Data C100, Final Exam

Summer 2022

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Instructions:

This final exam consists of **100 points** spread out over **7 questions** and must be completed in the **180 minute** time period from 6:00 PM to 9:00 PM, unless you have accommodations supported by a DSP letter.

Note that some questions have circular bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**. Please shade in the box/circle to mark your answer.
Please **complete the exam in the associated Gradescope assignment**!

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: _____

Proctoring Guidelines:

Please confirm that you have followed all proctoring guidelines to the best of your knowledge and ability. You must be recording your screen and video at all times while taking the exam.

◯ **I confirm that I have done so.**

◯ I have not done so.

1 First Question Troubles [11 Pts]

Anirudhan and Dominic are trying to write the final exam, but they have no idea what to ask! They decide to ask the students enrolled in the course for a few ideas in a Google Form survey as part of a weekly check. Even better, they realise that the construction of this Google Form survey could actually be the first question of the final exam!

(a) [2 Pts] Dominic messages Anirudhan about how he wants to conduct this survey, but it gets corrupted by two blanks! Fill in blanks A and B in the statement below to best complete Dominic's message. Each blank can only be filled in with a singular word corresponding to the **best** response possible in that blank.

> *Ideally, I [Dominic] want everyone enrolled in the course to fill the form, which is most well known as a _____A_____. If a significant portion of the class does not respond, we may incur _____B_____ bias.*

Solution: A: census, B: non-response

(b) [2 Pts] Suppose they are unable to survey everyone in the course. Instead, they decide to only poll people in the course that attend the final lecture live at 5:00 PM PT using a Zoom poll. Which of the following is true about the population and sampling frame?

- ☐ The population is the same as the sampling frame.
- ☐ The population is a subset of the sampling frame.
- ☐ **The sampling frame is a subset of the population.**
- ☐ **The sampling frame does not well represent the population.**

(c) [3 Pts] Suppose we want each student to only fill our Google Form survey once. To do this, Dominic decides to make students enter their student ID, which is a 9 or 10 digit numerical value. To make sure students have entered their SID correctly on the Google form, he wants to write a REGEX pattern to help him fully match all valid student IDs. If a student ID is valid, we wish to replace the **entire** SID with the message `'success!'`.

Fill in the REGEX pattern below such that the output below is generated.

```
>>> pattern = r'__________________'
>>> re.sub(pattern, 'success!', '123456789')
'success!'
>>> re.sub(pattern, 'success!', '2345678901')
'success!'
>>> re.sub(pattern, 'success!', '12345678a')
'12345678a'
```

```
>>> re.sub(pattern, 'success!', '!23456789')
'!23456789'
>>> re.sub(pattern, 'success!', '12345')
'12345'
>>> re.sub(pattern, 'success!', '123456789012')
'123456789012'
```

Solution:

```
pattern = r'^\d{9,10}$'
```

(d) [4 Pts] Rahul, a Data 100 TA, realizes that the form is still not secure enough! He decides to add a password to the form with two requirements shown below, both of which must be met for a valid password. The password must contain:

- 1 or more lowercase characters (i.e., a, b, c, ..., z)
- 1 or more uppercase letters (i.e., A, B, C, ..., Z)

To enforce the password requirements, he asks you to help him write a REGEX pattern to match all valid passwords and fail to match all invalid passwords. Fill in the REGEX pattern below such that the output below is generated.

Hint: Consider the different orders in which lowercase and uppercase letters may occur.

```
>>> pattern = r'__________________'
>>> test_passwords = ['Abc', 'abc', 'aBc', 'a@#', 'a!@#B']
>>> for password in test_passwords:
...     print(password, "matches?",
...           re.search(pattern, password) is not None)
...
Abc matches? True
abc matches? False
aBc matches? True
a@# matches? False
a!@#B matches? True
```

Solution:

```
pattern = r'[A-Z].*[a-z]|[a-z].*[A-Z]'
```

2 SQL to a Pandas Question [19 Pts]

Anirudhan and Dominic receive survey data from students about their preferences and interests which will help them write the final exam! Siddhant, a Data 100 TA, decides to store the data in a table named `survey_data` in a SQLite database since it is too large to open in Pandas on DataHub.

An arbitrary selection of 5 rows of the collected survey data is shown below when the following query is executed.

```
SELECT * FROM survey_data LIMIT 5;
```

score	country	city	n_movies_per_yr	pref_cartoon_character
2	Italy	Rome	3	Popeye
2	Germany	Berlin	6	Popeye
3	Ireland	Clonakilty	8	Homer Simpson
3	India	Coimbatore	1	Popeye
3	United States	Los Angeles	5	Popeye

Each row represents one student response to the survey. The `score` represents the student's score on the survey out of 3, the `city` and `country` represent the self-reported location where the student currently resides, `n_movies_per_yr` represents the number of movies the student watches in a year, and `pref_cartoon_character` represents the student's favourite animated character.

You may assume that there are **no null values**.

(a) [4 Pts] Siddhant wants to find out which cartoon characters are most popular among Data 100 students in the United States. Write a SQL query that returns the most popular preferred cartoon character among all students living in the `'United States'`, in descending order of number of occurrences in the survey data.

The output should resemble the table shown below, with rows corresponding to each preferred cartoon character and respective counts. **Note that the column names (aliases) must match.**

pref_cartoon_character	count
Mickey Mouse	18
Bugs Bunny	18
Popeye	17
Homer Simpson	14
Fred Flintstone	14

Solution:

```
SELECT pref_cartoon_character, COUNT(*) as count
FROM survey_data
WHERE country = 'United States'
GROUP BY pref_cartoon_character
ORDER BY count DESC;
```

(b) [4 Pts] Interestingly, Siddhant observes that all the favourite cartoon characters for students living in the United States from the previous subpart were all originating from the United States! He wishes to explore this relation further and loads a table `cartoon_origin` showing the origin for each cartoon character. A sample of 5 rows from this table are shown below.

character	origin_country
Mickey Mouse	United States
Arthur	Canada
Franklin	Canada
Homer Simpson	United States
Mr. Bean	United Kingdom

Write a SQL query that shows the proportion of students living in **each country from the survey data** that chose cartoon characters originating from that country, ordered from greatest to least. For cartoon characters not listed in the `cartoon_origin` table, assume by default they do not originate from the student's `country`.

A sample of the output is shown below, where rows correspond to each country and desired proportion per the question specification. **Note that the column names (aliases) must match.**

country	prop
Germany	0.322222
United States	0.283186
South Korea	0.192661
Ireland	0.157025

Hint: Use a combination of two aggregation functions. Recall that SQL allows you to apply aggregation functions to column expressions!

Solution:

```
SELECT s.country AS country,
       SUM(s.country = c.origin_country) / COUNT(*) AS prop
FROM survey_data AS s
LEFT JOIN cartoon_origin AS c
ON s.pref_cartoon_character = c.character
GROUP BY s.country
ORDER BY prop DESC;
```

(c) [3 Pts] Siddhant discovers that Data 100 students are from all over the world! He decides to find using SQL which students' cities contain the word `'City'` in them. If the city contains the word `'City'`, then we want to output `'yes'`; otherwise, we want to output `'no'`.

The output should resemble the table shown below, with rows corresponding to the city in each student's response and whether it contains the word `'City'`. Your output should have

the same number of rows as `survey_data`! **Note that the column names (aliases) must match.**

Fill in the blank below to accomplish this.

```
SELECT city, ____________________________________
FROM survey_data;
```

Hint: The following SQL syntax may be helpful.

```
CASE
    WHEN condition THEN result1
    ELSE result2
END
```

city	contains_city
Beijing	no
Oklahoma City	yes
Berlin	no
Incheon	no
Busan	no

Solution:

```
SELECT city, CASE WHEN city LIKE '%City%'
             THEN 'yes'
             ELSE 'no'
             END AS contains_city
FROM survey_data;
```

(d) [4 Pts] Instead of working with SQL, Siddhant decides to sample 100 rows of the SQL table into a Pandas DataFrame named `survey_sample` and use Pandas instead of SQL! Using a one-line Pandas expression, find the average number of movies watched per year at each unique student location (that is, combination of city and country) of the sampled survey data `survey_sample`.

A sample of the desired output is shown below as a Pandas DataFrame.

		n_movies_per_yr
country	**city**	
Germany	**Berlin**	4.400000
India	**Chennai**	1.666667
	Delhi	1.333333
	Kochi	2.000000
Ireland	**Clonakilty**	7.285714

Solution:

```
survey_sample.groupby(['country', 'city']) \
             [['n_movies_per_yr']] \
             .mean()
```

(e) [2 Pts] Siddhant does some research and discovers that Ireland, South Korea, the United States, and Australia visit the movies most on average per year. Suppose we are provided the (real!) data stored in a Pandas DataFrame `movie_data` shown below about average movie visits per year for **all** countries in the world.

	country	avg_n_movies_per_yr
0	United States	3.7
1	South Korea	4.2
2	Ireland	4.3
3	Australia	3.6
4	Poland	1.6

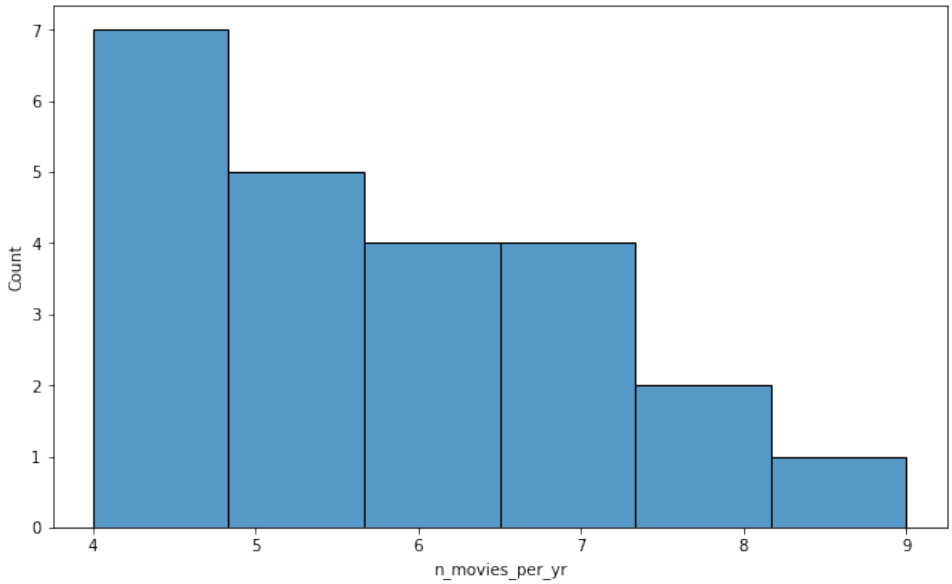
We wish to combine the average number of movies per year (`avg_n_movies_per_yr`) with the country information provided in the survey. That is, for each row in `survey_sample`, the corresponding average number of movies per year is provided for the country specified in that row. Write a one-line Pandas expression that merges the two DataFrames to accomplish this.

Solution:

```
survey_sample_w_avg = survey_sample.merge(movies,
                                          on = 'country')
```

(f) [2 Pts] Visualize using a histogram the number of movie visits for **only** students living in any of these countries using the sampled Pandas DataFrame `survey_sample`: `['Ireland', 'South Korea', 'United States', 'Australia']`. Fill in the blank to output the following figure.

```
c = ['Ireland', 'South Korea',
     'United States', 'Australia']
sns.histplot(data = ____________________________________________,
             x = 'n_movies_per_yr')
```



Solution:

```
c = ['Ireland', 'South Korea',
     'United States', 'Australia']
sns.histplot(data = survey_sample.loc[
                        survey_sample['country'].isin(c)
                    ],
             x = 'n_movies_per_yr')
```

3 Modeling with Mickey Mouse [20 Pts]

Anirudhan and Dominic invite the Data 100 staff to perform some initial modeling on some student survey data to understand student interests, preferences, and habits.

Specifically, we want to predict whether students live in the United States based on their favourite animated (cartoon) character (`pref_cartoon_character`) and number of movies watched per year (`n_movies_per_yr`). The following training data is sampled from student responses to the Data 100 survey!

	pref_cartoon_character	n_movies_per_yr	lives_in_united_states
0	Mickey Mouse	6	1
1	Homer Simpson	2	0
2	Fred Flintstone	4	1
3	Mickey Mouse	2	0
4	Franklin	9	0

(a) [2 Pts] Which of the following variable types best describes the `n_movies_per_yr`?

- ◯ Quantitative continuous
- ◯ **Quantitative discrete**
- ◯ Qualitative nominal
- ◯ Qualitative ordinal

(b) [2 Pts] Which of the following techniques are most suitable for this prediction task?

- ◯ **Classification**
- ◯ Clustering
- ◯ Regression
- ◯ Non-parametric modeling

(c) [2 Pts] Regardless of your previous answer, suppose we decide to use logistic regression. Using the LogisticRegression class in Scikit-learn with no regularization, we obtain 100% accuracy on the training set by training a model on one-hot encoded features specifying the cartoon character and the number of movies watched per year. Which of the following must be true if this has occurred?

- ☐ **The features are linearly separable based on the unique values of the response variable.**

- ☐ The response variable is linearly separable based on the unique values of the features.
- ☐ The design matrix is linearly independent.
- ☐ The response variable has a correlation coefficient of $r = 1$ or $r = -1$ with all of the design matrix features.

(d) [3 Pts] Instead of using one-hot encoding, suppose we encode the `pref_cartoon_character` with a 1 if the character originated from the United States and 0 if it did not. We train a logistic regression model on this feature along with the number of movies per year and a bias feature.

Suppose the optimal parameters obtained for our optimal logistic regression model trained on these three features (including a bias feature, listed as the first element of the vector) are $\hat{\theta} = [2, 1, 0.25]$. Unfortunately, we find that an outlier has strongly affected our cross-entropy loss and our model parameters!

Calculate the cross-entropy loss (using $\log$ base e) on the outlier data point $x_{\text{outlier}} = [1, 1, 4]$ if the corresponding $y_{\text{outlier}} = 0$. Round to the nearest 3 decimal places, and **do not show work**.

Hint: $\sigma(z) = \frac{1}{1+e^{-z}}$

Solution: Note that $P(\hat{Y} = 1|x) = \sigma(\theta^T x) = \sigma(2 + 1 + 1) = 0.98$. Then, calculating the cross-entropy loss:

$$-\left(y_i \log P(\hat{Y} = 1|x) + (1 - y_i) \log P(\hat{Y} = 0|x)\right) = 4.018$$

(e) [3 Pts] Suppose that we added a L regularization term to the logistic regression objective function from the previous function, shown below as $R(\theta)$, and we minimised this objective function to obtain $\hat{\theta}$ for our logistic regression model.

$$R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))] + \lambda \sum_{j=1}^{p} \theta_j^2$$

Which of the following is true?

- ☐ There exists an optimal analytical solution to minimising this objective function $R(\theta)$ that can be derived either using calculus or geometry.
- ☐ It is impossible to achieve an accuracy of 100% using any $\lambda > 0$.
- ☐ **If $\lambda > 0$, then the parameters in $\hat{\theta}$ for this objective function do not diverge to ∞.**
- ☐ **This objective function is convex since both the empirical risk function and the regularization penalty term are convex.**

(f) [2 Pts] After some tinkering, Dominic is able to train another logistic regression using regularization and obtains an optimal $\hat{\theta}$. Suppose our test set consists of the following 6 points, with our corresponding predictions shown. Calculate the **recall** of our model on the test set.

Round to the nearest 3 decimal places, and do not show work.

y	$\hat{y}$
1	1
1	0
1	1
0	1
0	0
0	1

Solution: We find TP = 2, FP = 2, TN = 1 and FN = 1.

Recall is equal to:

$$\frac{2}{2+1} = \frac{2}{3}$$

(g) [2 Pts] Suppose we decide to train a decision tree or random forest instead of a logistic regression model. Which of the following are true about the effect of the following model and hyperparameter choices on the estimated model bias and model variance of the resulting model on a training data point?

Note that unless specified otherwise, decision trees and random forests will not have any training restrictions (e.g., maximum depth).

☐ **A random forest will likely have lower model variance than a decision tree, when fully trained.**

☐ **A decision tree will likely have lower model bias than a random forest, when fully trained.**

☐ A decision tree with depth 10 will likely have lower model variance than a decision tree with depth 2, when fully trained.

☐ **A random forest with 10 trees will likely have lower model variance than a random forest with 2 trees, when fully trained.**

(h) [4 Pts] Rahul decides to train a decision tree for our prediction task without any training restrictions. Suppose he uses simply one feature and a small subset of our data for training, shown below. Which is the most optimal split point for the `n_movies_per_year` using weighted node entropy?

n_movies_per_yr	lives_in_united_states
6	1
2	0
4	1
2	0
9	0

◯ 1

◯ **3**

◯ 5

◯ 7

Justify your answer by calculating the numerical weighted node entropy of **only** this split, **the ideal split from your answer above** (**make sure to use** $\log_2$**, and not** $\log_e$ **or** $\log_{10}$). Round to 3 decimal places, and do not show work.

Solution: Anything less than 3 is a pure node! Greater than 3, 2/3 data points are 1. Hence:

$$H_R(X) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$$

Weighting it by the number of data points: $\frac{3}{5}H_R(X) = 0.551$

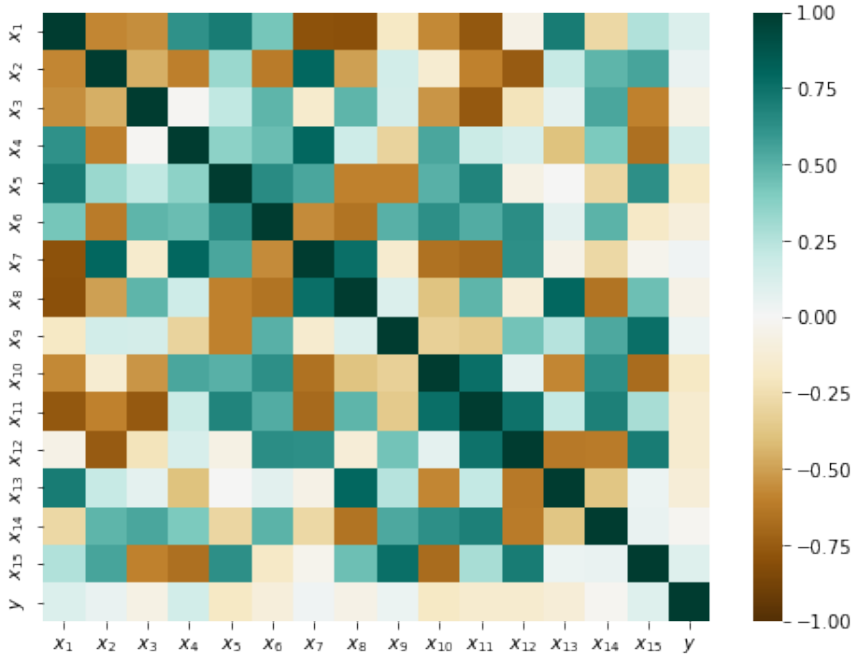
For this exam, we accepted logarithm of other bases too since it was unfamiliar to many how to access $\log_2$ on the online calculator.

4 Mmm... More Movie Modeling... [15 Pts]

Anirudhan and Dominic decide to focus on a different modeling problem: can we model and explain the number of movies our students have watched per year? They will try to predict students' number of movies watched per year using regression and as much data as possible from the questions in the survey.

	country	city	pref_cartoon_character	n_movies_per_yr
0	India	Mumbai	Homer Simpson	1
1	Germany	Berlin	Fred Flintstone	2
2	United States	New York City	Popeye	4
3	South Korea	Busan	Fred Flintstone	6
4	Italy	Rome	Mickey Mouse	3

(a) [2 Pts] To begin, Anirudhan and Dominic decide to one-hot encode all the qualitative features in the survey results. They sample a random subset of all p of these one-hot encoded features to obtain the 15 that they will use in their design matrix: x_1 to x_{15}. They measure the correlation between the 15 sampled features with each other and with the number of movies watched (y), and the correlation heatmap below shows the results. Which of the following is true?



- ☐ Since most of the correlation coefficients between the features and response variable are small, it is certain that the fit will be poor.
- ☐ **Since most of the the correlation coefficients between the features and response variable are small, it is possible that the fit will be poor.**
- ☐ **It is likely that there is multicollinearity among the features in the model.**
- ☐ The correlation coefficients would change if the number of movies watched was standardised.

(b) [3 Pts] Using another random sample of features, Rahul decides to train a multiple linear regression model optimizing the MSE that achieves an optimal average training loss of 0 and an average test loss of 12,000. Which of the following is true in this case?

- ☐ The residuals on the test set are orthogonal to the span of the training set's design matrix.
- ☐ **The residuals on the training set are orthogonal to the span of the training set's design matrix.**
- ☐ The residuals on the test set are orthogonal to the span of the test set's design matrix.
- ☐ **The residuals on the training set are all zero.**
- ☐ **The model is overfitting to the training data.**

(c) [2 Pts] Using the same $\hat{\theta}$ and the multiple linear regression model in the previous part, Rahul decides to interpret the parameters $\hat{\theta}$ that were obtained. One of his friends, Muthu, argues that this is a bad idea. Another friend, Suriya, argues that this is fine as long as Rahul uses bootstrap sampling and confidence intervals.

Who is correct? Note that neither or both may be correct. Justify your answer.

Solution: Both can be correct. See below:

Suriya is correct! Overfitting and multicollinearity both lead to poor interpretability of models.

Muthu is correct! Using bootstrap sampling confidence intervals, we can find which parameters are "important" for linear regression by inspecting whether 0 is contained within.

We were lenient in grading this question due to its ambiguity. Any reasonable and correct justification drawing upon these viewpoints received credit.

(d) [2 Pts] Rahul is unhappy with the multiple linear regression model's performance, and he decides to experiment with using the Ridge regression model using **all of the one-hot encoded qualitative features** to predict the number of movies per year. Suppose that the resulting design matrix is 200×190. Which of the following is true if he uses stochastic gradient descent to compute $\hat{\theta}$?

- ☐ **The $\hat{\theta}$ computed by stochastic gradient descent may differ from the optimal solution.**
- ☐ There isn't a unique optimal solution to minimise the Ridge objective function for this design matrix.
- ☐ **Given a good selection of λ, the Ridge regression model can reduce overfitting.**
- ☐ **Increasing the regularization hyperparameter λ will never decrease the training loss.**
- ☐ Increasing the regularization hyperparameter λ will never increase the test loss.

(e) [4 Pts] To find the optimal λ regularization parameter for his Ridge regression model, Rahul uses a cross-validation algorithm. Unfortunately, it's incomplete! Help him fill in the blanks below to complete the algorithm.

```
def cross_validate(X, y, model, fold_idxs):
    """
    Given a training design matrix X, response vector y,
    a model with fit and predict functions, and fold
    indices, return the average cross-validated mean
    square error.
    """
    mse = []
    for train_idx, test_idx in fold_idxs:
        X_k_train, X_k_test = X.loc[train_idx], X.loc[test_idx]
        Y_k_train, Y_k_test = Y.loc[train_idx], Y.loc[test_idx]
        model.fit(_____, _____)
        mse.append(_____)
    return np._____(mse)
```

Solution:

```
def cross_validate(X, y, model, fold_idxs):
    mse = []
```

```
    for train_idx, test_idx in fold_idxs:
        X_k_train, X_k_test = X.loc[train_idx], X.loc[test_idx]
        Y_k_train, Y_k_test = Y.loc[train_idx], Y.loc[test_idx]
        model.fit(X_k_train, y_k_train)
        mse.append(np.mean((Y_k_test - model.predict(X_k_test))**2))
    return np.mean(mse)
```

(f) [2 Pts] Assume you implemented the cross-validation algorithm in the previous subpart correctly. Based on the results from that cross-validation algorithm applied to determine an ideal regularization hyperparameter λ for our Ridge model shown below, which of the following are true?

Assume we used the Ridge optimal analytical solution to compute $\hat{\theta}$ in any `fit` call.

	CV MSE
λ = 0.001000	70.011580
λ = 0.010000	62.983541
λ = 0.100000	151.983646
λ = 1.000000	80.459269
λ = 10.000000	139.870794

☐ $\lambda = 0.01$ is necessarily optimal for minimizing the training mean squared error.

☐ $\lambda = 0.01$ is necessarily optimal for minimizing the test mean squared error.

☐ **For all the models shown, the sum of the residuals is non-zero even if there is a bias term.**

☐ **For all the models shown, the training predictions ($\hat{y} = X\theta$) are within the span of the column vectors of the design matrix X used to train the model.**

5 Slipping Down A New Loss Surface [16 Pts]

The Data 100 staff are trying to use modeling to predict Data 100 students' number of movies watched per year given data from a survey. After attempts at using Ridge regression, it ultimately does not converge quickly enough!

Anirudhan decides to use simple linear regression instead with **his own custom objective function**, which he hopes will converge sooner (and be easier for you to do calculus with)! However, they don't have much time or any slip days left, so they decide to experiment with gradient descent techniques! For subparts (a) and (b) of this question, please use this dataset shown below.

x	y
0	1
2	6
0	6
1	6
2	1

(a) [3 Pts] Suppose we wish to minimize Anirudhan's **custom objective function for SLR** (i.e., $\hat{y} = a + bx$) using batch gradient descent, and our initialization is $\theta^{(0)} = [a^{(0)}, b^{(0)}] = [0, 0]$. Apply one gradient step using the data shown above and a learning rate of $\alpha = 0.1$ to compute $\theta^{(1)}$.

$$R(\theta) = \frac{1}{2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Enter your answer as a vector containing two numbers, both rounded to 3 decimal places. For example, a sample response might be: `[3.141, 1.000]`.

Solution: The partial derivative with respect to a is:

$$\frac{dR}{da} = \sum_{i=1}^{n}(\hat{y}_i - y_i) = -20$$

The partial derivative with respect to b is:

$$\frac{dR}{db} = \sum_{i=1}^{n} x_i(\hat{y}_i - y_i) = -\sum_{i=1}^{n} x_i y_i = -20$$

Then, $[a^{(1)}, b^{(1)}] = [2, 2]$

(b) [2 Pts] Suppose the optimal solution is $\hat{\theta} = [4, 0]$. Using Anirudhan's objective function $R(\theta)$ for SLR (i.e., $\hat{y} = a + bx$) shown in the above subpart, does there **exist** a scalar learning rate α that helps batch gradient descent converge to $\hat{\theta}$ in one step initializing at $\theta^{(0)} = [a^{(0)}, b^{(0)}] = [0, 0]$?

◯ Yes, there is.

◯ **No, there is not.**

(c) [2 Pts] Using batch gradient descent to minimize $R(\theta)$ from the previous subparts with an arbitrary initialization and learning rate, which of the following is possible? Note that the optimal parameter in this context refers to the parameter which minimizes the empirical risk function on the training set.

☐ **The optimal parameter $\hat{\theta}$ is found in a finite number of gradient steps, and gradient descent returns $\hat{\theta}$.**

☐ The optimal parameter $\hat{\theta}$ is found in a finite number of gradient steps, but gradient descent does not return $\hat{\theta}$.

☐ **The optimal parameter $\hat{\theta}$ is not found.**

(d) [2 Pts] Using **stochastic** gradient descent to minimize $R(\theta)$ from the previous subparts with an arbitrary initialization, batch size, and learning rate, which of the following is possible? Note that the optimal parameter in this context refers to the parameter which minimizes the empirical risk function on the training set.

☐ **The optimal parameter $\hat{\theta}$ is found in a finite number of gradient steps, and gradient descent returns $\hat{\theta}$.**

☐ **The optimal parameter $\hat{\theta}$ is found in a finite number of gradient steps, but gradient descent does not return $\hat{\theta}$.**

☐ **The optimal parameter $\hat{\theta}$ is not found.**

(e) [3 Pts] Suppose we estimated the model bias and model variance using the bootstrap sampling technique in Homework 6 to understand which gradient descent (stochastic gradient gradient or batch gradient descent) technique will work better for this task!

Assume for this case that there exists no observational noise (that is, $\epsilon = 0$). Further, assume that we use a fixed initialization and that we use fixed, equivalent, and reasonable learning rates and the same number of gradient steps for both stochastic and batch GD. Which of the following is most likely to be true?

◯ **Models trained with stochastic GD will likely have larger model bias and larger model variance than batch GD.**

◯ Models trained with stochastic GD will likely have larger model bias and smaller model variance than batch GD.

◯ Models trained with stochastic GD will likely have smaller model bias and larger model variance than batch GD.

◯ Models trained with stochastic GD will likely have smaller model bias and smaller model variance than batch GD.

(f) [4 Pts] Assume for this case that there exists no observational noise (that is, $\epsilon = 0$), as before. Calculate the estimated model risk for the point $(x, Y) = (1, 1)$ given the following calculations from our bias-variance analysis using the same bootstrap sampling technique from Homework 6.

Round to the nearest 3 decimal places, and do not show work in the answer cell below.

	model predictions for $x = 1$ ($\hat{y}$)
count	100.000000
mean	4.299990
std	3.093373
min	-3.962432
50%	4.236718
max	13.532542

Show your work below.

Note: Please do not try to use LaTeX! You may describe your process using standard mathematical symbols on the keyboard (e.g., /, *, +, -, sum, mean, std). Please don't spend too long perfecting the way the math looks - if needed, use words to describe your steps. We will understand your workflow! Answers without justification will receive minimal or no credit. You **must** provide a numerical final answer; simply describing the process will receive no credit.

Solution: We know $Y = g(x)$. Recall the BVD:

$$E[(Y - \hat{Y}(x))^2] = \sigma^2 + (E[\hat{Y}(x)] - g(x))^2 + \text{Var}(\hat{Y}(x))$$

We know $\sigma = 0$ and that $g(x) = Y = 1$. From the statistics, $E[\hat{Y}(x)] = 4.3$ and $\text{Var}(\hat{Y}(x)) = 3.09^2$. Then:

$$E[(Y - \hat{Y}(x))^2] = (4.3 - 1)^2 + 3.09^2 = 20.459$$

Any answer near 20.4 was accepted due to rounding errors.

6 Preferences Clustering Analysis (PCA) [7 Pts]

Dominic has the following question: what kinds of clusters or groups of student preferences exist within Data 100 based on their location and preference of media (e.g., movies)? To answer this, he uses survey data collected from Data 100 students!

He one-hot encodes all the qualitative features such as country, city, etc. Then, he standardizes all the one-hot encoded "dummy" features along with all the numerical features into a matrix X. A sample of the rows and columns of X are shown below.

country_South Korea	country_Italy	country_Germany	country_Canada	country_Japan	country_France	...
-0.41	-0.36	-0.30	-0.32	-0.33	2.77	...
-0.41	-0.36	-0.30	3.08	-0.33	-0.36	...
-0.41	-0.36	3.27	-0.32	-0.33	-0.36	...
-0.41	-0.36	3.27	-0.32	-0.33	-0.36	...
-0.41	-0.36	-0.30	-0.32	-0.33	2.77	...
...	...	...	...	...	...	...

The dimensionality of X is 200×190. Dominic uses the power of `numpy` to find that the **rank of the dataset is 150**.

(a) [2 Pts] Dominic wants to cluster the standardized data received in the survey as shown above. However, DataHub cannot hold all the features in memory, so he decides to use the first k principal components! Which of the following is true if Dominic applies PCA to the standardized data shown and trains a multiple linear regression model on the first k principal components?

- ☐ **As k increases, the training loss decreases and then is constant.**
- ☐ As k increases, the training loss decreases and then increases.
- ☐ As k increases, the test loss decreases and then is constant.
- ☐ **As k increases, the test loss decreases and then increases.**

(b) [3 Pts] Recall that SVD returns a matrix decomposition of a standardized data matrix $X = U\Sigma V^T$ that is used to generate principal components. Which of the following is true regarding this matrix decomposition?

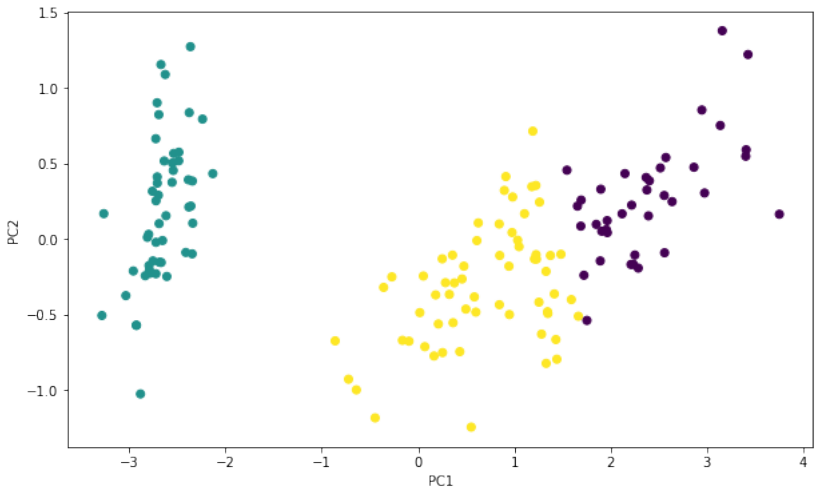
- ☐ The decomposition is unique (i.e., U, Σ, and V^T are unique).
- ☐ It is possible that σ_1 is less than the variance of all of the individual features (that is, column vectors) x_i in X.

☐ **The variance of Xv_1 is greater than or equal to the variance of Xv_2.**

☐ **The variance of $\sigma_1 u_1$ is greater than or equal to the variance of $\sigma_2 u_2$.**

☐ **All principal components are orthogonal to one another.**

(c) [2 Pts] Regardless of the previous subparts, suppose Dominic uses PCA on the data matrix X to select the first two principal components successfully (PC1, PC2) and applies K-Means clustering to the resulting two principal components, as shown below. Which of the following is true of this clustering?



☐ **In this particular clustering, $k = 3$.**

☐ Applying clustering to the first 2 principal components is equivalent to applying it to the entire data matrix X.

☐ This clustering is generated by minimizing the silhouette score using calculus.

☐ **This clustering is an application of an unsupervised learning algorithm.**

7 Finally... Some Extra Credit [11 Pts]

After doing all this data analysis, Anirudhan and Dominic realize that what students really want is extra credit! They ask Michelle and Siddhant, both Data 100 TAs, to devise a clever extra credit question for the final question on the final.

Michelle has decided to include a special extra credit question below in Data 100's final exam, but it has very specific conditions. Suppose the extra credit is given with conditions as specified below:

- Case 1: If no students answer the question, all students receive 1 point of extra credit.
- Case 2: If only two students answer the question, those two students receive 2 points of extra credit and the remaining students receive no extra credit.
- Case 3: If only one student answers the question, that student receives 3 points of extra credit and the remaining students receive no extra credit.
- Case 4: If more than two students answer the question, nobody receives extra credit.

Suppose that there are 201 students enrolled in Data 100 and that all 201 of the Data 100 students answer the question independently of one another. Each student's answer is chosen randomly based on the following probabilities.

- With probability 0.99, the student writes no answer.
- With probability 0.01, the student writes an answer.

Note that in some subparts of the question, when specified, not all students will answer randomly with the above probabilities. Also, note that all subparts except the final subpart are **not** extra credit; the last subpart is for extra credit.

(a) [2 Pts] Suppose that there are 201 students in Data 100. What is the probability that all students will receive 1 point of extra credit (case 1)? Round your answer to three decimal places.

Solution:

$$0.99^{201} = 0.133$$

(b) [3 Pts] Suppose that a Data 100 student, Morgan, decides to answer the question (i.e. this choice is not random!). We can model the number of extra credit points received by Morgan using a random variable, X. Assuming the **remaining** 200 students in the class answer the question randomly and independently with probability 0.01 (as above), compute the expected value of X, $\mathbb{E}[X]$. Round your answer to three decimal places.

Hint: Consider the definition of an expectation and the 3 cases that you will have to account for. Calculate the probability of each of these cases happening and use that to calculate the expectation.

Solution: If none of the remaining students answer, Morgan gets 3 points of extra credit. If one of them answer, she gets 2 points of extra credit. If any more than that answer, then she gets no extra credit.

$$\mathbb{E}[X] = 3(0.99^{200}) + 2(200 \cdot 0.99^{199} \cdot 0.01) = 0.943$$

(c) [2 Pts] Suppose that another Data 100 student, Ash, decides to **not** answer the question (i.e. this choice is not random!). We can model the number of extra credit points received by Ash using a random variable, Y, and we assume that all other students (including Morgan) answer the question randomly and independently with probability 0.01 (as above).

If we calculate that $\mathbb{E}[X] > \mathbb{E}[Y]$, which of the following is true for a student to maximize the number of extra credit points?

- ◯ In expectation, it is always better to never answer the question assuming other students choose to answer randomly.
- ◯ **In expectation, it is always better to answer the question assuming other students choose to answer randomly.**
- ◯ In expectation, it is sometimes better to answer the question assuming other students choose to answer randomly.

(d) [2 Pts] Which of the following is true about the random variable Y, the number of extra credit points received by Ash as defined in the previous question?

- ◯ **Y is a Bernoulli random variable.**
- ◯ Y is a binomial random variable such that $n > 1$.
- ◯ Y is neither a Bernoulli nor a binomial random variable.

Solution: Note that Y is a Bernoulli random variable with $p = 0.99^{200}$! It is either 1 with probability p or 0 with probability $1 - p$.

(e) [2 Pts] Suppose the extra credit points received by Ash, Y, are applied to the final exam category, which is worth 30% of her grade.

Which of the following expressions is equivalent to $\text{Var}[0.3Y]$ in terms of $\text{Var}[Y]$, if any?

- ◯ $0.3\text{Var}[Y]$
- ◯ $0.09\text{Var}[Y]$
- ◯ $0.3 + \text{Var}[Y]$
- ◯ $0.3 + 0.09\text{Var}[Y]$

(f) [0 Pts] This is the extra credit question for which the above conditions apply. Will you respond to this? If you want to respond, guess the total number of unique languages our course staff this semester can speak (you don't need to get it right). Remember that if you respond, all the conditions in the question **will** apply.

_____

Solution: Sorry! More than 20 students answered, so nobody got the extra credit :(

However, we got 90% response rate for the evaluation - so you will all receive 1 extra credit point for that! :D