Data 100/200, Final

Fall 2023

Name: ______________________________

Email: ______________________________@berkeley.edu

Student ID: ______________________________

Name and SID of the person on your right: ______________________________

Name and SID of the person on your left: ______________________________

Instructions:

This final exam consists of **90 points** spread out over **10 questions** and the Honor Code and must be completed in the **170 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**. These will always have at least one correct answer. Please shade in the box/circle to mark your answer.

You must write your Student ID number at the top of each page.

Points Breakdown:

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Points	14	7	10	7	5	7	11	12	8	8

Honor Code [1 Pt]:

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 Fizzy Vizzies [14 Pts]

This question involves coding. All code for each part, where applicable, must be written in Python. Assume that the `pandas` library is imported as `pd`, the `numpy` the library is imported as `np`, and the Python regular expression library is imported as `re`.

The professors want to throw a picnic for all of Data 100 course staff to celebrate the end of the semester but can't decide which type of soda to get. They choose to consult Data 100 staff about their favorite fizzy drinks.

(a) [2 Pts] To gauge the popularity of sodas, suppose the professors only ask the first 5 people who show up to the weekly course staff meeting for their favorite type of soda. Assuming that there are more than 5 members of course staff, which of the following statements are true regarding this process? **Select all that apply.**

☐ A. This is a simple random sample.

☐ B. This is a convenience sample.

☐ C. The results will likely have chance error.

☐ D. The results will likely have selection bias.

The professors realize it's easier to ask the entire staff to answer a poll about their favorite sodas. They record each staff member's favorite soda and its flavor and each staff member's year in school. Each staff member only responds once, and no two staff members have the same name. The results of this poll are stored in a `pandas DataFrame` named `staff_poll`. The first few rows are shown below.

	name	soda	flavor	year
0	Sean	Diet Coke	Cola	Graduate
1	Alina	Orange Fanta	Orange	5
2	Geovanni	Coca-Cola	Cola	4
3	Ishani	Mountain Dew	Lemon-Lime	3
4	Shreya	Sprite	Lemon-Lime	3
5	Stephanie	None	None	3
6	Charlie	A&W	Root Beer	3

(b) [1 Pt] What is the minimum set of columns that form the primary key of this dataset? Write the label(s) of the column or group of columns.

(c) [1 Pt] Which **variable type** best describes each of the following columns of `staff_poll`?

	Quantitative Continuous	Quantitative Discrete	Qualitative Ordinal	Qualitative Nominal
(i) `flavor`	○	○	○	○
(ii) `year`	○	○	○	○

(d) [2 Pts] The professors want to identify which sodas are a sub-type of Coca-Cola. Sodas that fit this category either contain the sub-strings `"Coke"` or `"Coca-Cola"`. The professors store a RegEx pattern in the variable `soda_pattern` which extracts the following sub-strings in the example below:

```
re.findall(soda_pattern, "Some people like Diet Coke or Cherry
Coke more than regular Coca-Cola")
```
Output: `["Coke", "Coke", "Coca-Cola"]`

Which of the following RegEx patterns could be `soda_pattern`? You only need to worry about the example text above and do not need to consider cases that are not present in the example. **Select all that apply**.

☐ A. `r"Co[ke|ca\-Cola]"`

☐ B. `r"Co\w{2}\-?\w*"`

☐ C. `r"Coke|Coca-Cola"`

☐ D. `r"Co(ke|ca-Cola)"`

(e) [2 Pts] The professors want to know which type of soda was most popular amongst all staff members, so they create a `Series` called `fav_flavors` using the following line of code:

```
fav_flavors = staff_poll["flavor"].value_counts()
```

Assume that the `seaborn` library is imported as `sns` and the `matplotlib pyplot` library is imported as `plt`. Which of the following lines of code will visualize the data stored in `fav_flavors`? **Select all that apply**.

☐ A. `sns.countplot(data=staff_poll, x="flavor")`

☐ B. `plt.hist(fav_flavors)`

☐ C. `sns.boxplot(x=fav_flavors.values)`

☐ D. `plt.bar(fav_flavors.index, fav_flavors.values)`

(f) [2 Pts] Which of the following types of visualizations are typically used to illustrate the distribution of a single quantitative variable? **Select all that apply.**

☐ A. Histogram

☐ B. Scatterplot

☐ C. Line Plot

☐ D. KDE Plot

(g) [4 Pts] The professors are interested in filtering `staff_poll` to only include rows which have an entry in the `soda` column that is at least 2 words and 4 characters long. We define having two words as anything with a space character, such as `"Mountain Dew"`.

(i) First, write a line of code to assign `has_space` to a `Series` containing boolean values for whether or not each entry of the `soda` column in `staff_poll` contains a space character (in your answer, please represent the space character as `"␣"`).

`has_space = ____________________________`

(ii) Next, write a line of code to create a `DataFrame`, which has the same structure as `staff_poll` but only contains rows that have an entry in the `soda` column that is at least 2 words and 4 characters long. You may assume `has_space` was defined correctly.

2 Gradient Gala [7 Pts]

(a) [4 Pts] Suppose Tina has the following model: $\hat{y} = 2\theta_0 + \theta_0^4\theta_1 x - e^{2\theta_1}x^2$. If Tina decides to use Mean Squared Error (MSE) as the loss function, *L*, to select the optimal choice of θ_0 and θ_1, which of the following expressions represents the **gradient vector** $\nabla_\theta L$?

Note: Grading will be done based on the work you show in the box below. Please indicate clearly how you calculate each item in the gradient vector.

○ A. $\nabla_\theta L = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n 2(y_i - 2\theta_0 - \theta_0^4\theta_1 x_i + e^{2\theta_1}x_i^2)(-2 - 4\theta_0^3\theta_1 x_i) \\ \frac{1}{n}\sum_{i=1}^n 2(y_i - 2\theta_0 - \theta_0^4\theta_1 x_i + e^{2\theta_1}x_i^2)(-\theta_0^4 x_i + 2e^{2\theta_1}x_i^2) \end{bmatrix}$

○ B. $\nabla_\theta L = \begin{bmatrix} \frac{2}{n}\sum_{i=1}^n (y_i + 2\theta_0 + \theta_0^4\theta_1 x_i - e^{2\theta_1}x_i^2)(2 + 4\theta_0^3\theta_1 x_i) \\ \frac{2}{n}\sum_{i=1}^n (y_i + 2\theta_0 + \theta_0^4\theta_1 x_i - e^{2\theta_1}x_i^2)(\theta_0^4 x_i - 2e^{2\theta_1}x_i^2) \end{bmatrix}$

○ C. $\nabla_\theta L = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n 2(y_i - 2\theta_0 - \theta_0^4\theta_1 x_i + e^{2\theta_1}x_i^2)(-4\theta_0^3\theta_1 x_i) \\ \frac{1}{n}\sum_{i=1}^n 2(y_i - 2\theta_0 - \theta_0^4\theta_1 x_i + e^{2\theta_1}x_i^2)(2e^{2\theta_1}x_i) \end{bmatrix}$

○ D. $\nabla_\theta L = \begin{bmatrix} \frac{2}{n}\sum_{i=1}^n (y_i + 2\theta_0 + \theta_0^4\theta_1 x_i - e^{2\theta_1}x_i^2)(4\theta_0^3\theta_1 x_i - 2e^{2\theta_1}x_i) \\ \frac{2}{n}\sum_{i=1}^n (y_i + 2\theta_0 + \theta_0^4\theta_1 x_i - e^{2\theta_1}x_i^2)(4\theta_0^3\theta_1 x_i - 2e^{2\theta_1}x_i) \end{bmatrix}$

(b) [2 Pts] Tina decides to try a new loss function, $\tilde{L}$, defined based on the two model parameters θ_0 and θ_1:

$$\tilde{L}(\theta) = 2\theta_0 + \theta_0^4\theta_1 - e^{2\theta_1}$$

Assume that we run gradient descent with $\theta_0^{(0)} = 1$ and $\theta_1^{(0)} = 2$. For a learning rate of $\alpha = 0.5$, calculate the value of $\theta_0^{(1)}$. Please simplify your answer.

$\theta_0^{(1)}$ = ______________________

(c) [1 Pt] Assuming that $\tilde{L}$ is a convex function, where is the optimal value of $\hat{\theta}_0$ for $\tilde{L}$ in relation to $\theta_0^{(0)}$ from the previous part?

- ◯ A. $\hat{\theta}_0 < \theta_0^{(0)}$
- ◯ B. $\hat{\theta}_0 = \theta_0^{(0)}$
- ◯ C. $\hat{\theta}_0 > \theta_0^{(0)}$
- ◯ D. Not enough information

3 Crust-Validation [10 Pts]

Yash feels mighty crusty and decides to bake some pies for the professors. The professors then rate their satisfaction with the pies on a scale of 1 to 10 (10 being the highest satisfaction). Yash records this satisfaction score and other data about the pies in a `DataFrame` called `pies`. The first 5 rows are given below.

	flavor	crust	lattice	open_face	diameter_in	diameter_cm	professor	satisfaction
0	apple	standard	0	0	9.2	23.368	Narges	9
1	chocolate	graham cracker	0	1	10.1	25.654	Fernando	7
2	peach	standard	1	0	8.5	21.590	Narges	6
3	banana cream	graham cracker	0	1	9.0	22.860	Narges	10
4	pumpkin	graham cracker	0	1	9.3	23.622	Fernando	8

(a) [3 Pts] Fill in the blanks to write a line of code that generates a `Series` that displays the top 5 `flavor` and `crust` combinations with the highest average `satisfaction`.

```
pies.________A________[____B____].mean()._______C______[:5]
```

(i) Fill in blank `A`:

(ii) Fill in blank `B`:

(iii) Fill in blank `C`:

(b) [2 Pts] Yash thinks that all pies with an `open_face` do not have a `lattice`. Both columns are binary, with `1` meaning that the pie has an `open_face` or `lattice` for that respective column, and `0` meaning that the pie does not have an `open_face` or `lattice` for that respective column. Select all the lines of code that return `True` if Yash is correct, and `False` otherwise.

☐ A. `sum(pies[pies["open_face"]==1]["lattice"])==0`

☐ B. `sum(pies[pies["open_face"]==0]["lattice"])==len(pies)`

☐ C. `len(pies[pies["lattice"]+pies["open_face"]>1])==0`

☐ D. `sum(pies["lattice"]+pies["open_face"])==pies.shape[0]`

(c) [3 Pts] Yash decides to train an Ordinary Least Squares (OLS) model to predict the `satisfaction` score.

(i) There are 6 types of `flavor` of pie, 2 types of `crust`, and 2 `professor` evaluators. Yash decides to use **ONLY** these columns to fit his OLS model. If he decides to one-hot encode these columns, what is the maximum number of columns that could exist in Yash's design matrix $\mathbb{X}$ for OLS to yield a unique solution and for the sum of residuals to equal 0?

Number of Columns = ☐

(ii) Now Yash wants to train a new model that utilizes **ONLY** the numerical columns (including `lattice` and `open_face`). What is the maximum number of columns that could exist in Yash's new design matrix $\mathbb{X}$ for OLS to yield a unique solution and for the sum of residuals to equal 0?

Number of Columns = ☐

(d) [2 Pts] Yash decides to use LASSO regression with the design matrix from part 3c(ii). He has 4 candidate values for the regularization parameter λ, and decides to use 5-fold cross-validation to find the best λ. How many validation errors will he need to calculate?

Number of Validation Errors = ☐

4 Leg-ularization [7 Pts]

Rohan is an upstanding citizen who likes to pick up pieces of litter on the street when he goes for a jog. He wishes to model the number of pieces he picks up (y_i) from the number of miles he runs (x_i) and decides to use the following model:

$$\hat{y}_i = \theta_0 + \theta_1 4^{x_i}$$

Rohan begins to track data from his last three runs and stores it in the table below.

x	y
0.5	3
1	7
2	31

(a) [2 Pts] With n as the number of data points and m as the number of features, which of the following are appropriate loss functions with L2 regularization? **Select all that apply.**

☐ A. $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{m}(\theta_i)^2$

☐ B. $\frac{1}{m}||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda||\theta||_2^2$

☐ C. $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\lambda \sum_{i=1}^{m}\theta_i)^2$

☐ D. $\frac{1}{n}||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2 + \lambda||\theta||_1$

(b) [3 Pts] Rohan decides to utilize LASSO Regression for his model and finds the ideal weights to be $\hat{\theta}_0 = -1, \hat{\theta}_1 = 2$. Next, given $\lambda = \frac{1}{2}$, **calculate the empirical risk.**

Answer = ______________________

(c) [2 Pts] Now Rohan wants to use a different model with even more features. For each of the following subparts, select the **best** choice for the following scenarios.

(i) Rohan has too many features and wants to narrow down the number of features.

◯ A. L1 Regularization

◯ B. L2 Regularization

◯ C. None of the above

(ii) The model has a very high validation error.

◯ A. We should increase λ

◯ B. We should decrease λ

◯ C. Not enough information

5 Variable Vibes [5 Pts]

Mir has begun to track the amount of time (in minutes) it takes for him to ride his electric skateboard to and from class. He notices that his commute time to class can be represented as a normal distribution with an expectation of 6 and a variance of 3. His commute time on his way back from class can be represented as a normal distribution with an expectation of 3 and a variance of 12. Assume that each commute is independent of all other commutes.

(a) [2 Pts] What is Mir's expected total commute time (both directions) over a 5-day week? Grading will be done based on the work you show in the box below.

Answer = ______________________

(b) [2 Pts] What is Mir's total commute time (both directions) variance over a 5-day week? Grading will be done based on the work you show in the box below.

Answer = ______________________

(c) [1 Pt] Mir decides to measure the difference between the total travel time going to class and the total travel time leaving the class over 5 days. Which of the following describes the relationship between the variance of this new metric and the variance in part (b)?

- ◯ A. It will be greater than the variance in part (b).
- ◯ B. It will be equal to the variance in part (b).
- ◯ C. It will be less than the variance in part (b).
- ◯ D. Not enough information.

6 Bias-Variance Trade Offer [7 Pts]

Matthew and Brandon have photos of their pets curled up and photos of bagels. They both decide to train models to predict whether an image shows an animal or a bagel.

(a) [1 Pt] **True or False.** Matthew opts to build a model to calculate the probability that each photo depicts a bagel. If he opts to use a constant model, with the proportion of bagel photos in the dataset as the constant, he will always achieve a model bias of 0.

◯ A. True

◯ B. False

(b) [1 Pt] **True or False.** If Matthew adds 10 more good-quality photos of bagels or pets to his training set for his model from part (a), the model is generally less likely to overfit.

◯ A. True

◯ B. False

(c) [2 Pts] Brandon trains a logistic regression model that classifies a photo as "animal" or "bagel". He notices that his training and validation errors are both too high. Which of the following are reasonable methods that Brandon can try to decrease **both** errors? **Select all that apply**.

☐ A. Reduce the number of features in the model.

☐ B. Increase the number of features in the model.

☐ C. Decrease the regularization parameter λ.

☐ D. Perform further feature engineering.

Matthew decides to try a new method. He first comes up with a model $h(x)$, which has a bias of B and a variance of V on the original dataset. Next, he builds a new model $f(x)$ using this process:

1. Obtain n independent, random samples from the original dataset.
2. For each dataset, train a new version of $h_i(x)$ for each i from 1 to n, where $h_1(x)$ was the $h(x)$ model trained on the first dataset and $h_n(x)$ was the $h(x)$ model trained on the n^{th} dataset.
3. Finally, define $f(x)$ as $f(x) = \frac{1}{n}\sum_{i=1}^{n} h_i(x)$.

(c) [3 Pts] What is the model's **bias squared** of the $f(x)$ model in terms of B and n? Grading will be done based on the work you show in the box below.

○ A. B^2

○ B. $\frac{1}{n}B^2$

○ C. $(\frac{B}{n})^2$

○ D. nB^2

7 The Original Was Better Than The SQL [11 Pts]

This question involves SQL databases. All code for this question, where applicable, must be written as SQL queries. ***In each blank, you may write as much code as is necessary, provided it fits the given skeleton code***.

Lillian is planning a Data 100 movie night! She asks each staff member to name their favorite movie and rate it on a scale of 1-10, and saves the results in a SQL table called `favorites`. Lillian gathers additional information about movies, including movies that weren't in `favorites`, and stores it in another SQL table called `movies`.

The first 5 rows of both tables are shown below:

	name	fav_movie	personal_rating
0	Angela	The Incredibles	9.5
1	Srikar	Avengers: Infinity War	9.1
2	Pragnay	Avengers: Infinity War	8.6
3	Rahul	Star Wars: The Empire Strikes Back	8.4
4	Celine	The Grand Budapest Hotel	10.0

`favorites`

	title	franchise	imdb	year
0	Star Wars: A New Hope	Star Wars	8.6	1977
1	Star Wars: The Empire Strikes Back	Star Wars	8.7	1980
2	Star Wars: Revenge of the Sith	Star Wars	7.6	2005
3	The Avengers	Avengers	8.0	2012
4	Avengers: Infinity War	Avengers	8.4	2018

`movies`

(a) [4 Pts] Fill in the blanks below to write a SQL query that identifies the 5 movie franchises with the highest average IMDB ratings and at least 3 entries.

An example output is given below. The output should contain the columns: `franchise` (name of the franchise), `avg_imdb` (average IMDB score of movies within a given franchise in the `movies` table), and `num_entries` (number of movies within a given franchise in the `movies` table).

	franchise	avg_imdb	num_entries
1	Avengers	8.025000	4
0	Star Wars	7.418182	11

```
SELECT franchise, _________________A____________________
FROM movies
____________________B_____________________
____________________C_____________________
____________________D_____________________
____________________E_____________________;
```

(i) Fill in Blank A:

(ii) Fill in Blank B:

(iii) Fill in Blank C:

(iv) Fill in Blank D:

(v) Fill in Blank E:

(b) [4 Pts] Lillian wants to pick a movie that features a famous actor and collects data for the table `actors`. The first 5 rows are shown below.

	title	actor
0	The Avengers	Chris Evans
1	The Avengers	Jeremy Renner
2	Avengers: Infinity War	Chris Evans
3	Avengers: Infinity War	Chris Hemsworth
4	Star Wars: The Empire Strikes Back	Mark Hamill

Fill in the blanks to write a SQL query to return a table that displays the name of each `actor` and the latest year they appeared in a staff member's favorite movie (name this column `most_recent`). **Note that actors may have starred in multiple movies**. An example output is shown below.

	actor	most_recent
0	Chris Evans	2018
1	Chris Hemsworth	2018
2	Jeremy Renner	NULL
3	Mark Hamill	1980

Hint: If an actor has never appeared in a staff member's favorite movie, they should have a `most_recent` entry of NULL.

```
SELECT a.actor, _____________A_____________
FROM favorites AS f
______B______ movies AS m ON __________C__________
______D______ actors AS a ON a.title = m.title
_________________E_________________;
```

(i) Fill in Blank A:

(ii) Fill in Blank B:

(iii) Fill in Blank C:

(iv) Fill in Blank D:

(v) Fill in Blank E:

(c) [3 Pts] Lillian stores the average IMDB score across all movies (6.8) in a variable called `avg_all_imdb`.

Fill in the blank to write a SQL query that returns a copy of the original `favorites` table with an additional column named `above_imdb_mean`, which contains a value of `"Above"` if the `personal_rating` is above the mean IMDB rating, `"Below"` if the `personal_rating` is below this mean, and `"Equal"` otherwise.

	name	fav_movie	personal_rating	above_imdb_mean
0	Angela	The Incredibles	9.5	Above
1	Srikar	Avengers: Infinity War	9.1	Above
2	Pragnay	Avengers: Infinity War	8.6	Above
3	Rahul	Star Wars: The Empire Strikes Back	8.4	Above
4	Celine	The Grand Budapest Hotel	10.0	Above

```
SELECT name, fav_movie, personal_rating,
_____________________ AS above_imdb_mean
FROM favorites;
```

Using as many lines as you need, fill in the blank:

8 Double Feature [12 Pts]

Yuerou can't believe everyone she knows only decided to watch one of Barbie or Oppenheimer. She collects data and gives each person a label of $y_i = 1$ if a person watched Barbie, and $y_i = 0$ if a person watched Oppenheimer. Assume that everybody in the dataset only watched one of the movies. The data Yuerou collected is displayed in the following table:

$\mathbb{X}_{:,1}$	$\mathbb{X}_{:,2}$	y
1	2	1
2	-1	1
0	0	0
2	0	0

(a) [1 Pt] Is this data linearly separable?

◯ Yes

◯ No

(b) [3 Pts] Yuerou trains a logistic regression model with an intercept term and finds the optimal model parameters to be $\hat{\theta} = [3, \frac{1}{2}, -2]^T$.

Suppose that we observe a new data point $x_{new} = [x_{new,1}, x_{new,2}]^T = [4, -1]^T$, with a corresponding $y_{new} = 0$. Based on the model, what is the probability that the person represented by x_{new} watched Oppenheimer?

Note: You may leave your final answer as an expression in terms of e.

Probability: ______________________________

(c) [2 Pts] Suppose that for just this part, Yuerou decides to minimize the following loss function:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} y_i(y_i - \hat{y}_i) + (1 - y_i)(y_i - \hat{y}_i)$$

Which of the following is correct regarding this loss function? **Select all that apply.**

☐ A. Under this loss function, worse model predictions can incur less penalty than better model predictions.

☐ B. This loss function is the complement of accuracy (i.e., $1 - L(\theta) = $ accuracy).

☐ C. A model that always predicts $\hat{y} = 0$ will always have a loss ≥ 0.

☐ D. The range of this loss function is [-1, 1].

(d) [2 Pts] Suppose Yuerou believes that wrongly predicting Barbie is worse than wrongly predicting Oppenheimer.

(i) Which quantity should she aim to **minimize**?

○ A. True Positives
○ B. False Positives
○ C. True Negatives
○ D. False Negatives

(ii) Which evaluation metric is the **BEST** option to **maximize** in this scenario?

○ A. Accuracy
○ B. Precision
○ C. True Positive Rate
○ D. False Positive Rate

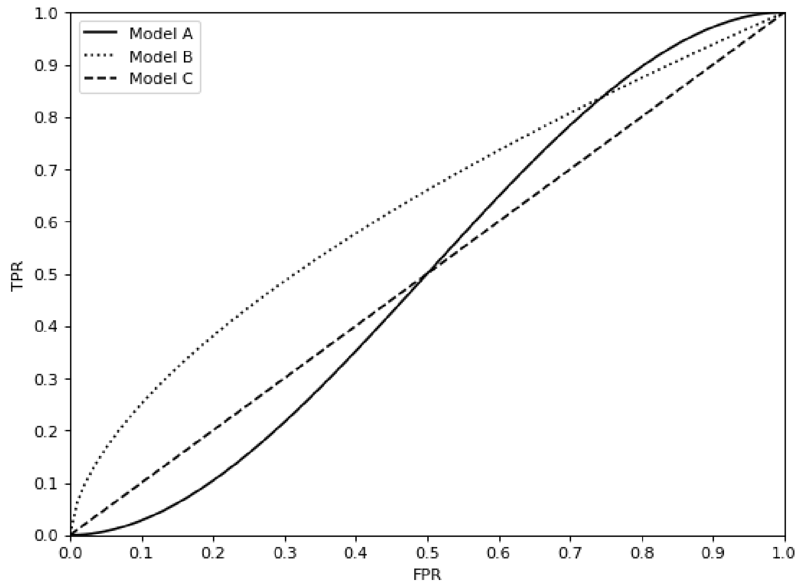
(e) [2 Pts] The table below shows a sample of validation data and predictions.

y_i	$\hat{P}(y = 1 \mid x_i)$
1	0.52
0	0.51
1	0.89
0	0.13
0	0.72
1	0.77

What is the range of classification thresholds that maximize accuracy while having no false positives on this sample of the validation set? Fill in this range's left and right bounds in the spaces below. Please keep your answers to 2 decimal places.

Threshold Range = (______ , ______)

(f) [2 Pts] Below are the ROC curves for 3 different models.



Suppose we fix the TPR to be 0.3.

(i) Which model would be most preferred?

◯ A. Model A
◯ B. Model B
◯ C. Model C
◯ D. Not enough information

(ii) Which model would be least preferred?

◯ A. Model A
◯ B. Model B
◯ C. Model C
◯ D. Not enough information

9 PieCe-A cake [8 Pts]

(a) [2 Pts] Milad wants to learn more about Principal Component Analysis (PCA). Which of the following statements are true? **Select all that apply.**

☐ A. Principal component vectors always have a mean of 0.

☐ B. PCA is sensitive to the scale of variables and can be influenced by outliers.

☐ C. The singular values along the diagonal of Σ can be used to explain how much variance is captured by a principal component.

☐ D. We should always pick the maximum number of principal components needed to capture all the model variance.

(b) [2 Pts] Use the options below to fill in the blanks and complete the following statement:

Milad trains an OLS model. As he increases the number of principal components used, the training loss will _____(i)_____ at first, then _____(ii)_____ later. The test loss will _____(iii)_____ at first, then _____(iv)_____ later.

(i) Fill in the Blank:

◯ A. Increase
◯ B. Mostly stays the same
◯ C. Decrease

(ii) Fill in the Blank:

◯ A. Greatly increase
◯ B. Very slightly increase
◯ C. Very slightly decrease
◯ D. Greatly decrease

(iii) Fill in the Blank:

◯ A. Increase
◯ B. Stay the same
◯ C. Decrease

(iv) Fill in the Blank:

◯ A. Increase
◯ B. Decrease
◯ C. Oscillate between increasing and decreasing
◯ D. Stay constant

(c) [2 Pts] Provided the full SVD of a square full-rank matrix X is written as $X = U\Sigma V^T$, which of the following matrixes are **always** symmetric? **Select all that apply.**

☐ A. $U^T X$

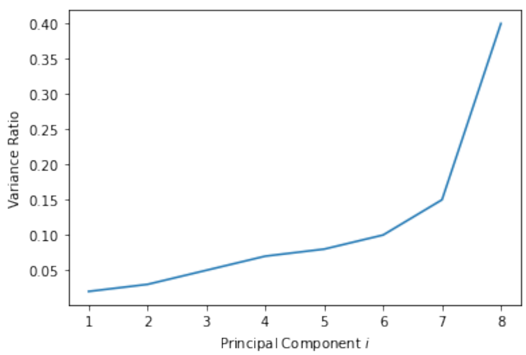
☐ B. XV

☐ C. $U^T XV$

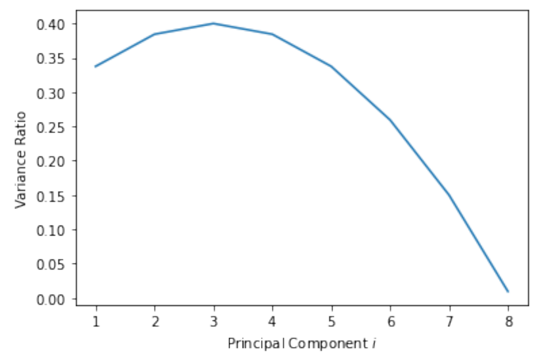
☐ D. $XU^T \Sigma V$

(d) [2 Pts] Milad makes the following plots:

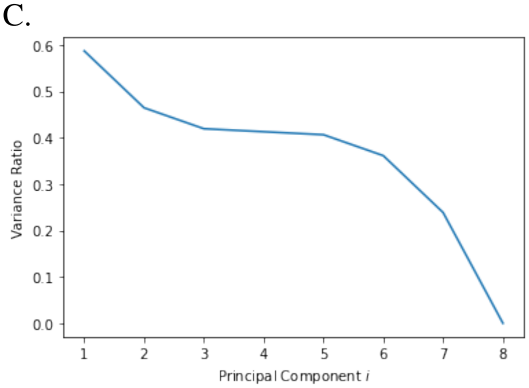
A.



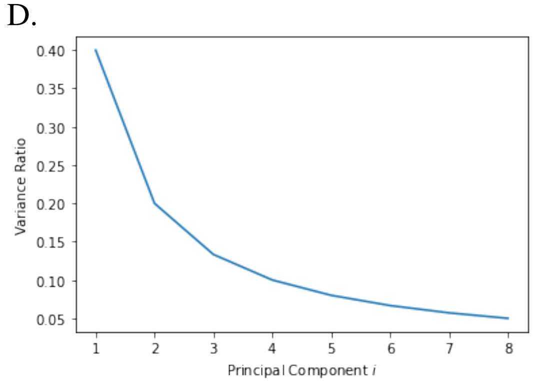
B.



C.



D.



Which of the above plots are **NOT** valid scree plots? **Select all that apply.**

☐ A

☐ B

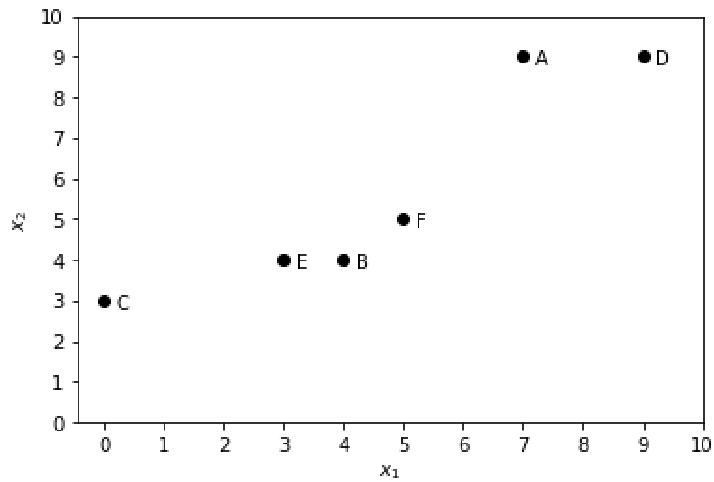
☐ C

☐ D

10 k-Nice [8 Pts]

(a) [4 Pts] Shiny has the following dataset, where each point has two features, x_1 and x_2:

x_1	x_2	Point Label
7	9	A
4	4	B
0	3	C
9	9	D
3	4	E
5	5	F



Shiny decides to perform k-means clustering on this dataset with 2 clusters: the left cluster has an initial center at (3, 3), and the right cluster has an initial center at (10, 10).

(i) Which points belong to the left cluster after the first iteration? Format your answer as a list of point labels separated by commas in alphabetical order.

(ii) Where are the clusters centered after the first iteration?

Left Cluster Center = (____ , ____)

Right Cluster Center = (____ , ____)

(iii) Which points belong to the left cluster after the second iteration? Format your answer as a list of point labels separated by commas in alphabetical order.

(iv) How many points do we expect to change clusters if Shiny were to perform a third iteration?

Number of Points = ☐

(b) [2 Pts] Suppose now, Shiny is working with the same dataset, but with new clusters. One cluster has points [B, C, E] and the other cluster has points [A, D, F]. For both parts, please write your answers in the given blanks.

(i) What is the distance between these two clusters by single linkage?

Distance = ____________________

(ii) What is the distance between these two clusters by complete linkage?

Distance = ____________________

(c) [2 Pts] Which of the following statements are **TRUE** regarding clustering? **Select all that apply.**

☐ A. If we run agglomerative clustering with the same k and the same type of linkage, we will get the same results each time.

☐ B. Distortion is a version of inertia weighted by cluster size.

☐ C. When choosing the optimal k for k-means clustering, we should pick the k with the lowest loss.

☐ D. Within one iteration of k-means clustering, some clusters' centers may move even if other centers do not.

You are done with the Final! Congratulations!

Use this page to draw your favorite Data 100 moment!