Data Science (CS 148)
February 12, 2024

# Midterm Exam

Duration: 90 minutes

**Write your name and UID:**

*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 (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.

**Solution:** A, under-coverage bias
Reason: If examples with missing values belong to certain groups (e.g. some people may not be comfortable answering certain questions), dropping those examples introduce bias.

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

**Solution:** B
Reason: If there is existing bias in the data the predictions of a model trained on the data will be biased towards the majority groups and using those predictions would amplify the bias in the 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.

**Solution:**
A, Non-response Bias
Reason: Only people who choose to respond to the emails are represented in the data.

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

**Solution:**
A, Response Bias
Reason: Since the survey is non-anonymous, people tend to respond dishonestly (e.g. what the boss likes to hear).

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

**Solution:**
A, response-bias
Reason: People change their behavior when they're observed.

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

**Solution:**
A, undercoverage Bias
Reason: Since only UCLA students' faces are included, the data may not include people from under-represented races, genders, etc.

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

**Solution:**
A, undercoverage bias
Reason: Most of the software engineers in google are male, and the model would learn to predict based on gender-related features.

**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?

**Solution:** No, because KNN finds neighbors based on distance, and features with larger scale dominate the distance.

To improve the performance, all features should be standardized.

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

**Solution:** KNN works well with a small number of input features, but struggles when the number of predictors is very large (distances become similar).

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

**Solution:** When you decrease the k the variance will increase, and vice versa.

## 2 Linear Regression (24 points)

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

(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.*

**Solution:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$
$X_1$ (exercise): real-valued (hours per week).
$X_2$ (smoking): binary (1 = smokes, 0 = does not smoke).
$X_1 X_2$ (interaction term): captures the different effect of exercise on life expectancy for smokers vs non-smokers.

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

**Solution:**
$\beta_0$: average life expectancy for non-smokers who do not exercise.
$\beta_1$: one unit increase in exercise changes life expectancy for non-smokers by $\beta_1$.
$\beta_2$: average difference between life expectancy of smokers vs non-smokers, who do not exercise.
$\beta_3$: one unit increase in exercise changes the expected life expectancy for smokers $\beta_3$ more than non-smokers.

(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 ($\beta_i$) in your model?

**Solution:**
$\beta_0 > 0, \beta_1 > 0, \beta_2 < 0, \beta_3 < 0$

(d) **(4 points)** (i) If the standard error of $\beta_1$ is $\sigma$, what is the 95% confidence interval for $\beta_1$? (ii) If the 95% confidence interval for $\beta_1$ contains zero, interpret the effect of excessive on life expectancy.

**Solution:** $\beta_1 \pm 2\sigma$. If the confidence interval contains zero, then there is no statistically significant evidence that exercise affects life expectancy (this is similar to p-value>0.05).

(e) **(4 points)** Mention *two* ways that you can modify the data to reduce the standard error of $\beta_1$.

**Solution:** 1. Collect more data. 2. Increase the coverage of data points (i.e., increase the diversity of predictors).

# 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?

**Solution: Although MSE is low and $R^2$ is high, we cannot conclude that this model has the right complexity for the data.**
$R^2$ of 0.90 means that 90% of the variance in the data is explained by the model and the model is better than the naive baseline of predicting average $y$ values. Low MSE means the complexity is not too small for the data and the noise in the data is small, but it doesn't mean that the model has the right complexity.

(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?

**Solution:**

(i) There is multicollinearity between some of the predictors.

(ii) Lack of interpretability because the coefficients are not unique.

(iii) Drop all but one of the correlated predictors, ideally keep the predictor that is the cause.

(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?

**Solution:**

(i) The model is too simple (underfitting). **Uniform** distribution means there are similar number of large and small residuals.

(ii) The model has a high bias, and not enough complexity.

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

**Solution:** Large variance in predictions indicates that the model is highly sensitive to small changes in the input data, which means that the model is too complex (overfitting).

(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?

(i) L2. L2 is faster/easier to compute and makes the model smoother but does not necessarily make any of the coefficients equal to 0. Hence, is not useful for interpretability.

(ii) L1. Using L1 regularization makes some of the coefficients exactly equal to zero, reducing the number of features, which makes it easier to understand the model's decisions.

# 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 ($X_1$) 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 $X_2$.

**Solution:**

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

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

**Solution:**

- $e^{\beta_0}$ - the odds of passing for students who don't study and didn't take the prerequisite.
- $e^{\beta_1}$ - the multiplicative change in the odds of passing for one extra hour of study time, for students who didn't take the prerequisite.
- $e^{\beta_2}$ - the odds ratio of passing for students who took the prerequisite and don't study over those who didn't take the prerequisite and don't study.
- $e^{\beta_3}$ - for every extra hour of studying, $e^{\beta_3}$ is the ratio of the multiplicative change in odds of passing for students who took the prerequisite vs those who didn't.

(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?

**Solution:**
***For a fixed study time*** $X_1$, $\frac{e^{\beta_0+\beta_2+(\beta_1+\beta_3)X_1}}{e^{\beta_0+\beta_1 X_1}}$ is the odds ratio of passing the exam for students who took the prerequisite vs those who didn't.

(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?

**Solution:**
**Decision Boundary:**

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

**No Prerequisite:**

$$\beta_0 + \beta_1 X_1 = 0$$

$$t \geq X_1 = -\frac{\beta_0}{\beta_1}$$

**Taken Prerequisite:**

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

$$t \geq X_1 = -\frac{\beta_0 + \beta_2}{\beta_1 + \beta_3}$$

# 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? (ii) How do you explain the discrepancy between accuracy and F1 score? (iii) Which metric is better to evaluate the performance of the classifier?

**Solution:**

(i) Data is imbalanced.

(ii) A high accuracy means the majority of the model's predictions are correct. A low F1 score indicates that either precision or recall are low. For an imbalanced data (e.g. most examples have $Y = 0$), the classifier that just predicts $Y = 0$ will have a high accuracy but fails to identify the $Y = 1$ class. In this case, even though the accuracy is high, the F1 score is low.

(iii) F1 Score is better when the dataset is imbalanced.

(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?

**Solution:**

(i) Data is imbalanced. Classifier is good but the threshold 0.5 is not optimal for the data.

(ii) The classifier ranks examples correctly (based on the probability of having $Y = 1$), but since threshold 0.5 is not optimal for classification, accuracy is low.

(iii) Change the decision threshold.

(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 potential cancer patients?

**Solutions:** We should decrease the threshold to get a high Recall while ensuring an acceptable False Positive Rate (FPR).

(d) **(3 points)** How do you use the output probabilities of the above classifier to predict $Y = 1$ to minimize the number of patients who are flagged by mistake?

**Solutions:** We should increase the threshold that maximizes precision while keeping recall at an acceptable level.