DATA 100 Fall 2020 Final-Exam

INSTRUCTIONS

This is your exam. Complete it either at exam.cs61a.org or, if that doesn't work, by emailing course staff with your solutions before the exam deadline.

This exam is intended for the student with email address `<EMAILADDRESS>`. If this is not your email address, notify course staff immediately, as each exam is different. Do not distribute this exam PDF even after the exam ends, as some students may be taking the exam in a different time zone.

For questions with **circular bubbles**, you should select exactly *one* choice.

○ You must choose either this option

○ Or this one, but not both!

For questions with **square checkboxes**, you may select *multiple* choices.

□ You could select this choice.

□ You could select this one too!

You may start your exam now. Your exam is due at <DEADLINE> Pacific Time. Go to the next page to begin.

Preliminaries

You can complete and submit these questions before the exam starts.

(a) What is your full name?

(b) What is your Berkeley email?

(c) What is your student ID number?

(d) When are you taking this exam?

○ Tuesday 7pm PST

○ Wednesday 8am PST

○ Other

(e) Honor Code: *All work on this exam is my own.*

By writing your full name below, you are agreeing to this code:

(f) Important: You must copy the following statement exactly into the box below. Failure to do so may result in points deducted on the exam.

"I certify that all work on this exam is my own. I acknowledge that collaboration of any kind is forbidden, and that I will face severe penalties if I am caught, including at minimum, harsh penalties to my grade and a letter sent to the Center for Student Conduct."

1. (a) (9.0 points)

Consider sampling students from the audience of a comedy show at UC Berkeley. The theater, which is currently at full capacity, is divided into three sections: Front, Middle, and Back. The following table contains the capacity of each section:

Section	Capacity
Front	20
Middle	35
Back	25

In the first two subparts of this question, we sample 5 students uniformly at random **with replacement**.

Ai. (1.0 pt) In our sample of 5 students, what is the expected number of students sitting in the middle?

○ $\frac{9}{4}$

○ $\frac{5}{4}$

○ $\frac{35}{16}$

○ $\frac{7}{16}$

○ $\frac{25}{16}$

○ None of the above

B. (2.0 pt) In our sample of 5 students, what is the probability that everyone is *not* in the same section? Select all that apply.

□ $\sum_{i=0}^{5} \left(\frac{1}{4}\right)^i \left(\frac{5}{16}\right)^i \left(\frac{7}{16}\right)^i$

□ $\left(\frac{1}{4}\right)^5 \left(\frac{5}{16}\right)^5 \left(\frac{7}{16}\right)^5$

□ $1 - \left(\frac{1}{4}\right)^5 - \left(\frac{5}{16}\right)^5 - \left(\frac{7}{16}\right)^5$

□ $1 - \sum_{i=0}^{5} \left(\frac{1}{4}\right)^i \left(\frac{5}{16}\right)^{5-i} \left(\frac{7}{16}\right)^{5-i}$

□ None of the above

ii. Consider the population of UC Berkeley students. We are interested in finding the expectation and variance of the number of students that have a driver's license in a sample from this population. We are given the following information:

- 70% of students are in-state and 30% of students are out-of-state
- 60% of in-state students have driver's licenses and 30% of out-of-state students have driver's licenses

We sample 120 students uniformly at random **with replacement**.

A. **(2.0 pt)** Define the random variable X_i to be 1 if the ith student in our sample has a driver's license, and 0 otherwise.

What is $P(X_i = 1)$? Please answer as a decimal rounded to two decimal places.

B. **(1.0 pt)** How many students do we expect to hold a driver's license in our sample? Your answer should be an algebraic expression involving *prevletter*, where *prevletter* is the correct answer to the previous part.

C. **(1.0 pt)** What is the variance of the number of students that hold a driver's license in our sample? Again, your answer should be an algebraic expression involving *prevletter*, as defined above.

D. **(2.0 pt)** In the previous two parts, we assumed that we were sampling with replacement. How would your answers to the above two parts change if we were instead sampling without replacement?

○ Expectation and variance would both stay the same

○ Expectation and variance would both be different

○ Expectation would stay the same while variance would be different

○ Expectation would be different while the variance would stay the same

(9.0 points)

Consider sampling students from the audience of a comedy show at UC Berkeley. The theater, which is currently at full capacity, is divided into three sections: Front, Middle, and Back. The following table contains the capacity of each section:

Section	Capacity
Front	35
Middle	20
Back	25

In the first two subparts of this question, we sample 5 students uniformly at random **with replacement**.

(b) Ai. (1.0 pt) In our sample of 5 students, what is the expected number of students sitting in the middle?

○ $\frac{9}{4}$

○ $\frac{5}{4}$

○ $\frac{35}{16}$

○ $\frac{7}{16}$

○ $\frac{25}{16}$

○ None of the above

B. (2.0 pt) In our sample of 5 students, what is the probability that everyone is *not* in the same section? Select all that apply.

□ $\sum_{i=0}^{5} \left(\frac{1}{4}\right)^i \left(\frac{5}{16}\right)^i \left(\frac{7}{16}\right)^i$

□ $\left(\frac{1}{4}\right)^5 \left(\frac{5}{16}\right)^5 \left(\frac{7}{16}\right)^5$

□ $1 - \left(\frac{1}{4}\right)^5 - \left(\frac{5}{16}\right)^5 - \left(\frac{7}{16}\right)^5$

□ $1 - \sum_{i=0}^{5} \left(\frac{1}{4}\right)^i \left(\frac{5}{16}\right)^{5-i} \left(\frac{7}{16}\right)^{5-i}$

□ None of the above

ii. Consider the population of UC Berkeley students. We are interested in finding the expectation and variance of the number of students that have a driver's license in a sample from this population. We are given the following information:

- 30% of students are in-state and 70% of students are out-of-state
- 20% of in-state students have driver's licenses and 80% of out-of-state students have driver's licenses

We sample 150 students uniformly at random **with replacement**.

A. (2.0 pt) Define the random variable X_i to be 1 if the ith student in our sample has a driver's license, and 0 otherwise.

What is $P(X_i = 1)$? Please answer as a decimal rounded to two decimal places.

B. (1.0 pt) How many students do we expect to hold a driver's license in our sample? Your answer should be an algebraic expression involving *prevletter*, where *prevletter* is the correct answer to the previous part.

C. (1.0 pt) What is the variance of the number of students that hold a driver's license in our sample? Again, your answer should be an algebraic expression involving *prevletter*, as defined above.

D. (2.0 pt) In the previous two parts, we assumed that we were sampling with replacement. How would your answers to the above two parts change if we were instead sampling without replacement?

○ Expectation and variance would both stay the same

○ Expectation and variance would both be different

○ Expectation would stay the same while variance would be different

○ Expectation would be different while the variance would stay the same

2. (6.0 points)

Throughout this question, we are dealing with pandas DataFrame and Series objects. All code for this question, where applicable, must be written in Python. You may assume that `pandas` has been imported as `pd`.

The following DataFrame `cars` contains the names of car models from 1970 to 1982. The `name` column is the primary key of the table.

The first five rows are shown below.

name	mpg	horsepower	weight	acceleration	year	origin	brand
toyota corolla 1200	32.0	65	1836	21.0	1974	Japan	toyota
buick skylark 320	15.0	165	3693	11.5	1970	USA	buick
fiat 128	29.0	49	1867	19.5	1973	Europe	fiat
ford mustang gl	27.0	86	2790	15.6	1982	USA	ford
ford torino	17.0	140	3449	10.5	1970	USA	ford

(a) (2.0 pt) Below, write a line of Pandas code that creates a **Series** of the **names** of cars created by brand "carbrand" with greater than mpgnum mpg. The resulting Series should be assigned to the variable `varname`.

(b) (4.0 pt) Below, write a line of Pandas code to create a **DataFrame** containing data only for those car models whose brands have at least mpgnum2 mpg for **each** of their models. The resulting DataFrame must have the same structure and format as `cars`. The resulting DataFrame should be assigned to the variable `varname2`.

3. (8.0 points)

In this question, we're interested in finding the number of classes taken by students at Zoom University. We will be working with two DataFrames, `students` and `enrollment`. Throughout this question, you may assume that `pandas` has been imported as `pd`.

Each row in the `students` DataFrame represents a student. The `students` DataFrame contains the following columns:

- student_name: the student's name
- SID: the student's ID
- major: the student's major

Here are the first four rows in `students`:

	student_name	SID	major
0	Alice Red	123	Computer Science
1	Bob Lime	128	Biology
2	Susie Orange	209	Anthropology
3	Frank Blue	212	History

Each row in the `enrollment` DataFrame represents an enrollment record for a specific student in a single class. If a student is enrolled in multiple classes, each class taken by the student is a separate row in `enrollment`. The `enrollment` DataFrame contains the following columns:

- SID: the student's ID
- class_name: the name of the **class** the student is enrolled in
- class_id: the ID of the class

Here are the first five rows in `enrollment`:

	SID	class_name	class_id
0	123	Intro to Data Science	200
1	128	Organic Chemistry	145
2	128	Intro to Data Science	100
3	209	US History	185
4	212	US History	185

Note: It is possible for rows with different `class_id` to share the same `class_name` in the `enrollment` DataFrame. For example, there is an "Intro to Data Science" with `class_id` 100 and another "Intro to Data Science" with `class_id` 200.

(a) **(4.0 pt)** Suppose you are asked to add a column `num_class` to the `students` DataFrame that indicates the number of classes each student is enrolled in. If a student does not have any enrollment records, they should have a value of 0 in `num_class`. You are allowed to change the index of `students`, but the number of rows should stay the same after adding the column, and the `name` and `major` columns should be kept the same.

Which of the following accomplishes this task? There is only one correct answer.

A:

```
num_class = students.merge(enrollment, left_on='student_name', right_on='class_name', how='right')
                    .groupby('SID').count()
num_class = num_class.drop(columns=['class_name', 'student_name', 'major'])
num_class = num_class.rename(columns={'class_id': 'num_class'})
students = students.merge(num_class, left_on='SID', right_index=True)
```

B:

```
num_class = enrollment.groupby('SID').count()
num_class = num_class.set_index('SID')
num_class = num_class.rename(columns={'class_id': 'num_class'})
students['num_class'] = num_class['class_id']
```

C:

```
num_class = students.merge(enrollment, on='SID', how='outer').groupby('SID').count()
num_class = num_class.drop(columns=['class_name', 'student_name', 'major'])
num_class = num_class.rename(columns={'class_id': 'num_class'})
students = students.merge(num_class, left_on='SID', right_index=True)
```

○ A

○ B

○ C

○ None of the above

(b) (4.0 pt) Now you are asked to find all unique majors across all students enrolled in `Intro to Data Science`. Specifically, you need to create a Series `ds_majors` that has majors as the index and the counts of students enrolled in `Intro to Data Science` in each major as the values.

Which of the following accomplishes this task? There is only one correct answer.

A:

```
ds = enrollment[enrollment['class_name'] == 'Intro to Data Science']
ds_majors = ds.merge(students, on='SID', how='outer').groupby('major')['SID'].count()
```

B:

```
ds = enrollment[enrollment['class_name'] == 'Intro to Data Science']
ds_majors = ds.merge(students, on='SID', how='left').groupby('major')['SID'].count()
```

C:

```
major_count = students.groupby('major').count()
merged = enrollment.merge(major_count, on='SID')
ds = merged[merged['class_name'] == 'Intro to Data Science']
ds_majors = ds['major']
```

○ A

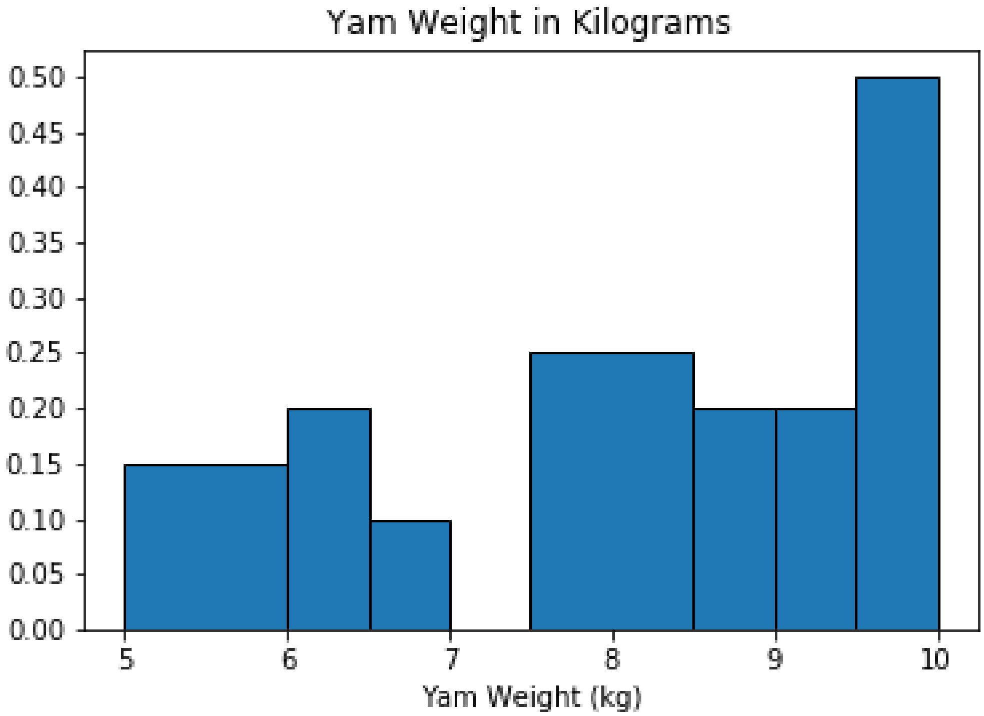
○ B

○ C

○ None of the above

4. (6.0 points)

A biology class grows and weighs yams as part of a class project. Some yams were grown in hot water and some were grown in cold water. A student, Shirley, decides to create a histogram of the yam weights.



(a) (2.0 pt) Professor Kane decides that yams weighing between 8 and 9 kilograms are his favorite. What percentage of yams weigh between 8 and 9 kilograms?

○ 20%

○ 25%

○ 30%

○ 35%

○ Impossible to tell

(b) (2.0 pt) Another student, Jeff, suspects that the the yams grown in hot water didn't grow as well as the yams grown in cold water and as such ended up weighing less. If 20 yams were grown in total and weighed, how many yams weigh less than 7 kilograms?

○ 3

○ 4

○ 5

○ 6

○ Impossible to tell

(c) (2.0 pt) A third student, Todd, wants to compare the maximum bin (9.5 to 10 kilograms) with the median bin (7.5 to 8.5 kilograms). Which bin contains more yams?

○ Median bin (7.5 to 8.5 kg bin)

○ Maximum bin (9.5 to 10 kg bin)

○ They contain the same number of yams

○ Impossible to tell

5. (15.0 points)

(a) Suppose we have the following dataset from the neighborhood CVS store on Shattuck. The table shows total rain (mm) for each quarter and total number of umbrellas sold for each quarter. **Note:** For the first three parts of this question, our dataset only has these four rows.

Quarter	Total rain (mm)	Total number of umbrellas sold
Jan-Mar	300	200
Apr-Jun	50	40
Jul-Sep	10	10
Oct-Dec	200	100

i. (2.0 pt) We first decide to model umbrella sales using the constant model $\hat{y} = \theta$. We will use squared loss as our loss function (no regularization).

Which expression below correctly gives the average loss of our fitted model on the given dataset? **Select the closest answer**.

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - 87.5x_i)^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - 140)^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - 87.5)^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - 140x_i)^2$

ii. (3.0 pt) Now we decide to fit a simple linear model with an intercept term $\hat{y} = \theta_0 + \theta_1 x$ that predicts total number of umbrellas sold (y) given total rain (x). We will use squared loss as our loss function, and we will not use regularization.

We are given $r = 0.979$, $\sigma_x = 116.40$, and $\sigma_y = 72.59$, which are the correlation coefficient, standard deviation of x, and standard deviation of y, respectively.

Which expression below correctly gives the average loss of our fitted model on the given dataset? **Select the closest answer**.

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - (10 + 0.61x_i))^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - (0.61 + 2x_i))^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - (2.57 + 1.57x_i))^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(y_i - (2 + 0.61x_i))^2$

iii. (3.0 pt) For whatever reason, we decide to reverse our model. That is, we decide to predict total rain (x) given total number of umbrellas sold (y) using a simple linear model with an intercept term $\hat{x} = \theta_0 + \theta_1 y$. Again, we will use squared loss as our loss function, and we will not use regularization.

Which expression below correctly gives the average loss of our fitted model on the given dataset? **Select the closest answer**.

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(x_i - (10 + 1.57y_i))^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(x_i - (0.61 + 2y_i))^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(x_i - (2.57 + 1.57y_i))^2$

○ $R(\hat{\theta}) = \frac{1}{4}\sum_{i=1}^{4}(x_i - (2 + 0.61y_i))^2$

(b) Now, we are back to predicting total number of umbrellas sold (y). For the remainder of the question, **assume that we have many more rows of data, not just the four given originally.**

In the first part of this question, we didn't use the `Quarter` column. Let's suppose we want to one-hot encode Quarter for use in our model, but with a twist - we only want to encode whether or not the current Quarter is Jul-Sep, since that's when rainfall is at a low.

The resulting design matrix, along with an intercept column, is provided below. (Note, the "Total number of umbrellas sold" column is no longer visible since it's not part of our design matrix.)

Intercept	Quarter=Jul-Sep	Quarter!=Jul-Sep	Total rain (mm)
1	0	1	300
1	0	1	50
1	1	0	10
⋮	⋮	⋮	⋮
1	0	1	200

We fit two different linear models using ordinary least squares, both of which use a subset of the columns of the above design matrix:

- We fit a linear model on all columns except `Quarter!=Jul-Sep`. After doing so, we end up with the following fitted model, where our optimal model parameter is $\hat{\theta} = [letter1, letter2, letter3]^T$:

$$\hat{y} = letter1 + letter2 \cdot (\text{Quarter=Jul-Sep}) + letter3 \cdot (\text{Total rain})$$

- We fit a linear model on all columns except `Quarter=Jul-Sep`. After doing so, we end up with the following fitted model, where our optimal model parameter is $\hat{\beta} = [D, E, F]^T$:

$$\hat{y} = D + E \cdot (\text{Quarter!=Jul-Sep}) + F \cdot (\text{Total rain})$$

In this problem, you will express D, E, and F in terms of $letter1$, $letter2$, and $letter3$. Your answers should all be algebraic expressions, for instance "100 * $letter1$ * $letter2$ * $letter3$" (that is not the correct answer to any of these parts). **If you don't believe it's possible to determine the answer, just write "not possible".**

i. (2.0 pt) What is D in terms of $letter1$, $letter2$, and $letter3$?

ii. (2.0 pt) What is E in terms of $letter1$, $letter2$, and $letter3$?

iii. (2.0 pt) What is F in terms of $letter1$, $letter2$, and $letter3$?

iv. (1.0 pt) Suppose we now regularize the previous two models using L_2 regularization with some fixed value of $\lambda > 0$.

We denote the optimal regularized model parameters by $\hat{\theta}_{\text{ridge}}$ and $\hat{\beta}_{\text{ridge}}$, corresponding to the first and second models in the previous part, respectively. All three of our features, including our intercept term, are regularized.

True or False: The relationships involving D, E, F, $letter1$, $letter2$, and $letter3$ from the previous part still hold true, even though our model is now regularized.

○ True

○ False

6. (7.0 points)

(a) In class, we derived the following bias-variance decomposition under a specific set of conditions.

$$\text{model risk} = \sigma^2 + (\text{model bias})^2 + \text{model variance}$$

We assume that there is an unknown underlying function $g(x)$ that generates the points we observe. Specifically, we observe $Y_i = g(x_i) + \epsilon_i$, where ϵ_i is a zero-mean noise term with variance σ^2 that is independent for each observation. Our model's goal is to approximate $g(x)$ as best as possible.

i. (1.0 pt) Does this decomposition hold true for linear models and squared loss?

○ Yes

○ No

ii. (1.0 pt) Does this decomposition hold true for non-linear models and squared loss?

○ Yes

○ No

iii. (1.0 pt) Does this decomposition hold true for linear models and absolute loss?

○ Yes

○ No

iv. (1.0 pt) Does this decomposition hold true for classification decision trees and zero-one loss?

(Zero-one loss is equal to 1 if a prediction is correct, and 0 if it is incorrect.)

○ Yes

○ No

(b) Recall that we discussed the technique of pruning a decision tree, which involves removing certain branches. What effect does pruning a decision tree have on its

i. (1.0 pt) Bias?

○ Increases it

○ Decreases it

ii. (1.0 pt) Variance?

○ Increases it

○ Decreases it

iii. (1.0 pt) Complexity?

○ Increases it

○ Decreases it

○ Depends on the splitting rule

7. (4.0 points)

For each of the following prompts, answer true if the given modification to k-fold cross-validation will result in overfitting, and false if it will not. Assume that we have a large dataset that we have split into a training set and test set.

(a) (1.0 pt) The test set is divided into k folds. For each fold of the test set, we use the entire training set to train the model, and use the given fold/subset of the test set for validation. The average error among all k folds is the cross-validation error.

True or False: This modification will result in overfitting.

○ True

○ False

(b) (1.0 pt) We use normal k-fold cross-validation, but for each fold we only use half of the validation set for validation.

True or False: This modification will result in overfitting.

○ True

○ False

(c) (1.0 pt) We use normal k-fold cross-validation, but for each fold we use the entire training set for training.

True or False: This modification will result in overfitting.

○ True

○ False

(d) (1.0 pt) We use normal k-fold cross-validation, but after the train-test split, we standardize the training set before running cross-validation so that each column has mean 0 and variance 1.

True or False: This modification will result in overfitting.

○ True

○ False

8. (14.0 points)

Consider the following model:

$$f_\theta(x) = \theta_0 + 2^{\theta_1} x + \theta_1 \theta_2 x^2$$

We have a training dataset with two observations (x_i, y_i): $\{(1, 1), (2, 3)\}$.

In order to determine optimal model parameters $\hat{\theta}_0$, $\hat{\theta}_1$, and $\hat{\theta}_2$, we choose squared loss with L_2 regularization. Assume that the regularization hyperparameter $\lambda = \frac{1}{2}$ for the entirety of this question, and assume that we regularize the intercept term θ_0. Our objective function is the sum of our loss function averaged across our entire dataset and a regularization penalty.

We decide to use gradient descent to help us solve for the optimal parameters.

(a) **(3.0 pt)** Which of the following is equal to the objective function for our model, loss, regularization, and training data?

○ $$R(\theta) = \left[(\theta_0 + 2^{\theta_1} + \theta_1\theta_2 - 1)^2 + (\theta_0 + 2^{\theta_1+1} + 4\theta_1\theta_2 - 3)^2\right] + \frac{1}{2}(\theta_0^2 + \theta_1^2 + \theta_2^2)$$

○ $$R(\theta) = \frac{1}{2}\left[(1 - (\theta_0 + 2^{\theta_1} + \theta_1\theta_2))^2 + (3 - (\theta_0 + 2^{\theta_1+1} + 4\theta_1\theta_2))^2 + |\theta_0|^2 + |\theta_1| + |\theta_2|\right]$$

○ $$R(\theta) = \frac{1}{2}\left[(1 - (\theta_0 + 2^{\theta_1} + \theta_1\theta_2))^2 + (3 - (\theta_0 + 2^{\theta_1+1} + 4\theta_1\theta_2))^2\right] + 2(\theta_1^2 + \theta_2^2)$$

○ $$R(\theta) = \frac{1}{2}\left[(\theta_0 + 2^{\theta_1} + \theta_1\theta_2 - 1)^2 + (\theta_0 + 2^{\theta_1+1} + 4\theta_1\theta_2 - 3)^2 + \theta_0^2 + \theta_1^2 + \theta_2^2\right]$$

Suppose we start our gradient descent procedure at the initial guess $\theta^{(0)} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, where a, b, c are some constants.

Then, $\left.\frac{\partial R}{\partial \theta_0}\right|_{\theta=\theta^{(0)}}$, the partial derivative of our objective function with respect to θ_0 evaluated at our initial guess $\theta^{(0)}$, is of the form

$$Ga + H \cdot 2^b + 5bc - 4$$

where G and H are integers.

(b) **i.** **(3.0 pt)** What is G?

○ -3

○ -2

○ -1

○ 0

○ 1

○ 2

○ 3

ii. **(3.0 pt)** What is H?

○ -3

○ -2

○ -1

○ 0

○ 1

○ 2

○ 3

(c) (1.0 pt) Recall that our model is $f_\theta(x) = \theta_0 + 2^{\theta_1} x + \theta_1 \theta_2 x^2$, or equivalently $f_\theta(x_i) = \theta_0 + 2^{\theta_1} x_i + \theta_1 \theta_2 x_i^2$.

Suppose we define $\gamma = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix}$ such that

$$\gamma_0 = \theta_0, \gamma_1 = 2^{\theta_1}, \gamma_2 = \theta_2$$

Can we use ridge regression to find $\hat{\gamma}$?

○ Yes

○ No

(d) (1.0 pt) Suppose our model is instead

$$f_\theta(x_i) = \theta_0 + 2^{\theta_1} x_{i,1} + \theta_2 x_{i,1} \cdot x_{i,2}$$

where $x_{i,1}$ and $x_{i,2}$ are scalars corresponding to feature 1 and feature 2 for observation i, respectively. Let γ be as defined in the previous part.

Can we use ridge regression to find $\hat{\gamma}$?

○ Yes

○ No

(e) **(3.0 pt)** Note: This part is independent of the previous parts of this question.

Below is a buggy implementation of `sgd`, a function which is supposed to perform stochastic gradient descent with batch size `B` on the training dataset `X` and `y` by applying the gradient `gradient_function` with learning rate `alpha`.

```
def sgd(X, y, theta0, gradient_function, alpha, B, max_iter=100000):
    """
    Performs stochastic gradient descent.

    Args:
        X: A 2D array, the dataset, with features stored in columns
           and observations stored in rows
        y: A 1D array, the outcome values
        theta0: A 1D array, the initial weights
        gradient_function: A function that takes in a vector
            of weights, a dataset, and outcome values and
            returns the value of the gradient
        alpha: A float, the learning rate
        B: An integer, the batch size
        max_iter (optional): The maximum number of iterations
            to attempt during SGD

    Returns:
        A 1D array of optimal weights

    Notes:
        gradient_function takes 3 arguments: a 1D array of weights,
        a 2D array of data points, and a 1D array of outcomes. It
        returns a 1D array of the same shape as the weights, the
        value of the gradient evaluated with those parameters.
    """

    theta = theta0
    for _ in range(max_iter):
        idx = np.random.choice(X.shape[1], size=B, replace=True)
        Xb, yb = X[idx,:], y[idx]
        grad = gradient_function(theta, Xb, yb)
        theta = theta - alpha*grad
    return theta
```

Which of the following edits need to be made to the implementation of `sgd` above so that it works correctly (as specified in class)? Select all that apply.

□ `X.shape[1]` should be replaced with `X.shape[0]`

□ `size=B` should be replaced with `size=X.shape[0]`

□ `replace=True` should be replaced with `replace=False`

□ `theta - alpha*grad` should be replaced with `theta + alpha*grad`

□ `gradient_function(theta, Xb, yb)` should be replaced with `gradient_function(theta, X, y)`

□ `X[idx,:], y[idx]` should be replaced with `X[:, idx], y`

□ None of the above

9. (13.0 points)

In this problem, we'll be using logistic regression to build a classifier that differentiates between 2 varieties of wine produced in the same region of Italy.

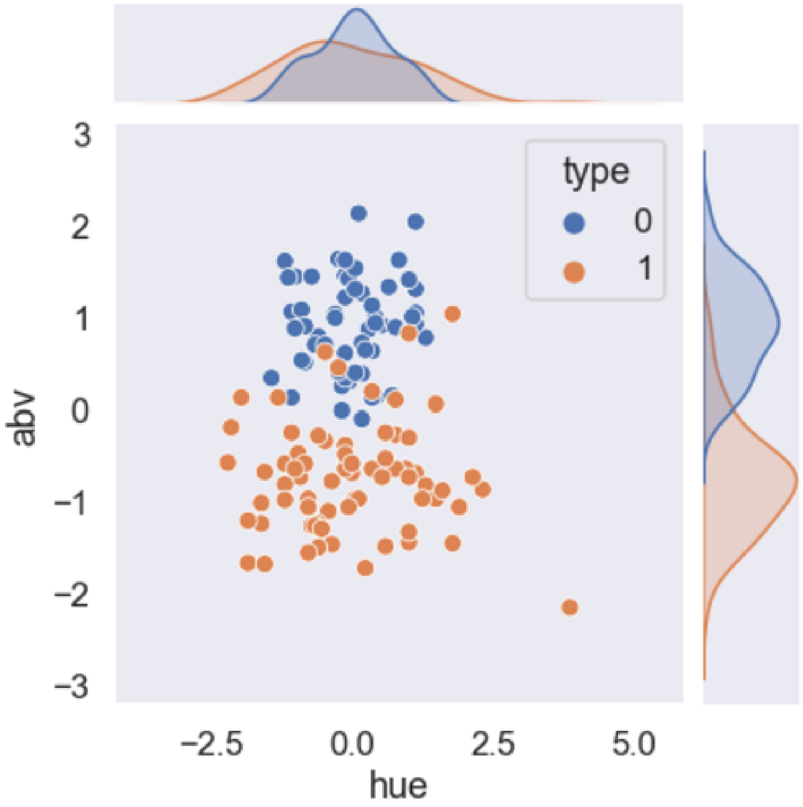
In this problem, assume the following:

- We are working with a design matrix $\mathbb{X}$ with two features: the hue of the wine (hue, x_1) and its alcohol by volume (abv, x_2). Note that both hue and abv are quantitative (hue is a quantitative measure of a wine's color).
- $\mathbb{X}$ is standardized.
- All wines are either type 0 or 1 (y).

We are modeling the probability that a particular wine is of type 1 using

$$P(Y = 1|x) = \sigma(\theta_1 \cdot \text{hue} + \theta_2 \cdot \text{abv})$$

(a) (2.0 pt) Consider the following scatter plot of our two (standardized) features. Note, this scatter plot is only relevant in this subpart of the question.



Which of the following statements are true about an unregularized logistic regression model fit on the above data? Select all that apply.

□ After performing logistic regression, the weight for the hue feature will very likely have a negative sign.

□ After performing logistic regression, the weight for the abv feature will very likely have a negative sign.

□ After performing logistic regression, the abv feature will have very likely a higher magnitude weight than the hue feature.

□ This data is linearly separable between the two wine types without any feature transformations.

(b) (2.0 pt) Consider the following three rows from our training data, along with their predicted probabilities $\hat{y}$ for some choice of θ:

hue	abv	y	$\hat{y}$
-0.17	0.24	0	0.45
-1.18	1.61	0	0.19
1.25	-0.97	1	0.80

What is the mean cross-entropy loss on just the above three rows of our training data?

○ $-\frac{1}{3}\big(\log(0.45) + \log(0.19) + \log(0.20)\big)$

○ $-\frac{1}{3}\big(\log(0.55) + \log(0.19) + \log(0.80)\big)$

○ $-\frac{1}{3}\big(\log(0.45) + \log(0.81) + \log(0.80)\big)$

○ $-\frac{1}{3}\big(\log(0.55) + \log(0.81) + \log(0.80)\big)$

○ None of the above

(c) (3.0 pt) After thresholding $\hat{y}$, we compute a confusion matrix for our model's predictions. As a reminder, type 0 and type 1 refer to wine types.

	Predicted Type 0	Predicted Type 1
Actual Type 0	57	???
Actual Type 1	???	62

For some reason, our confusion matrix is corrupted, and doesn't contain the information on the off-diagonals. However, we somehow know that our model's accuracy is $\frac{119}{130}$ and our model's precision is $\frac{31}{32}$.

What is our model's recall? Give your answer as a reduced fraction with no spaces, i.e. in the form a/b (no decimals or spaces).

(d) Suppose we choose $\hat{\theta} = [2, 1]^T$. Consider the wine "Billywine" with hue $\frac{1}{4}$ and abv -2.

i. (2.0 pt) Let β be the odds that Billywine is a type 1 wine under our model. What is β? There is only one correct answer.

○
$$\beta = \frac{3}{2}$$

○
$$\beta = -\frac{3}{2}$$

○
$$\beta = e^{\frac{3}{2}}$$

○
$$\beta = e^{-\frac{3}{2}}$$

○
$$\beta = \sigma(-\frac{3}{2})$$

○
$$\beta = \log\left(\frac{-\frac{3}{2}}{1+\frac{3}{2}}\right)$$

ii. (2.0 pt) Let γ be the probability that Billywine is a type 1 wine under our model. What is γ? Select all that apply. (β is as defined in the previous subpart.)

□
$$\gamma = e^{-\frac{3}{2}}$$

□
$$\gamma = \sigma(-\frac{3}{2})$$

□
$$\gamma = \sigma(\frac{3}{2})$$

□
$$\gamma = \frac{\beta - 1}{\beta}$$

□
$$\gamma = \frac{\beta}{\beta + 1}$$

iii. (2.0 pt) Suppose that we choose a threshold T such that the decision boundary of our model is $2 \cdot \text{hue} + \text{abv} = \frac{3}{2}$. What value of T results in this decision boundary? There is only one correct answer. (β and γ are as defined in the previous two subparts.)

○ $T = \beta$

○ $T = e^{-\gamma}$

○ $T = \gamma$

○ $T = -\beta$

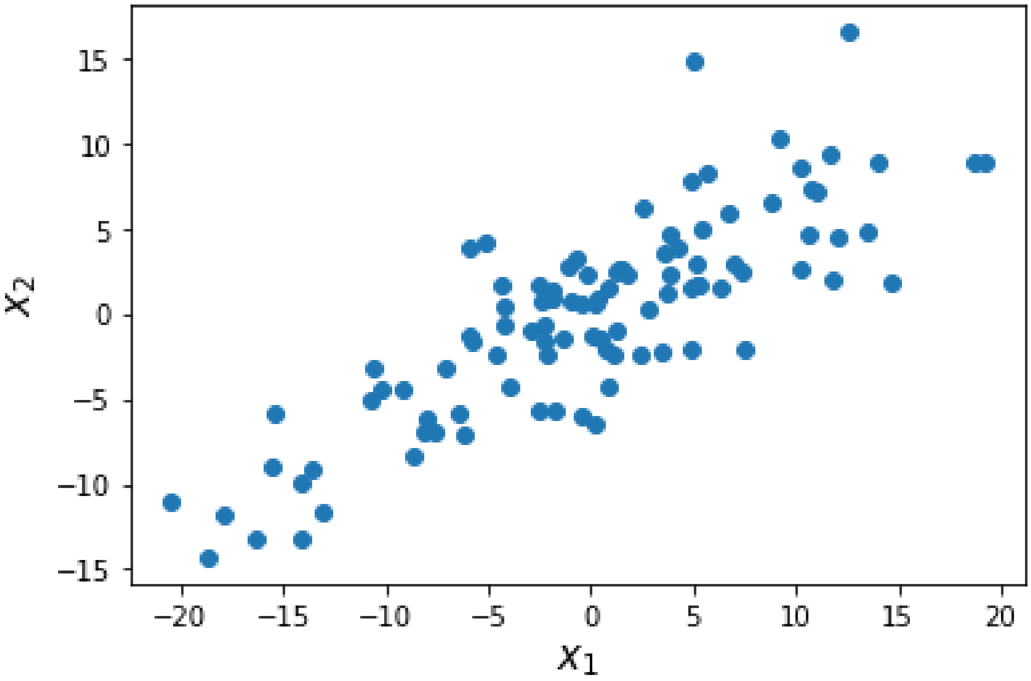
○ $T = 1 - \beta$

○ $T = \log(\frac{\gamma}{1-\gamma})$

○ $T = 1 - \gamma$

10. (7.0 points)

(a) Suppose we are given the following scatter plot.



We have data that is plotted in the space of features x_1 and x_2. Suppose we want to perform PCA on these two features.

i. (1.0 pt) Which of the following is most likely to be the equation of the line representing PC 1?

○ $$x_2 = \frac{11}{3}x_1 - 9$$

○ $$x_2 = 3x_1$$

○ $$x_2 = -\frac{20}{3}x_1 + 5$$

○ $$x_2 = \frac{2}{3}x_1$$

○ $$x_2 = -3x_1$$

ii. (1.0 pt) Which of the following is most likely to be the equation of the line representing PC 2?

○
$$x_2 = -3x_1$$

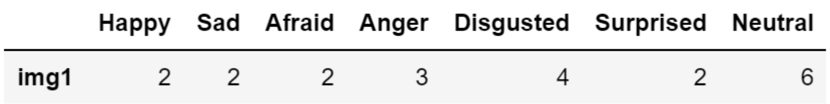
○
$$x_2 = \frac{1}{3}x_1 + 5$$

○
$$x_2 = -4x_1 + 10$$

○
$$x_2 = 3x_1$$

○
$$x_2 = -\frac{3}{2}x_1$$

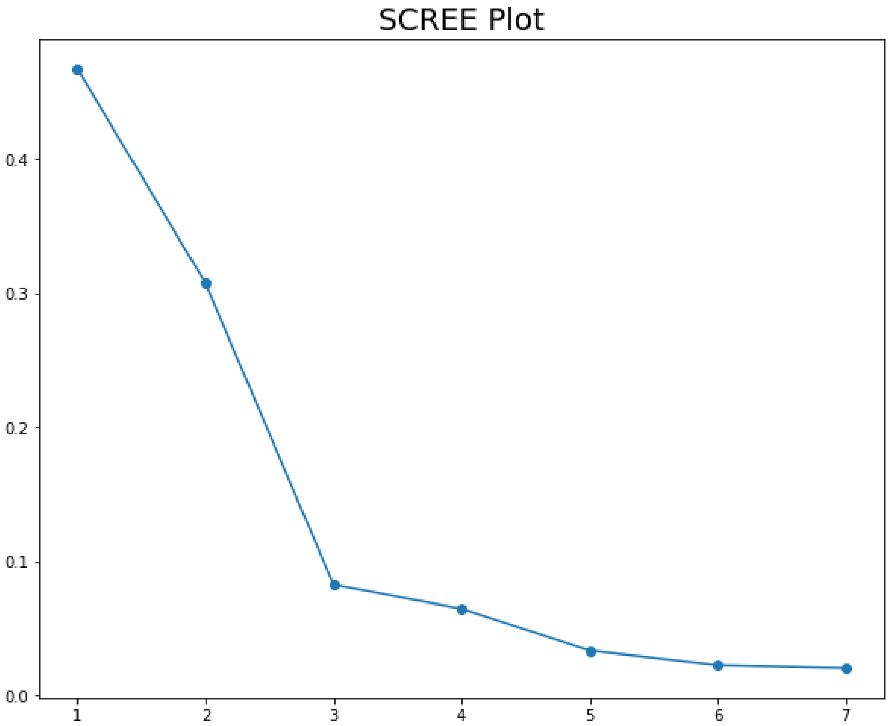
(b) In this part of this question, we will look at emotion ratings of images for a psychology experiment. Each row of the DataFrame F represents an image, and each column represents an emotion. There are **940 images and 7 emotions**. An example row of F is provided below.



	Happy	Sad	Afraid	Anger	Disgusted	Surprised	Neutral
img1	2	2	2	3	4	2	6

Say we perform the SVD on F using the following code:

```
X = (F - np.mean(F, axis = 0))
u, s, vt = np.linalg.svd(X, full_matrices=False)
```



i. (1.0 pt)

The above scree plot depicts the proportion of variance captured by each PC. Ignoring the plot's title, which of the following lines of code could have created the above plot?

○ `plt.plot(s**2/np.sum(s**2), u)`

○ `plt.plot(F[:, :7]),  s**2/np.sum(s))`

○ `plt.plot(np.arange(1, F.shape[1]+1), s**2/np.sum(s**2))`

○ `plt.plot(np.arange(1, F.shape[1]+1), s**2/np.sum(s))`

○ `plt.plot(u@s, s**2/np.sum(s**2))`

ii. (2.0 pt) Suppose we know that `np.sum(s**2)` evaluates to 121. Which of the following is closest to `s[1]`?

○ 0.3

○ 3.3

○ 6

○ 8

○ 36

iii. (1.0 pt) Which of the following statements evaluates to `True`?

○ `(u @ np.diag(s)).shape == (940, 7)`

○ `(u @ np.diag(s)).shape == (7, 7)`

○ `(u @ np.diag(s)).shape == (940, 940)`

○ `(u @ np.diag(s)).shape == (7, 940)`

○ None of the above

iv. (1.0 pt) True or False: Ignoring numerical precision issues, the expression

`np.var((X @ vt.T)[:, i]) == s[i]**2 / len(X)`

evaluates to `True` for all integers `i` between `0` and `X.shape[1] - 1`.

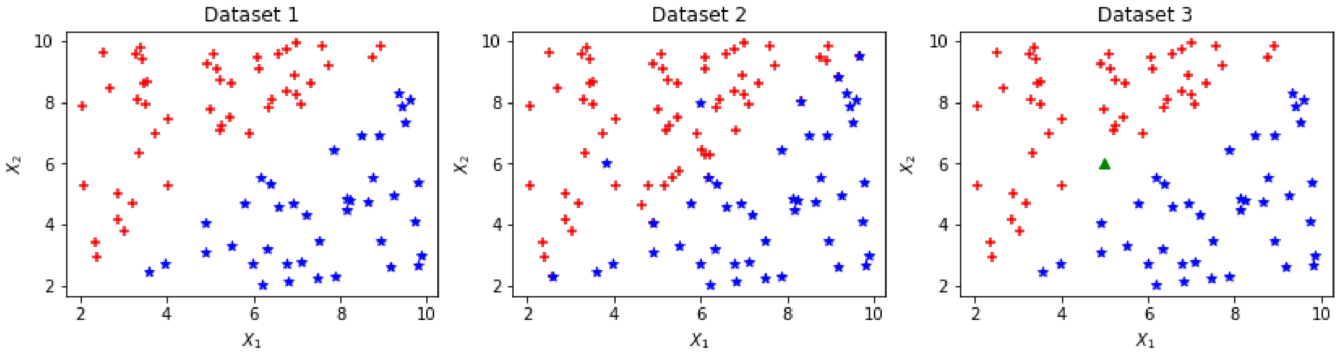
○ True

○ False

○ Impossible to tell

11. (8.0 points)

(a) Consider the following three datasets, each consisting of two features (x_1 and x_2) and a class label (red crosses and blue stars).



The green triangle in Dataset 3 represents a point with an overlapping red cross and blue star point at the same position. Assume that otherwise, there are no overlapping points of different classes in any of the above datasets.

i. (2.0 pt) On which of the above datasets could logistic regression (fit with no regularization) achieve 100% training accuracy? Select all that apply.

□ Dataset 1

□ Dataset 2

□ Dataset 3

□ None of the above

ii. (2.0 pt) On which of the above datasets could a decision tree achieve 100% training accuracy? Select all that apply.

□ Dataset 1

□ Dataset 2

□ Dataset 3

□ None of the above

iii. (2.0 pt) On which of the above datasets could a random forest achieve 100% training accuracy? Select all that apply.

□ Dataset 1

□ Dataset 2

□ Dataset 3

□ None of the above

(b) (2.0 pt) Suppose we have a training dataset with $n = 2^6$ observations, consisting of some design matrix $\mathbb{X}$ and binary response variable y, and we want to train a binary classifier.

The all-zero classifier is a classifer that predicts 0 for all observations, regardless of input. The training accuracy of the all-zero classifier on our training data is $\frac{1}{8}$.

If we were to build a decision tree for classification, what would be the entropy of the tree at the root node, where all observations begin?

○ $-\frac{1}{8}\left[7\log_2\frac{1}{8} + \log_2\frac{7}{8}\right]$

○ $-\frac{1}{8}\left[\log_2\frac{1}{8} + \log_2\frac{7}{8}\right]$

○ $-\frac{1}{64}\left[8\log_2\frac{1}{8} + 56\log_2\frac{7}{8}\right]$

○ $-\frac{1}{8}\left[\log_2\frac{1}{8} + 7\log_2\frac{7}{8}\right]$

○ $-8\left[\log_2\frac{1}{8} + 7\log_2\frac{7}{8}\right]$

○ Impossible to tell

12. (7.0 points)

Consider a DataFrame `people` containing the height, weight, and BMI (body mass index) of several individuals. Our dataset has three columns:

- `height (cm)`: Height in centimeters
- `weight (kg)`: Weight in kilograms
- `bmi`: Body Mass Index, calculated as

```
people['bmi'] = people['weight (kg)'] / (people['height (cm)'] / 100) ** 2
```

The first five rows of `people` might look something like:

height (cm)	weight (kg)	bmi
185.42	109.545	31.8626
172.72	73.6364	24.6835
187.96	96.3636	27.2761
180.34	100	30.7479
175.26	93.6364	30.4845

(a) (2.0 pt) Let `r(x, y)` be a function that computes the correlation coefficient r for two Series of numbers `x` and `y`.

Suppose, just for this part, that the values in `height (cm)` and `weight (kg)` are generated using an uncorrelated random number generator (that is, `r(people['height (cm)'], people['weight (kg)']) == 0`).

What is the most likely value of `R = r(people['height (cm)'], people['bmi'])`?

○ `R < -0.2`

○ `-0.2 <= R < 0.2`

○ `R >= 0.2`

(b) (2.0 pt) For whatever reason, we decide to add Imperial units to our dataset, which we will now call `humans`. That is, we add the columns `height (in)` and `weight (lb)`, where `humans['height (in)'] = humans['height (cm)'] / 2.54` and `humans['weight (lb)'] = humans['weight (kg)'] * 2.2`.

The first five rows of `humans` might look something like:

height (in)	height (cm)	weight (lb)	weight (kg)	bmi
73	185.42	241	109.545	31.8626
68	172.72	162	73.6364	24.6835
74	187.96	212	96.3636	27.2761
71	180.34	220	100	30.7479
69	175.26	206	93.6364	30.4845

Which of the following sets of columns are linearly independent and have a span that is equal to the span of the columns of `humans`? Select all that apply.

□ `height (in)`, `height (cm)`, `weight (lb)`, `weight (kg)`, `bmi`

□ `height (in)`, `weight (lb)`, `bmi`

□ `height (cm)`, `weight (lb)`, `bmi`

□ `height (in)`, `height (cm)`, `weight (lb)`, `bmi`

□ `height (cm)`, `bmi`

□ None of the above

(c) **(2.0 pt)** Now suppose we fit two linear models on the `humans` data.

Model A:

$$\hat{\text{bmi}} = \theta_0 + \theta_{\text{in}} \cdot \text{height (in)} + \theta_{\text{cm}} \cdot \text{height (cm)} + \theta_{\text{lb}} \cdot \text{weight (lb)} + \theta_{\text{kg}} \cdot \text{weight (kg)}$$

Model B:

$$\hat{\text{bmi}} = \beta_0 + \beta_{\text{cm}} \cdot \text{height (cm)} + \beta_{\text{kg}} \cdot \text{weight (kg)}$$

Suppose we create 95% confidence intervals for each of the above non-intercept parameters using the bootstrap method. Which of the following parameters' confidence interval will likely contain the value 0? Select all that apply.

□ θ_{in}

□ θ_{cm}

□ θ_{lb}

□ θ_{kg}

□ β_{cm}

□ β_{kg}

□ None of the above

(d) **(1.0 pt)** Suppose we add random noise to all columns in `humans` except for `bmi`. Assume that our random noise is drawn from the Normal distribution with mean 0 and variance 2, and that the noise for each element in the DataFrame is independent. We call this new DataFrame `noisy_humans`.

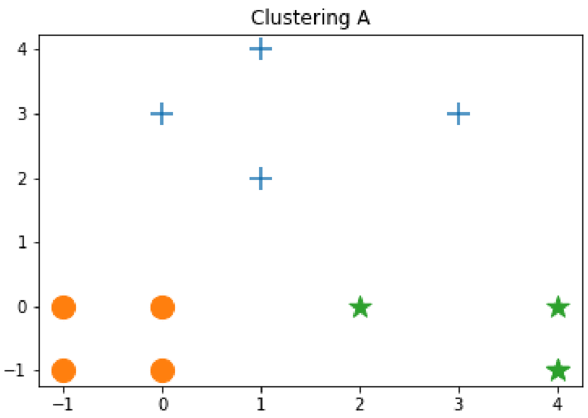
Suppose we fit Model A and Model B on `noisy_humans` and create bootstrapped confidence intervals for each of the above six parameters. True or False: our answer to the previous part remains the same.

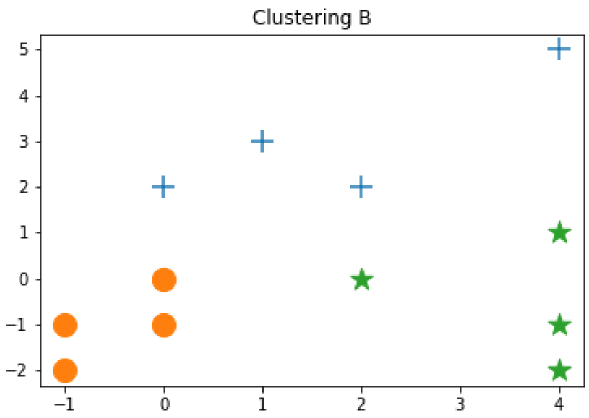
○ True

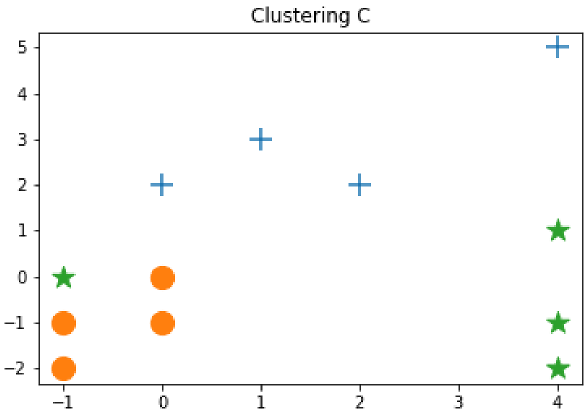
○ False

13. (3.0 points)

Below, we've clustered three different datasets each into three classes (orange circles, blue crosses, and green stars). Assume that there are no overlapping points anywhere.







(a) (2.0 pt) In which of the above dataset/clustering combinations is it true that

$$\text{inertia} = n \cdot \text{distortion}$$

where n is a positive integer? Select all that apply.

□ Clustering A

□ Clustering B

□ Clustering C

□ None of the above

(b) (1.0 pt) In which of the above dataset/clustering combinations is there a point with a negative silhouette score? Select all that apply.

□ Clustering A

□ Clustering B

□ Clustering C

□ None of the above

14. (3.0 points)

(a) (1.0 pt) Fill in the blanks: In the star schema for data storage, the fact table contains ____ that refer to ____ in ____.

○ primary keys, secondary keys, dimension tables

○ integers, primary keys, dimension tables

○ primary keys, dimension tables, foreign keys

○ primary keys, foreign keys, dimension tables

○ foreign keys, primary keys, dimension tables

(b) (1.0 pt) Fill in the blanks: ____ is/are designed to manipulate small amounts of data. ____ is/are designed to manipulate large amounts of data. ____ do/does both.

○ numpy and pandas, Hadoop and Spark, Modin

○ Hadoop and Spark, Modin, numpy and pandas

○ Hadoop and Spark, numpy and pandas, Modin

○ Modin, numpy and pandas, Hadoop and Spark

(c) (1.0 pt) True or False: Hadoop, Spark, and Modin were all created at Berkeley.

○ True

○ False

No more questions.