# Data C100, Midterm Exam

Fall 2021

**Important Note: this exam was administered under very different conditions to the typical Data 100 semester. Its difficulty level and topic coverage are likely not representative of the exams you may receive in your semester's offering of the class.**

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Exam Room: _____

## Instructions:

**Do not** open the examination until instructed to do so.

This exam consists of **85 points** and must be completed in the **110 minute** time period ending at 9:00, unless you have accommodations supported by a DSP letter.

For multiple-choice questions with circular bubble options **select one choice**. For multiple-choice questions with box options, **select all choices that apply**. In both cases, please **shade in** full the box/circle to mark your answer. Do not use checkmarks or Xs.

**Make sure to write your student ID on each page** to ensure that your exam is graded.

## 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: _____

The difficulty level and topic coverage of this exam are likely not representative of future semesters.

# 1 Sparse Data and Even Sparser Structure

Connie is working on fitting a linear model to a Data 100 dataset of 50 DNA strings. Each DNA string contains 3,000 letters, and each letter is one of `T`,`G`,`C`,`A`. Here is an example string:

```
TCGAGGTACGATAGATGCAGATG...
```

DNA contains data for building proteins: In particular, ***disjoint* segments of 3 letters called "codons"** are each translated into protein building blocks called "amino acids". There 64 possible codons ($4^3$) but only 20 possible amino acids, so **many codons may map to one amino acid**. Finally, biologists have shown that **adjacent codons may affect how one another are translated**.

*N-gram review: Recall from the Feature Engineering lecture that to encode data using n-grams, we extend a bag-of-words encoding with all consecutive sequences of n words as "words". For example, say we encode "I love ice cream" with a 3-gram. Our resulting encoding is [1, 1], where the 2 dimensions correspond to "I love ice" and "love ice cream".*

Connie has asked for your advice on how to start: we will help her featurize data, design a model, and evaluate. Our goal is to understand how both sample and feature sparsity inform model design.

(a) [3 Pts] **Encoding your Data**: We can't naively feed letters, non-numerical data, into our linear regression model. As a result, we need to decide on a way to convert non-numerical letters into a numerical representation. Given your domain knowledge above, pick the "best" encoding strategy for the data. Here, we define "best" as the most compact representation that *preserves the meaning in your data*. More compact representations that lose meaning are invalid.

- ◯ Apply one one-hot encoding to the entire DNA string. (The resulting encoding would only have *one* entry with value 1)
- ◯ One-hot encode each of the 1,000 disjoint codons, per sample.
- ◯ One-hot encode each of the 500 disjoint 6-letter combinations, per sample.
- ◯ Bag-of-words encoding, where each "word" is a codon

(b) [6 Pts] **Compute Dimensionality**: For each of the options above, report the number of dimensions each sample will have. For large numbers, you may leave the expression unevaluated (e.g., $e^{30}$). *Explain your work in plain English for partial credit.*

- **One-hot encode whole**:
- **One-hot encode codons**:
- **One-hot encode 6-letter combos**:
- **Bags-of-words**:

(c) [2 Pts] **N-gram Downfalls**: Why is a vanilla N-gram (with $N$ up to 3) encoding a poor strategy for this data? Select all that apply.

- ☐ Segments of non-consecutive letters would be included, meaning `ATGCA` could suggest the word `AGA`.
- ☐ Segments with 1 or 2 letters have no meaning, which N-gram would capture as "words".
- ☐ Overlapping segments would be included, meaning `ATGCA` is broken up into 3 overlapping words.
- ☐ DNA does not "wrap around", but N-gram would capture "wrap-around" segments, like `TGG` from `GGTT`.

(d) [2 Pts] **Design Encoding**: Connie is confident there is a more compact representation that remains faithful to the data (and she's right). Describe an encoding strategy more compact than all of the strategies above. **Additionally include the dimensionality of your encoded data**. This should be one boxed number.

(e) [1 Pt] **Evaluation Robustness** The validation accuracy of Connie's model is extremely unstable, with each run producing wildly different numbers. She is rightfully concerned but doesn't know how to fix this. What should she do?

- ☐ Retrain with more data in the training set allocated to the validation set.
- ☐ Retrain with more data in the validation set allocated to the training set.
- ☐ Add more features to the dataset, to increase model complexity.
- ☐ Use cross-validation.

(f) [1 Pt] **Responding to Overfitting**: Connie trains a model that overfits significantly and is considering regularizing her model. Which of the following regularization techniques should she apply? There is *one* best answer.

- ◯ Apply L2 regularization, due to the redundancies in codon meaning.
- ◯ Apply L1 regularization, due to the redundancies in codon meaning.
- ◯ Apply L2 regularization, due to the high-dimensionality of the samples.
- ◯ Apply L1 regularization, due to the high-dimensionality of the samples.

(g) [1 Pt] **Picking Hyperparameters**: Connie is now using cross validation to pick a hyperparameter for her regularization term above. She provides you with the errors she has computed, along with the fitted parameters for each fold and the training data points. Assume that the training and validation sets for each fold are consistent across all three choices of $\lambda$. What value of $\lambda$ should Connie choose?

| Fold Num | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 3$ |
|---|---|---|---|
| 1 | 1 | 0.5 | 9 |
| 2 | 3 | 1.5 | 1 |

| Fold Num | Training Data | $\theta_1$ | $\theta_2$ |
|---|---|---|---|
| 1 | Rows 1 and 2 | -1 | 1 |
| 2 | Rows 3 and 4 | 0 | 2 |

**Your Answer:**

(h) [1 Pt] Mark all the statements that are true, about cross-validation.

- ☐ Increasing k in k-fold cross validation decreases the bias of the model.
- ☐ We use cross validation because it is less computationally intensive than regular validation.
- ☐ We use cross validation over regular validation because cross validation allows us to use the entire training data set to test the model
- ☐ Cross-validation cannot be used in production, as the model is cheating by looking at validation samples.

# 2 Bias-Variance Tradeoff

Parth is creating a linear model with ridge regression and no intercept to predict Netflix weekly watch-time of consumers. He gets his training data from a company that collected 500 consumer surveys recording data about their demographics and watching habits.

(a) [2 Pts] He starts by training a model using only demographic information. Parth experimentally determines the bias and variance of his model: he finds the bias is 5 and the variance is 4. Parth read that the measurement error from the surveys has mean 0. If he wants to keep his model risk to be no higher than 38, what is the maximum standard deviation in his data that would be acceptable?

_____

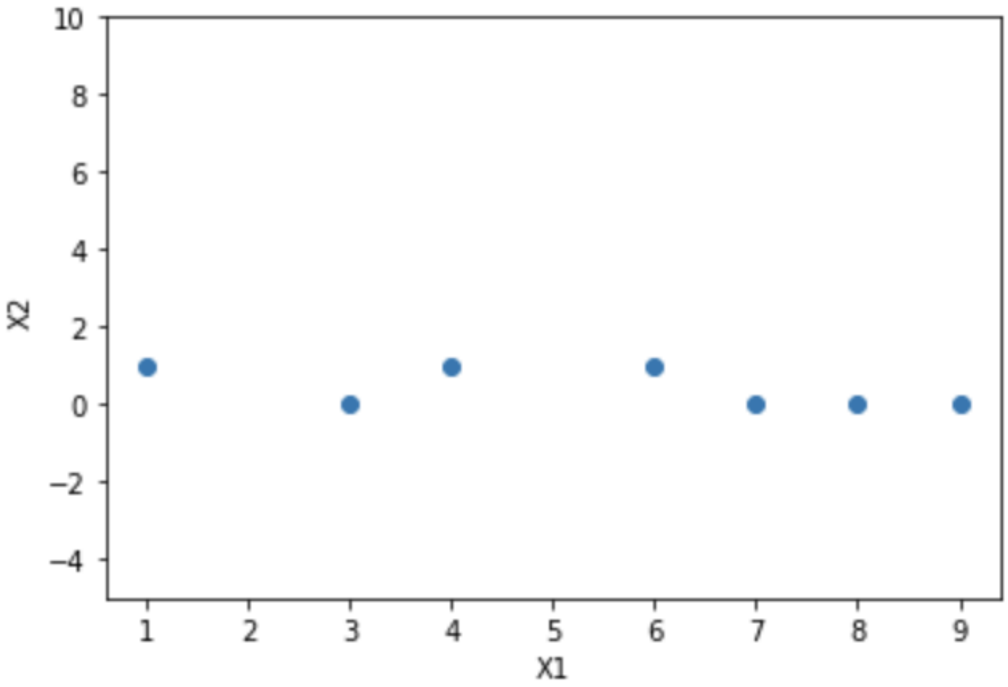
(b) [2 Pts] Parth's model from part a is overfitting and he wants to correct that. Select all of the following strategies that Parth could implement to decrease model variance.

- □ Increase $\lambda$.
- □ Decrease $\lambda$.
- □ Add new features like cable watch time.
- □ Remove some demographic features.
- □ Add an intercept term.
- □ Gather more training data.

(c) [2 Pts] Parth notices that the model doesn't perform so well on the test set. Thus, in an attempt to improve the model, Parth removes an "Avg Hours Watched Per Day" feature from his model. As a consequence, the model now performs equally poorly on the training data. What is the most likely consequence that this change had on model bias or variance?

- ◯ Model bias increases.
- ◯ Model bias decreases.
- ◯ Model variance increases.
- ◯ Model variance decreases.

(d) [2 Pts] Parth now decides to use only the two most important features, $X_1$ and $X_2$, in the dataset to predict weekly watch time. Given below is the scatterplot of the two feature values along a representative sample of the dataset. Recall Parth's model was based on ridge regression. From the plot below, how would the model bias be affected if Parth were to use a LASSO regression model instead?




◯ Model bias increases.

◯ Model bias decreases.

◯ Model bias stays the same.

# 3 Obligatory Outliers

We have discussed techniques to train linear models with loss functions to be resistant to outliers. Suppose Wanda the Panda and Anirudhan wish to predict extreme weather events such as tornadoes using wind data, where accuracy on outliers is just as (if not even more!) important than on typical data points.

They train a standard linear regression model with an intercept term using a mean square loss on wind speed (1 feature), but they find that it performs poorly on outlier points. Assume $\theta$ is the parameter corresponding to the model.

(a) [3 Pts] Which of the following loss functions, when optimized, could perform better on outliers in this scenario than mean square loss? Select all that apply

- □ $L(\theta) = \frac{1}{n}\sum_i |y_i - f_\theta(x_i)|$
- □ $L(\theta) = \max_i (y_i - f_\theta(x_i))^2$
- □ $L(\theta) = \frac{1}{n}\sum_i (y_i - f_\theta(x_i))^3$
- □ $L(\theta) = \frac{1}{n}\sum_i e^{(y_i - f_\theta(x_i))^2}$

(b) [3 Pts] Regardless of your answer to the previous question, suppose she decides to implement an approach that optimizes the following loss function.

$$L(\theta) = \frac{1}{n}\sum_i ((f_\theta(x_i) - \ln y_i)^2 + e^{x_i y_i})$$

Choose the optimal value of the parameter $\hat{\theta}$ if $f_\theta(x_i) = \theta$ for all $i \in 1 \ldots n$ out of the following options:

- ◯ $\hat{\theta} = \frac{1}{n}\sum_i e^{y_i}$
- ◯ $\hat{\theta} = \frac{1}{n}\sum_i \ln y_i$
- ◯ $\hat{\theta} = \frac{1}{n}\sum_i e^{y_i} \ln y_i$
- ◯ $\hat{\theta} = \frac{1}{n}\sum_i x_i \ln y_i$

(c) [1 Pt] Suppose Anirudhan and Wanda modify the loss function as shown below, where they minimize the maximum possible exponential loss on the dataset. Which of the following is true about this modification?

$$L(\theta) = \max_i e^{(y_i - f_\theta(x_i))}$$

- ☐ No outliers will be penalized heavily.
- ☐ Some outliers will be penalized heavily.
- ☐ All outliers will be penalized heavily.
- ☐ This loss function is impossible to optimize in a general machine learning framework with calculus.

(d) [1 Pt] Anirudhan and Wanda decide that it could be helpful for the machine learning algorithm to know whether the current data point is an outlier. Recall that they use 1 feature, the wind speed, with an ordinary least squares model with an intercept term. They plan to use mean square loss with the normal equations to solve for his optimal $\theta$. Which of the following ideas are likely to decrease errors on outliers?

- ☐ Include the range of the wind speeds as a constant feature.
- ☐ Include the standard deviation of the wind speeds as a constant feature.
- ☐ Include the each data point's standardized wind speed as a feature.
- ☐ Include a binary label corresponding to whether each wind speed is inside of the 1.75 * IQR range of the wind speed distribution.
- ☐ Include 2 binary labels corresponding to whether each wind speed is inside and outside of the 1.75 * IQR range of the wind speed distribution.

(e) [4 Pts] They use a simple linear regression model with **standardized** wind speed as the feature and wishes to find the optimal parameters. Unfortunately, they have misplaced their $x$ values! Instead, they recorded some statistics about the product between $x$ and $y$ as well as the target values $y$. Calculate the correlation coefficient, $r$, slope, $\theta_0$, and intercept, $\theta_1$.

Your answer should be 3 numbers (with no leading/trailing 0's), separated by a single comma and no spaces. For example, if you believe $r = .8$, $\theta_0 = 1$, $\theta_1 = 2$, your answer should be $.8, 1, 2$. Please round your answer to 2 decimal places.

*hint:* Recall what standardized means for our $x$!

| | xy | y |
|---|---|---|
| **count** | 100.000000 | 100.000000 |
| **mean** | -0.525185 | 6.479142 |
| **std** | 12.052144 | 10.620935 |
| **min** | -43.102587 | -10.137668 |
| **25%** | -6.216035 | -3.168057 |
| **50%** | -0.426497 | 5.584462 |
| **75%** | 4.179957 | 16.905455 |
| **max** | 24.321534 | 23.049212 |

# 4 Random?

Two Data 100 students, Agnibho (Player A) and Bella (Player B), have finished their lab early and have decided to play a simple game involving coins and dice. First, they each flip a coin. If their coin lands on Heads, they roll a die that has equal probability of landing on a 2, 4, or a 6 (and 0 probability of landing on a 1, 3, or a 5). If their coin lands on Tails, they roll a die that has an equal probability of landing on a 1, 3, or a 5 (and 0 probability of landing on a 2, 4, or a 6). Each player's final score is the outcome of their dice roll. (Note: each player separately flips their coin and then separately rolls the corresponding die). Whoever has the higher score wins the round. They play 10 rounds.

Note: for all parts of this question, when asked for a probability, give your answer as a decimal rounded to the nearest thousandth. If you are asked for an integer, make sure to give the number without any spaces or trailing 0's.

(a) [2 Pts] Assuming they are playing with fair coins (50 percent chance of heads and 50 percent chance of tails), what is the expected score for Player A in a given round?

(b) [1 Pt] Suppose Player B cheats and plays with a weighted coin where the chance of heads is now 0.6 and the chance of tails is 0.4. Let H represent the number of heads Player B gets over the 10 rounds with the weighted coin. What is the distribution of H? Your answer should be a singular distribution with the relevant parameters.

(c) [2 Pts] What is the probability that H is at least 1?

(d) [2 Pts] With Player B playing with this weighted coin, what is the probability that neither player gets a score (outcome of their die roll) strictly above 3 in a given round?

(e) [2 Pts] With Player B playing with the weighted coin, what is the probability that the players tie in any given round?

(f) [3 Pts] With this new weighted coin, what is the expected score differential between the two players in each round?

Select the player that is expected to win.

◯ Player A

◯ Player B

By how much will the player you selected in the above subpart win? In other words, what's the expected score differential as a decimal to the nearest tenth?

(g) [3 Pts] Regardless of your previous answer, assume that Player B wins in expectation by 0.2 points. Suppose the players agree that the loser has to pay 5 dollars for every point in the difference between the two scores. For example, if Player A wins by 1 point, Player B will pay them 5 dollars; if they win by 2, Player B will pay them 10 dollars. What is the expected outcome at the end of the 10 rounds?

Which of the following is true?

◯ Player A pays Player B.

◯ Player B pays Player A.

How much money is expected to be exchanged? If your answer is $\$X$, please give your answer as $X$ to the nearest whole number.

# 5 Re-staurant Regex

The Data 100 staff decided to raise some funding through starting a restaurant and created a text generator to assist with the text messaging orders, but there's an issue. The generator would create run-on sentences when prompting the customers.

(a) [2 Pts] When asking the customer for drink orders, the generator asks this question:

```
s1 = "do you want to add apple juice to your order
or how about a can of coke and what about water but
would you rather have chocolate milk instead"
```

We want to separate the run-on sentence by the conjunctions so we can ask each question separately. Specifically, separate `s1` on `"and"`, `"or"`, `"but"` such that our output consists of all the sentences separated as items in a list as shown below. Do now worry about any potential whitespace before or after any of the strings of the resulting list.

Note: Do not worry about the edge cases in the English language. We will only grade based on the strings we provide in the problem.

```
[
    "do you want to add apple juice to your order ",
    "how about a can of coke ",
    "what about water ",
    "would you rather have chocolate milk instead",
]
```

Fill in the blanks in the code below such that it outputs the above:

```
pattern = r"_____________________"
re._______(pattern, s1)
```

(b) [2 Pts] The generator machine actually misbehaves even more as it inserts extra whitespaces around the conjunctions (`"and"`, `"or"`, `"but"`) when merging sentences.

```
s2 = "do you want to add apple juice to your
order \n\t  \nor\t\t how about a can of
coke \t  \nand    what about water   \t\n\nbut
\t\t would you rather have chocolate milk instead"
```

Edit your regular expression from the previous part such that it separates the conjunctions and excludes the extra whitespaces. The output should look exactly the same as the previous part,

but using `s1` instead of `s2`. The escape character on "t" and "n" denote a tab and a newline, respectively.

```
updated_pattern = r"____________________"
```

(c) [3 Pts] Say Wallace responds back to the generator with a large order.

```
s3 = "I want to order 50,000 water and 75 apples
and 40 pizzas and 100,999 bobas and 4,321 tacos but
I also want 23 juices and 456 aloes and 9,876
burgers"
```

The staff wants to identify all the orders Wallace requested in tuples:

```
>>> [('50,000', 'water'), ('75', 'apples'), ...  ]
```

Fill out the code below such that the above list of tuples is generated.

Note:

- All numbers will be followed by a single whitespace character
- All objects following the numbers will only be 1 alphabetic word
- Numbers in the thousands will have a comma
- Range of possible numbers will be between 1 - 999,999

Hint: use multiple capturing groups to split into tuples.

```
re._______(r"_____________________", s3)
```

# 6 SQL

The Data 100 staff is holding a mid-semester dinner, and Andrew was assigned to pick the restaurant. Since it's a Data 100 event, he wants to have evidence to support his decision, so he gathers some data on the staff members' experiences of different restaurants and constructs a table named `staff_reviews`, the first 4 lines of which are shown below:

Write your answer in the provided box for each part.

(a) [2 Pts] Andrew wants to aggregate some of this information to make it easier to make a decision. Fill in the blanks to construct SQL query that creates a table `agg_reviews` containing the restaurant name, type of food, average rating, and average price paid for each restaurant. The first two lines of `agg_reviews` are shown below:

```
CREATE TABLE agg_reviews AS
    SELECT _____A_____
    FROM staff_reviews
    _____B_____;
```

Fill in the blanks in each part as indicated above.

1. What goes in the blank indicated by the letter `A`?

2. What goes in the blank indicated by the letter `B`?

(b) [6 Pts] Andrew now wants to pick the restaurant with the highest average rating. Write a query that outputs the restaurant with the highest average rating, along with the rating. The final output should be a single row.

```
CREATE TABLE andrews_choice AS
    SELECT _____C_____
    FROM agg_reviews AS_____D_____
    _____E_____
        (SELECT _____F_____
        _____G_____
    WHERE _____H_____;
```

Fill in the blanks in each part as indicated above.

1. What goes in the blank indicated by the letter `C`?

2. What goes in the blank indicated by the letter `D`?

   _____

3. What goes in the blank indicated by the letter `E`?

   _____

4. What goes in the blank indicated by the letter `F`?

   _____

5. What goes in the blank indicated by the letter `G`?

   _____

6. What goes in the blank indicated by the letter `H`?

   _____

(c) [4 Pts] Kelly doesn't trust Andrew's process, and thinks it would be better to choose a restaurant with a rating of at least 4.5 and with at least 8 staff member reviews. However, Kelly doesn't like eating at restaurants that have the letter "t" in their name. Write a query to find all the restaurants that Kelly would consider eating at, along with their average rating.

```
CREATE TABLE kellys_choice AS
    SELECT _____I_____
    FROM _____J_____
    WHERE _____K_____
    _____L_____;
```

Fill in the blanks in each part as indicated above.

1. What goes in the blank indicated by the letter `I`?

   _____

2. What goes in the blank indicated by the letter `J`?

   _____

3. What goes in the blank indicated by the letter `K`?

   _____

4. What goes in the blank indicated by the letter `L`?

   _____

# 7 Pandas

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 `ath` contains the names of athletes who participated in the Olympic Games, including all the Games from Athens 1896 to Tokyo 2020. The first 5 lines of the table are shown below. You may assume that the `ID` column is the primary key of the table.

(a) [2 Pts] Choose the line of code that correctly sorts the number of unique Olympic events per year, in descending order of events. The result should show a series with the year and the number of events.

◯
```
ath.groupby('Year')['Event'].value_counts()
.sort_values(ascending=False)
```

◯
```
ath.groupby('Event')['Year'].unique()
.sort_values(ascending=False)
```

◯
```
ath.groupby('Year')['Event'].unique().agg(len)
.sort_values(ascending=False)
```

◯
```
ath.groupby('Year')['Event'].unique()
.sort_values(ascending=False)
```

◯
```
ath.groupby('Year')['Event'].unique().sort_values()
```

◯
```
ath.groupby(['Year','Event']).value_counts()
```

(b) [2 Pts] Choose the line of code that correctly identifies the athlete with the most medals of all time. The result should show a series with one row of the name of the athlete and the corresponding number of medals.

◯
```
ath.groupby(['Name','Medal']).count()
.sort_values(ascending=False)
```

◯
```
ath.groupby('Name')['Medal'].count()
.sort_values(ascending=False).head(1)
```

◯
```
ath.groupby('Medal')[Name'].count()
.sort_values(ascending=False).head(1))
```

◯
```
ath.groupby('Name')['Medal'].value_counts().head(1)
```

◯
```
ath.groupby('Medal')['Name'].value_counts().head(1)
```

◯
```
ath.groupby('Name')['Medal'].count().max()
```

(c) [2 Pts] Fill in the blanks below in order to answer the question: What is the average age of female athletes who won gold medals, across all years?

```
ath[(______A______)& (______B______)][______C______].______D______
```

Fill in the blanks in each part as indicated above.

1. What goes in the blank indicated by the letter `A`?

   _____

2. What goes in the blank indicated by the letter `B`?

   _____

3. What goes in the blank indicated by the letter `C`?

   _____

4. What goes in the blank indicated by the letter `D`?

   _____

(d) [2 Pts] Write a line of code to answer the following question: What is the name of the athlete that won the bronze medal in the event ”Swimming Men’s 400 metres Freestyle” in 2008? Your answer should be a single String.

# 8 Climate change and physical data

The behavior of the climate is governed by physical laws, and the data we have about the climate is also connected to the physical properties of the Earth.

(a) [1 Pt] If the absolute temperature of a body increases by a factor of two, how much energy does it emit relative to its original temperature? ($\frac{1}{2}$ means half as much, 1 means the same, etc).

- ◯ $\frac{1}{2}$
- ◯ 1.
- ◯ 2.
- ◯ 4.
- ◯ 10.
- ◯ 16.
- ◯ 32.
- ◯ 100.

(b) [1 Pt] If we compute the global average air temperature based on a regular latitude x longitude grid, we obtain:

- ◯ An under-estimate, because latitude lines are not actually equally spaced on Earth.
- ◯ An under-estimate, because we have too few samples along the tropical regions relative to the polar ones.
- ◯ The correct value.
- ◯ An over-estimate, because longitude lines are not actually equally spaced on Earth.
- ◯ An over-estimate, because we have too few samples along the tropical regions relative to the polar ones.