Data C100/200 - Midterm 2

Fall 2025

Name: ____

Email: __________________________________@berkeley.edu

Student ID: ______________________________________

Name and SID of the person on your left: __________________________

Name and SID of the person on your right: __________________________

Exam Room:______________________ Seat Number: ________________

Instructions:

This exam consists of **30 points** spread out over **7 questions**. The exam must be completed in **50 minutes** unless you have accommodations supported by a DSP letter.

- Note that some questions have circular bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**. Please **shade in** the circle/box **fully** to mark your answer.
- Blank answers and incorrect answers are graded identically, so it's in your best interest to answer every question.
- **You MUST write your Student ID number at the top of each page.**
- You should not use a calculator, scratch paper, or notes you own other than the reference sheets distributed at the beginning of the exam.

For all Python questions, you may assume `Pandas` has been imported as `pd`, `NumPy` as `np`, the Python RegEx library as `re`, `matplotlib.pyplot` as `plt`, and `seaborn` as `sns`.

Honor Code:
As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: ______________________________________

This page has been intentionally left blank.

1 Ridge of the Machines [1.5 Pts]

Suppose we want to fit a multiple regression model using **Ridge** regularization. The model has **3 features and an intercept**. Before fitting the final model, we want to select the best value of the regularization hyperparameter for our particular dataset. Using **8-fold cross validation** (CV), we compute the CV error for **11 different values** of the hyperparameter.

For each of the following questions, select the correct option out of the choices below:

A. 8×11
B. $8 + 11$
C. $(3 + 1) \times 1 \times 11$
D. $(3 + 1) \times (8 - 1) \times 11$
E. 1×11
F. $(3 + 1) \times (8 + 11)$
G. $(8 - 1) \times 11$
H. $(3 + 1) \times 8 \times 11$

(a) [0.5 Pts] To select the best performing value of the regularization hyperparameter via 8-fold CV, how many models must be fit?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

(b) [0.5 Pts] To select the best performing value of the regularization hyperparameter via 8-fold CV, how many times must **each data point** be used as part of a **held-out validation** fold?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

(c) [0.5 Pts] To select the best performing value of the regularization hyperparameter via 8-fold CV, how many times must **each data point** be used in the **training dataset** of a fitted model?

○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

2 I Think You're Projecting... [5 Pts]

You fit an **OLS model with an intercept** using a design matrix $\mathbb{X}$ with n rows and $p+1$ columns. The bias (intercept) column is the first column of $\mathbb{X}$, and the remaining columns correspond to the p features in the model. Let $\mathbb{Y}$ be the vector of true outcomes and $\hat{\mathbb{Y}}$ be the OLS predictions.

Note: $\mathbb{1}$ is a vector of length n where all elements are 1. Assume $\mathbb{X}$ is full column rank.

(a) [1 Pt] $\mathbb{Y}$ is _________ $\hat{\mathbb{Y}}$.

○ Orthogonal to

○ In the span of

○ NOT orthogonal to and NOT in the span of

(b) [1 Pt] $\mathbb{1}$ is _________ $\mathbb{X}$.

○ Orthogonal to

○ In the span of

○ NOT orthogonal to and NOT in the span of

(c) [1 Pt] $\mathbb{Y} - \hat{\mathbb{Y}}$ is _________ $\mathbb{X}_{:,p}$.

○ Orthogonal to

○ In the span of

○ NOT orthogonal to and NOT in the span of

(d) [1 Pt] $\mathbb{1}$ is _________ $\mathbb{Y} - \hat{\mathbb{Y}}$.

○ Orthogonal to

○ In the span of

○ NOT orthogonal to and NOT in the span of

(e) [1 Pt] $\mathbb{Y} - \hat{\mathbb{Y}}$ is _________ $\hat{\mathbb{Y}}$.

○ Orthogonal to

○ In the span of

○ NOT orthogonal to and NOT in the span of

3 Big Steppers [6 Pts]

The table below shows several iterations of **batch gradient descent** for a **constant model**. Complete the missing entries in the table.

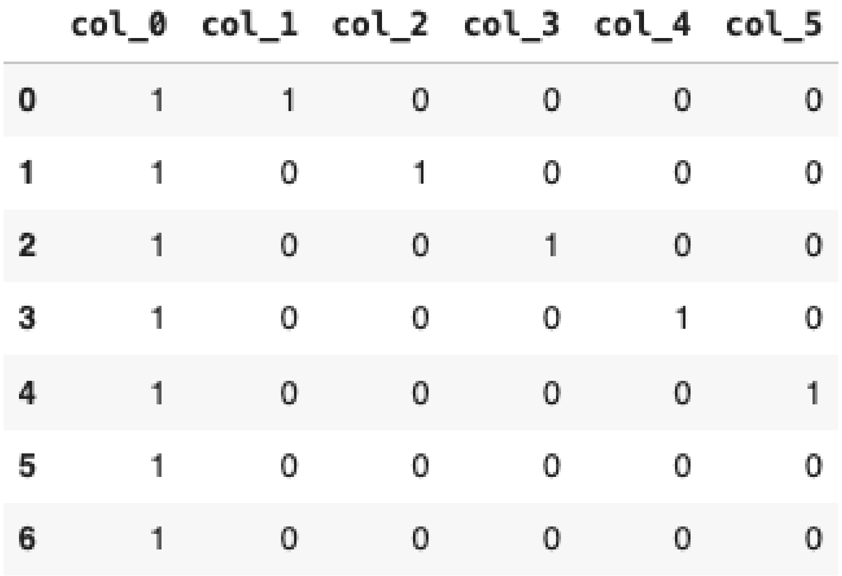
- Assume that the learning rate $\alpha = 1$.
- If the value of a table entry is ambiguous or impossible to know, write **NA**.

t	$\theta^{(t)}$	$\theta^{(t+1)}$	$L(\theta^{(t)})$	$\frac{d}{d\theta}L(\theta^{(t)})$
0	10	(a) ______________	1	(b) ______________
1	11	(c) ______________	(d) ______________	(e) ______________
2	(f) ______________	9	0.5	-2

4 One-Hot Set of Data [1.5 Pts]

The design matrix below was used to fit an **OLS model with an intercept** and one categorical feature. The categorical feature has been one-hot encoded.

Note: Assume that the categorical feature has no missing values and that no categories have been combined.



	col_0	col_1	col_2	col_3	col_4	col_5
0	1	1	0	0	0	0
1	1	0	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	0	1	0
4	1	0	0	0	0	1
5	1	0	0	0	0	0
6	1	0	0	0	0	0

How many **unique values** of the categorical feature are there?

- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ Not enough information to answer

5 Expect the Expected [4 Pts]

Consider the categorical probability distribution in the table below.

x	$P(X = x)$
1	0.25
3	0.5
5	0.25

Suppose we generate 10 independent random variables $X_1, X_2, \ldots, X_{10}$, where each random variable is drawn from the distribution above.

Note: All answers should be algebraic expressions that contain ***only numbers*** *and* ***no variables****. For example,* 10, $20 + 5$, *and* $(5 + 10)^2$ *are acceptable answers.* $50x$ *and* $n^2 + n$ *are not.*

Note: Each part is assessed independently; errors in prior parts will not carry forward.

(a) [1 Pt] What is $\mathbb{E}[X_1]$?

Note: Answers without work will not receive credit.

(b) [1.5 Pts] What is the expected value of the **average** of the 10 random variables?

Note: Be sure to derive your answer using the rules of expectation. Answers without work will not receive credit.

(c) [1.5 Pts] $\mathrm{Var}(X_1) = 2$. What is the variance of the **average** of the 10 random variables?

Note: Be sure to derive your answer using the rules of variance. Answers without work will not receive credit.

6 Live, Love, LASSO [6 Pts]

Suppose we want to fit an OLS model with p features and an intercept. Call this **Model A.** For each proposed change to Model A in the table below, **select all possible effects** on four measures: (model bias)^2, model variance, MSE on the training data, and MSE on held-out test data.

- For each box, select Increase, Decrease, or both. **At least one option will always apply**.
- For each box, the effect(s) you select **do not have to be guaranteed**, they just **have to be possible**.
- For each row, you should assume there are **no changes** to the model **except for the proposed change**.
- You should assume that the data-generating process does not change, and that the model fitting process always converges.
- You do not have to consider whether the proposed change has no effect on the measure, though it may be possible.

Change to Model A	**(Model Bias)^2**	**Model Variance**	**MSE on training data**	**MSE on held-out test data**
Add one new feature to Model A that is the square of an existing feature	□ Increase □ Decrease	□ Increase □ Decrease	□ Increase □ Decrease	□ Increase □ Decrease
Remove one feature from Model A	□ Increase □ Decrease	□ Increase □ Decrease	□ Increase □ Decrease	□ Increase □ Decrease
Use LASSO regularization to fit the model, with $\lambda > 0$	□ Increase □ Decrease	□ Increase □ Decrease	□ Increase □ Decrease	□ Increase □ Decrease

7 Sleep Now, Predict Later [4 Pts]

The Data 100 team decides to explore the **relationship** between **hours of sleep**, whether or not the student has a major **affiliated with CDSS**, and **GPA**. The team collected the following data for 100 UC Berkeley undergraduates:

- `hours_sleep`: The **average** amount of sleep the student gets every night. (type: `np.float`)
- `is_cdss`: An **indicator** of whether the individual has a major affiliated with CDSS. A value of 1 indicates that the major is affiliated with CDSS, and 0 indicates that it is not. (type: `int`)
- `gpa`: The GPA of the student. (type: `np.float`)

The team fits the following multiple regression model to the collected data:

$$\widehat{\text{gpa}} = \theta_0 + \theta_1 \cdot \text{hours\_sleep} + \theta_2 \cdot \text{is\_cdss}$$

Suppose the optimal estimated parameters are $\hat{\theta}_0 = 1.2$, $\hat{\theta}_1 = 0.23$, and $\hat{\theta}_2 = 0.5$.

(a) [2 Pts] What is the interpretation of $\hat{\theta}_0$? Answer in exactly one sentence.

(b) [2 Pts] What is the interpretation of $\hat{\theta}_1$? Answer in exactly one sentence.

Please state any relevant assumptions in the box below (Optional).

You are done with the midterm- Congratulations!

Draw your favorite DATA 100/200 memory so far!

Regex Crossword (Optional, not graded)

Fill each square with a single CAPITAL letter so that every row and column matches its corresponding regular expression.

<table>
<tr><th></th><th>H[GIN\s]+</th><th>[UMN\s]*</th><th>\w\s(U|M|N)+</th><th>([GIN]|&|\s)+</th><th></th></tr>
<tr><td>\w*</td><td></td><td></td><td></td><td></td><td>(HU|GI)+</td></tr>
<tr><td>(\w)N\W&</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>\WM[M-Z]+</td><td></td><td></td><td></td><td></td><td>[^O-T]+U.</td></tr>
<tr><td>[GIN]+.</td><td></td><td></td><td></td><td></td><td>[H-K].+\s</td></tr>
</table>