uniform and centered around zero, (i) what can we conclude about the complexity of the model? (ii) what can we conclude about the bias of this model?

Uniform residuals indicates an underfitted model i.e. it's complexity is not high enough. when models are underfitted their bias is very high as most predictions will be off by alot.

(d) **(2 points)** If the model's predictions have a large variance, what can we conclude about the complexity of the model?

A large variance usually means a model has been overfit & is too complex. we should regularize it.

(e) **(4 points)** To find the right complexity for our model, we use regularization. (i) Do you choose L1 or L2 regularization, if we only care about the model's performance, and why? (ii) Do you choose L1 or L2 regularization, if we care about model's performance and interpretability, and why?

If we only care about performance we should go with L2 regularization as it's easier to compute Ridge regularization
If we care about interpretation we should go with L1 regularization as it makes high order weff 0 making it alot easier to interpret.

4 Logistic Regression & Decision Boundary (22)

We use Logistic regression to model the probability for students to pass a course ($Y = 1$), based on their study time in hours (X_1) and if they have taken a prerequisite ($X_2 = 1$) or not ($X_2 = 0$). The effect of study time on the probability of passing the course depends on if the student has taken the prerequisite.

(a) **(4 points)** Write the logistic regression formulation to model log (use ln) odds of passing the course, based on X_1 and X_2.

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

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

when not taken prereq $X_2 = 0$

$$\frac{P(Y=1 \mid X_2=0)}{P(Y=0 \mid X_2=0)} = e^{\beta_0} \times e^{\beta_1 X_1}$$

- e^{β_0} is the base odds of passing the course when hours studied is 0 & the student hasn't taken prereq
- e^{β_1} is the amt multiplied to the odds of passing the course for each additional hour studied when the preq hasn't been taken

when taken prereq $X_2 = 1$

$$\frac{P(Y=1 \mid X_2=1)}{P(Y=0 \mid X_2=1)} = e^{\beta_0+\beta_2} \times e^{(\beta_1+\beta_3)X_1}$$

- e^{β_2} is the multiplicative to the base odds of passing when hours studied is 0 when the preq is taken as compared to when it's not. (Base odds for no prereq $\times e^{\beta_2}$ = Base odds for prereq)
- e^{β_3} is the extra amt multiplied to the odds of passing for each additional hour studied for students who took the prereq as compared to those who didn't.

- $e^{\beta_1+\beta_3}$ is the multiplicative for every hour studied to the odds of passing when student has taken prereq

(c) **(4 points)** How do you compare the odds ratio of passing the course for students who have taken the prerequisite with those who have not (for various study times)?

taken : $e^{\beta_0} \times e^{\beta_2} \times e^{\beta_1 X_1} \times e^{\beta_3 X_1}$

not taken : $e^{\beta_0} \times e^{\beta_1 X_1}$

$: e^{\beta_2 + \beta_3 X_1}$

the odds of passing the worse when you take the pre req is $e^{\beta_2 + \beta_3 X_1}$ times the odds of passing when you dont take the pre req

You can calculate it when holding X_1 constant.

(d) **(6 points)** Write the formulation for the decision boundary. How many hours a student who has taken the prerequisite and who has not taken the prerequisite needs to practice to pass the course?

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = 0$ decision boundary

taken prereq :

$\beta_0 + \beta_2 + (\beta_1 + \beta_3) X_1 = 0$

$X_1 = \frac{-\beta_0 - \beta_2}{\beta_1 + \beta_3}$ is the min num of hours of practice needed when taken the pre req

Not taken

$\beta_0 + \beta_1 X_1 = 0$

$X_1 = \frac{-\beta_0}{\beta_1}$ is the min num of hours of practice needed when not taken pre req

5 Classification Metrics (14 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. Answer each of the following questions *independently*, i.e., later questions are not follow ups on the previous ones.

(a) **(4 points)** If the classifier has a high accuracy but a low F1 score, (i) what do you conclude about the data? (ii) how do you explain the discrepancy between accuracy and F1 score? (iii) which metric is better to evaluate the performance of the classifier?

(i) We can conclude the data is skewed. There is a larger num of observations of true false as compared to true positive. This is because the F_1 score takes into account both Recall & precision so it's a better score when it comes to skewed data as it considers positive predictions (true & false better).

. F_1 score is a better metric to evaluate the performance.

(b) **(4 points)** If the classifier has a high AUC but a low accuracy, (i) what do you conclude? (ii) how do you explain the discrepancy between AUC and accuracy? (iii) how can we improve the accuracy in this case?

A low accuracy would indicate a small % of correct predictions out of all & a high AUC would indicate low false positive prediction rate. This can occur due to skewed data where the number of true positives & negative predictions are low. We can improve the accuracy by looking at the num of true negative & true positive predictions.

(c) **(3 points)** Consider the classifier in the main question. How can we make sure to identify *all* the potential cancer patients, without modifying the classifier?

To identify all potential cancer patients we need to increase predictions of $y=1$ so we lower the bar needed to make a positive prediction $P(y=1|x) \geq \pi$, $\pi < 0.5$

(d) **(3 points)** Consider the classifier in the main question. How can we minimize the number of patients who are flagged by mistake, without modifying the classifier?

We want to reduce the num of false positives so we need to $P(y=1|x) \geq \pi$, $\pi > 0.5$ ↓ FP

	PP	PN
AP	TP	FN
AN	FP	TN