DS-100 Final Exam

Spring 2018

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Instructions:

- This final exam must be completed in the **3 hour time** period ending at **11:00AM**, unless you have accommodations supported by a DSP letter.
- Note **all questions on this exam are single choice only.**
- Please put your student id at the top of each page to ensure that pages are not lost during scanning.
- When selecting your choices, you must **fully shade** in the circle. Check marks will likely be mis-graded.
- You may use a two-sheet (two-sided) study guide.
- Work quickly through each question. There are a total of 199 points on this exam.

Honor Code:

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

Signature: _____

Syntax Reference

Regular Expressions

"^" matches the position at the beginning of string (unless used for negation "[^]")

"$" matches the position at the end of string character.

"?" match preceding literal or sub-expression 0 or 1 times. When following "+" or "*" results in non-greedy matching.

"+" match preceding literal or sub-expression *one* or more times.

"*" match preceding literal or sub-expression *zero* or more times

"." match any character except new line.

"[ ]" match any one of the characters inside, accepts a range, e.g., "[a-c]".

"( )" used to create a sub-expression

"{n}" preceding expression repeated n times.

"\d" match any *digit* character. "\D" is the complement.

"\w" match any *word* character (letters, digits, underscore). "\W" is the complement.

"\s" match any *whitespace* character including tabs and newlines. \S is the complement.

"\b" match boundary between words

Some useful Python functions and syntax

re.findall(pattern, st) return the list of all sub-strings in st that match pattern.

np.random.choice(n, replace, size) sample size numbers 0 to n with replacement.

Useful Pandas Syntax

```
df.loc[row_selection, col_list]  # row selection can be boolean
df.iloc[row_selection, col_list] # row selection can be boolean
df.groupby(group_columns)[['colA', 'colB']].sum()
pd.merge(df1, df2, on='hi') # Merge df1 and df2 on the 'hi' column
```

Variance and Expected Value

The expected value of X is $\mathbf{E}[X] = \sum_{j=1}^{m} x_j p_j$. The variance of X is $\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$. The standard deviation of X is $\mathbf{SD}[X] = \sqrt{\mathbf{Var}[X]}$.

Problem Formulation

1. [3 Pts] In 1936, the Literary Digest ran a poll to predict the outcome of the Presidential election. They constructed a sample of over 10 million individuals by aggregating lists of
   - `magazine subscribers`
   - `registered automobile owners`
   - `telephone records`

   and received responses from about 2.4 million individuals from this 10 million sample.

   (a) What kind of sample is this?

   √ **Convenience Sample** ◯ SRS ◯ Stratified Sample ◯ Census

   (b) Which is likely a more serious concern for the Literary Digest estimate of the proportion of voters who support FDR?

   √ **Bias** ◯ Variance

   (c) Including more registered magazine subscribers would more likely have helped reduce

   ◯ Bias √ **Variance**

2. [5 Pts] Which kind of statistical problem is associated with each of the following tasks?

   (a) Filtering emails according to whether they are spam.

   ◯ Estimation √ **Prediction** ◯ Causal Inference

   (b) Determining whether a new feature will improve a website's revenue from an A/B test.

   ◯ Estimation ◯ Prediction √ **Causal Inference**

   (c) Investigating whether perceived gender has any effect on student teaching evaluations.

   ◯ Estimation ◯ Prediction √ **Causal Inference**

   (d) Building a recommendation system from historical ratings to serve personalized content.

   ◯ Estimation √ **Prediction** ◯ Causal Inference

   (e) Determining the growth rate of yeast cells in a petri dish.

   √ **Estimation** ◯ Prediction ◯ Causal Inference

3. Suppose we observe a sample of n runners from a larger population, and we record their race times $X_1, \ldots, X_n$. We want to estimate the maximum race time θ^* in the population. When comparing estimates, we **prefer whichever is closer to θ^* without going over**. We consider the following three estimators based on our sample:

$$\widehat{\theta}_1 = \max_i X_i$$
$$\widehat{\theta}_2 = \frac{1}{n}\sum_i X_i$$
$$\widehat{\theta}_3 = \max_i X_i + 1$$

(a) [2 Pts] $\widehat{\theta}_1$ is never an over estimate but could be an underestimate of θ^*.

√ **True** ◯ False

(b) [2 Pts] $\widehat{\theta}_1$ is never a worse estimate of θ^* than $\widehat{\theta}_2$.

√ **True** ◯ False

(c) [2 Pts] $\widehat{\theta}_3$ is never a worse estimate of θ^* than $\widehat{\theta}_1$.

◯ True √ **False**

(d) [3 Pts] Which loss $\ell(\widehat{\theta}, \theta^*)$ best reflects our goal of *"closest without going over"*? (where $\widehat{\theta}$ represents any $\widehat{\theta}_i$, $i = 1, 2, 3$)

◯ $\ell(\widehat{\theta}, \theta^*) = (\widehat{\theta} - \theta^*)^2$

√ $\ell(\widehat{\theta}, \theta^*) = \begin{cases} \theta^* - \widehat{\theta}, & \textbf{if } \widehat{\theta} \leq \theta^* \\ \infty, & \textbf{otherwise} \end{cases}$

◯ $\ell(\widehat{\theta}, \theta^*) = |\widehat{\theta} - \theta^*|$

◯ $\ell(\widehat{\theta}, \theta^*) = \begin{cases} \theta^* - \widehat{\theta}, & \text{if } \widehat{\theta} \leq \theta^* \\ 0, & \text{otherwise} \end{cases}$

Data Collection and Cleaning

4. For each scenario below, mark the sampling technique used.

   (a) [1 Pt] A researcher wants to study the diet of California Residents. The researcher collects a dataset by asking her family members.

   ◯ SRS ◯ Stratified Sample ◯ Cluster Sample √ **Convenience Sample**

   (b) [1 Pt] Bay Area Rapid Transit (BART) wants to survey its customers one day, so they randomly select 5 trains and survey all of the customers on these trains

   ◯ SRS ◯ Stratified Sample √ **Cluster Sample** ◯ Convenience Sample

   (c) [1 Pt] In order to survey drivers in a certain city, the police set up checkpoints at randomly selected road locations, then inspected every driver at those locations.

   ◯ SRS ◯ Stratified Sample √ **Cluster Sample** ◯ Convenience Sample

   (d) [1 Pt] To study how different student organizations perceive campus issues, a professor surveyed 3 students at random from each student organization.

   ◯ SRS √ **Stratified Sample** ◯ Cluster Sample ◯ Convenience Sample

5. [1 Pt] The date $01/01/1970$ is typically associated with which data anomaly:

   ◯ Outliers √ **Missing Values** ◯ Leap Years ◯ Roundoff Error

6. [1 Pt] When would it be safe to drop records with missing values?

   ◯ When less than ten percent of the records have missing values.
   ◯ When the missing value occurs in a field that is not being studied.
   √ **When the missing value implies that the entire record is corrupted.**
   ◯ When the missing values are encoded using 999.

7. [1 Pt] When loading a comma delimited file which of the following is a parsing concern?

   ◯ Unquoted tab characters in strings
   √ **Unquoted newline characters in strings**
   ◯ Dates with negative values
   ◯ Capitalization

SQL

Consider the following *database schema* used to track doctor visits to animals in a zoo (note: all the questions in this **SQL** section are based on this schema):

```
CREATE TABLE animals (
    aid INT PRIMARY KEY,
    animal_type TEXT,
    name TEXT,
    age INTEGER,
    color TEXT);

CREATE TABLE visits (
    vid INT PRIMARY KEY,
    aid INT REFERENCES animals(aid),
    did INT REFERENCES doctors(did));
```

```
CREATE TABLE doctors (
    did INT,
    name TEXT,
    PRIMARY KEY (did));
```

Each row of the `animals` table describes a distinct animal. Each row of the `doctors` table describes a distinct doctor at the zoo. Each row of the `visits` table describes a distinct visit of an animal to a doctor. The entire dataset is contained in the following tables:

aid	animal_type	name	age	color
0	rabbit	Bugs	2	white
1	bear	Air	5	golden
2	bear	Care	1	golden
3	cat	Grumpy	75	gray

(a) `animals`

did	name
0	Turk
1	House
2	Dre
3	Bailey

(b) `doctors`

vid	aid	did
132	1	0
145	2	1
167	0	3
168	2	3
169	2	2

(c) `visits`

8. Mark each statement below as True or False.

   (a) [1 Pt] An animal who visits the same doctor multiple times is recorded under the same doctor id (`did`) in the `visits` table.

   √ **True** ◯ False

   (b) [2 Pts] More than one animal could have the same animal type, age and color

   √ **True** ◯ False

   (c) [2 Pts] Each `aid` in the visits table must be present in the `animals` table.

   √ **True** ◯ False

9. [4 Pts] What does the following query compute?

```
SELECT animal_type, color, AVG(age) AS avg_age
FROM animals
GROUP BY animal_type, color;
```

◯ The average age of each type of animal.

◯ The average age of the animals for each type of color.

√ **The average age for each combination of animal type and color.**

◯ This query throws an error.

10. [4 Pts] Which of the following SQL queries computes the names of the animals in our zoo who visited the doctor more than 2 times?

◯
```
SELECT name FROM animals, visits
WHERE COUNT(*) > 2
GROUP BY animals.aid, animals.name;
```

◯
```
SELECT name FROM animals, visits
GROUP BY animals.aid, animals.name
HAVING COUNT(*) > 2;
```

◯
```
SELECT name FROM animals, visits
WHERE visits.aid = animals.aid AND COUNT(*) > 2
GROUP BY animals.aid, animals.name
```

√
```
SELECT name FROM animals, visits
WHERE visits.aid = animals.aid
GROUP BY animals.aid, animals.name
HAVING COUNT(*) > 2;
```

11. [3 Pts] When run on above data what does the following compute:

```
SELECT animal_type
FROM animals JOIN visits ON animals.aid = visits.aid
GROUP BY animal_type
ORDER BY COUNT(*) DESC
LIMIT 1;
```

◯ rabbit √ **bear** ◯ 3 ◯ The above query is invalid.

Solution: Output table:

animal_type
bear

Pandas

12. Using each of the zoo tables (from the SQL questions) as Pandas dataframes:

aid	animal_type	name	age	color
0	rabbit	Bugs	2	white
1	bear	Air	5	golden
2	bear	Care	1	golden
3	cat	Grumpy	75	gray

(a) `animals`

did	name
0	Turk
1	House
2	Dre
3	Bailey

(b) `doctors`

vid	aid	did
132	1	0
145	2	1
167	0	3
168	2	3
169	2	2

(c) `visits`

Evaluate each of the following Python expressions:

(a) [2 Pts]
```
(doctors[doctors['name'] == 'Dre'][['did']]
    .merge(visits)
    .merge(animals)['name'][0])
```

◯ `'Air'` √ **`'Care'`** ◯ `'Grumpy'` ◯ `None`

(b) [3 Pts]
```
len(animals.merge(visits, on='aid')
    .groupby('color')[['age']].mean())
```

◯ 0 ◯ 1 √ **2** ◯ 3

(c) [3 Pts]
```
list(visits.groupby('did')
    .filter(lambda g: len(g) > 1)
    .merge(animals, on='aid')['name'])
```

◯ `['Bugs']` ◯ `['Care']` ◯ `['Air']` √ **`['Bugs', 'Care']`**

(d) [3 Pts]
```
list(animals.merge(visits, how='outer')
    .groupby(['aid'])
    .filter(lambda g: len(g['vid'].dropna()) == 0)
    ['name'])
```

◯ `['Bugs']` ◯ `['Care']` √ **`['Grumpy']`** ◯ `['Air']`

13. The following is the output of the command `taxi_df.head()` run on the `taxi_df` dataframe containing taxi trips in NYC. You may assume that there are **no missing values** in the dataframe and the **duration is measured in seconds**.

id	vendor_id	start_timestamp	passenger_count	duration
0	2	2016-06-08 07:36:19	1	1040
1	2	2016-04-03 12:58:11	1	827
2	2	2016-06-05 02:49:13	5	614
3	2	2016-05-05 17:18:27	2	867
4	1	2016-05-12 17:43:38	4	4967

(a) [1 Pt] Which of the following lines *returns a dataframe* containing *only* the rides with a *duration less than 3 hours.*

◯ `taxi_df[taxi_df['duration'] < 3]`

◯ `taxi_df.set_index('duration') < 3`

◯ `taxi_df['duration'] < 3 * 60 * 60`

√ `taxi_df[taxi_df['duration'] < 3 * 60 * 60]`

(b) [4 Pts] We would like to know the average and `duration` and for each `passenger_count` for each `vendor_id`, excluding (`vendor_id`, `passenger_count`) pairs for which we have less than 10 records. Which of the following returns this dataframe.

√
```
(taxi_df.groupby(['vendor_id', 'passenger_count'])
 .filter(lambda x: x.shape[0] >= 10)
 .groupby(['vendor_id', 'passenger_count'])
 .agg({'duration': 'mean'}))
```

◯
```
(taxi_df.groupby(['passenger_count'])
 .filter(lambda x: x.shape[0] >= 10)
 .groupby(['vendor_id', 'passenger_count'])
 .agg({'duration': 'mean'}))
```

◯
```
(taxi_df.groupby(['vendor_id', 'passenger_count'])
 .filter(lambda x: x.shape[0] < 10)
 .groupby(['vendor_id', 'passenger_count'])
 .agg({'duration': 'mean'}))
```

◯
```
(taxi_df.groupby(['vendor_id', 'passenger_count'])
 .filter(lambda x: x.shape[1] >= 10)
 .groupby(['vendor_id', 'passenger_count'])
 .mean())
```

Big Data

14. We want to store a big file on a distributed file system by splitting it into smaller fragments. Assuming the file divides evenly into **800** fragments and we use **4-way replication** answer the following questions.

    (a) [2 Pts] If the distributed file system contains 8 separate nodes, how many fragments of the file will be stored on each node?

    ◯ 100 √ **400** ◯ 800 ◯ 3200

    (b) [2 Pts] What is the maximum number of machines that can fail and still guarantee that we can read the entire file.

    ◯ 1 ◯ 2 √ **3** ◯ 4 ◯ 5

15. Which of the following statements are correct?

    (a) [1 Pt] In the star schema, the *dimension table* contains the relationships between different facts in the separate *fact tables*.

    ◯ True √ **False**

    (b) [1 Pt] Star schemas can help eliminate update errors by reducing duplication of data.

    √ **True** ◯ False

    (c) [1 Pt] During the *reduce phase* of MapReduce all the records associated with a given key are sent to the same machine.

    √ **True** ◯ False

    (d) [1 Pt] Because files are spread across multiple machines, reading a large file from a distributed file system is usually slower than reading the same large file from a single drive.

    ◯ True √ **False**

    (e) [1 Pt] When using MapReduce, we need to have a memory buffer that is big enough to load all the data from disk to memory.

    ◯ True √ **False**

Regular Expressions

16. [3 Pts] Given that we are using the regular expression `r"ta.*c"`, which option specifies the starting and ending position of the first match in the string: `"tacocat"`

◯ 0-2 ◯ 0-3 √ **0-5** ◯ 0-6 ◯ The string contains no matches.

17. Evaluate the following Python expressions that use the `re` module. Notes: (1) assume `import re` was already run; (2) The character "␣" represents a single space:

(a) [2 Pts]
```
re.findall(r"\d{3}\.\d{3}\.\d{4}",
           "123.456.7890␣and␣Fax␣800\999\0000")
```

◯ `None`

√ **`['123.456.7890']`**

◯ `['800\999\0000']`

◯ `['123.456.7890', '800\999\0000']`

(b) [2 Pts]
```
re.findall(r"\{[^}]*\}", "{begin}␣{and␣{end}}")
```

◯ `None`

◯ `['{begin}']`

◯ `['{begin}', '{end}']`

√ **`['{begin}', '{and␣{end}']`**

(c) [2 Pts]
```
len(re.findall(r"[cat]+|dog",
               "cat␣catch␣dog␣attack"))
```

◯ 1 ◯ 2 ◯ 3 √ **4** ◯ 5

(d) [2 Pts]
```
re.findall(r"<p>.*?</p>",
"<html><p>stuff?</p><p><body>more␣stuff</p></body>")
```

◯ `['<p>stuff?</p>']`

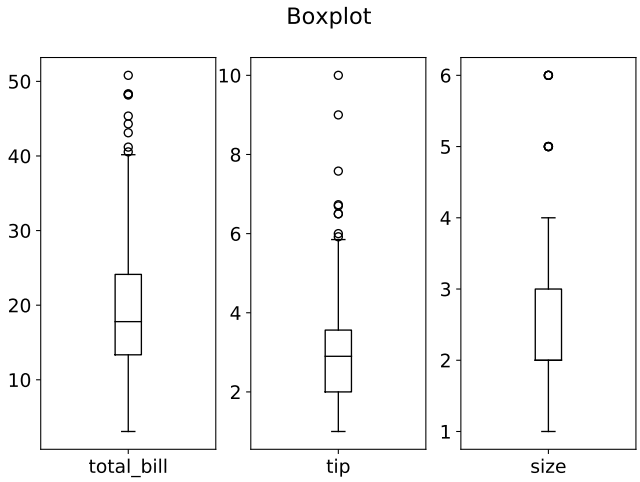
◯ `['<p>stuff?</p><p><body>more␣stuff</p>']`

√ **`['<p>stuff?</p>', '<p><body>more␣stuff</p>']`**

◯ `['<p>more␣stuff</p>']`

Visualization & EDA

18. Using the following box plots from lecture answer the following questions:



(a) [1 Pt] Which of piece of information is not communicated by these box plots.

√ **The number of observations in each box.**

◯ Quartiles (lower and upper) and median.

◯ A simple comparison between distributions across different groups.

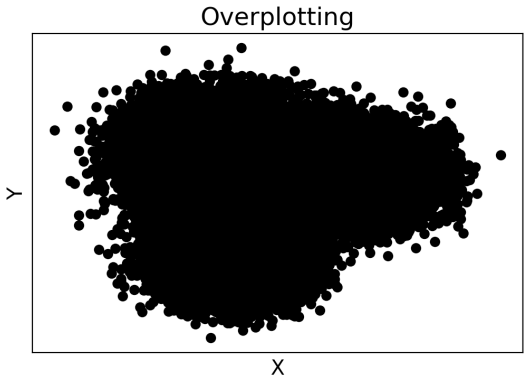
◯ Outliers values for each category.

Solution: Some box plots might only contain negligible amounts of data compared to others.

(b) [1 Pt] Using *only* the box plot we can tell that the `tip` distribution appears to be:

◯ Bimodal ◯ Unimodal ◯ Skewed left √ **Skewed right**

19. [2 Pts] Which of the following changes will *most* effectively improve the following plot to communicate the relationship between the two variables `x` and `y`?



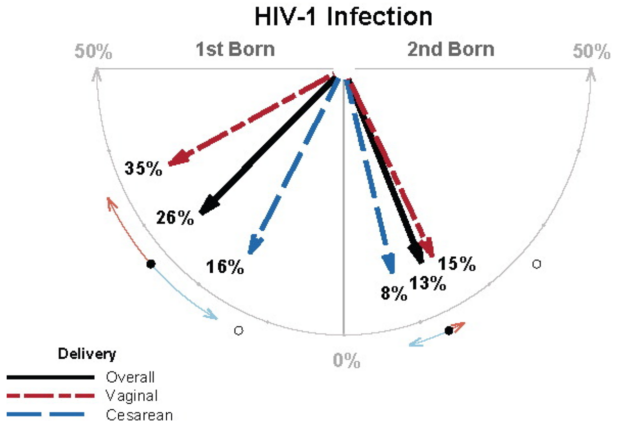
√ **Change the dot size and/or transparency.**

◯ Remove the outliers.

◯ Display a 1D histogram.

◯ Change the scale of `x` and `y`.

20. For this question consider the following dataframe (`compass_df`) and data visualization.



	born_order	delivery	perc
0	2nd	Cesarean	0.08
1	2nd	Overall	0.13
2	2nd	Vaginal	0.15
3	1st	Cesarean	0.16
4	1st	Overall	0.26
5	1st	Vaginal	0.35

(a) [1 Pt] Which plotting mistake best characterizes a problem with this plot.

- ◯ Use of stacking
- √ **Use of angles to compare magnitudes**
- ◯ Chart junk
- ◯ Overplotting

(b) [3 Pts] The original intent of this plot was to demonstrate that the 2^{nd} born child has a lower risk of HIV-1 infection. Which of the following snippets of code would generate a plot that best illustrates this trend?

√
```
sns.barplot(y='perc', x='delivery',
            hue='born_order', data=compass_df);
```

◯
```
sns.barplot(y='perc', x='born_order',
      hue='delivery', data=compass_df);
```

◯
```
sns.boxplot(y='perc', x='delivery',
      hue='born_order', data=compass_df);
```

◯
```
compass_df.plot(y='perc', kind='pie');
```

21. Suppose we have constructed a dataset about meals at restaurants around Berkeley. The following is just a *sample* of the dataset.

	total_bill	tip	day	party_size	date	place
0	16.99	1.01	Sun	2	2017-01-01	Taqueria El Buen Sabor
1	10.34	1.66	Mon	3	2017-01-02	Burger King
3	23.68	3.31	Wed	2	2017-01-04	Ichiraku Ramen
4	24.59	3.61	Thu	4	2017-01-05	Akatsuki Hideout

For each of the following scenarios, determine which plot type is **most appropriate** to reveal the distribution of and/or the relationships between the following variable(s).

(a) [1 Pt] The spread of the `total_bill` for each day of the week:

◯ Bar plot √ **Side-by-side boxplots** ◯ Scatter plot ◯ Contour Plot

Solution: This allows to compare distributions of total bill amount per gender and per day of week

(b) [1 Pt] The distribution of the `tip` field for meals at Taqueria El Buen Sabor:

√ **Histogram** ◯ Bar plot ◯ Line plot ◯ Contour plots

Solution: We are interested in the distribution of a continuous variable

(c) [1 Pt] Average tip for meals on each day from January 2017 to January 2018:

◯ Histogram √ **Line plot** ◯ Side-by-side boxplots ◯ Contour plots

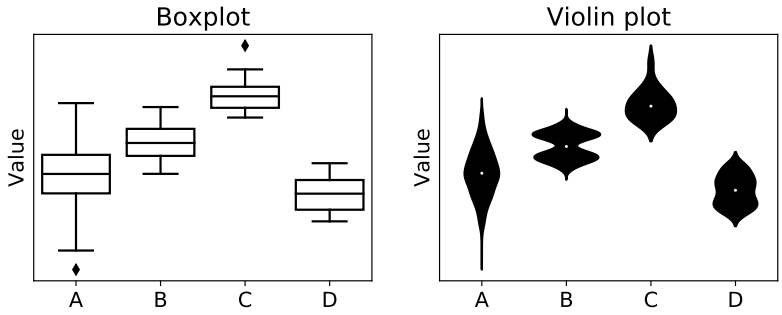
Solution: This allows us to see the trends over time

(d) [1 Pt] Number of meals for each place in 2017:

◯ Histogram √ **Bar plot** ◯ Side-by-side boxplots ◯ Scatter plot

Solution: This allows us to visualize counts of a categorical variable.

22. [2 Pts] What additional information does the violin plot on the right provide that is not present in the boxplot on the left.



◯ The violin plot displays the number of observations, which is hidden in the boxplot.

√ **The violin plot shows the underlying distribution into each group, e.g. we observe a bimodal distribution in group B.**

◯ The violin plot shows the number of missing values, which is hidden in the boxplot.

◯ The violin plot does not provide any additional information.

Modeling and Estimation

23. [6 Pts] What parameter estimate would minimize the following regularized loss function:

$$\ell(\theta) = \lambda(\theta - 4)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \tag{1}$$

- ◯ $\hat{\theta} = \frac{1}{\lambda n}\sum_{i=1}^{n} x_i$
- ◯ $\hat{\theta} = 4 + \frac{1}{\lambda n}\sum_{i=1}^{n} x_i$
- ◯ $\hat{\theta} = \frac{1}{n(\lambda+1)}\sum_{i=1}^{n} x_i$
- ◯ $\hat{\theta} = \frac{\lambda}{\lambda+1} + \frac{1}{n(\lambda+1)}\sum_{i=1}^{n}(x_i - 4)$
- √ $\hat{\theta} = \frac{4\lambda}{\lambda+1} + \frac{1}{n(\lambda+1)}\sum_{i=1}^{n} x_i$

You may use the space below for scratch work (not graded, no partial credit).

Solution:

Taking the derivative of the loss function we get:

$$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{\partial}{\partial\theta}\lambda(\theta - 4)^2 + \frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \tag{2}$$

$$= 2\lambda(\theta - 4) - \frac{2}{n}\sum_{i=1}^{n}(x_i - \theta) \tag{3}$$

$$\tag{4}$$

Setting the derivative equal to zero and solving for θ:

$$2\lambda(\theta - 4) = \frac{2}{n}\sum_{i=1}^{n}(x_i - \theta) \tag{5}$$

$$\lambda\theta - 4\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) - \theta \tag{6}$$

$$\lambda\theta + \theta = 4\lambda + \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$\theta = \frac{4\lambda}{\lambda + 1} + \frac{1}{n(\lambda + 1)}\sum_{i=1}^{n} x_i \tag{8}$$

$$\tag{9}$$

24. [8 Pts] Suppose $X_1, \ldots, X_n$ are random variables with $\mathbb{E}[X_i] = \mu^*$ and $\mathbf{Var}[X_i] = \theta^*$. Consider the following loss function

$$\ell(\theta) = \log(\theta) + \frac{1}{n\theta} \sum_{i=1}^{n} X_i^2.$$

Let $\widehat{\theta}$ denote the minimizer for $\ell(\theta)$. What is $\mathbb{E}[\widehat{\theta}]$?

◯ θ^* ◯ $\theta^* + \mu^*$ ◯ $\theta^* + \mu^*/2$ ◯ $\mathbb{E}[\theta^* + \mu^*]$ √ $\theta^* + (\mu^*)^2$ ◯ $(\theta^* + \mu^*)^2$

You may use the space below for scratch work (not graded, no partial credit).

Solution:

Taking the derivative of the loss function we get:

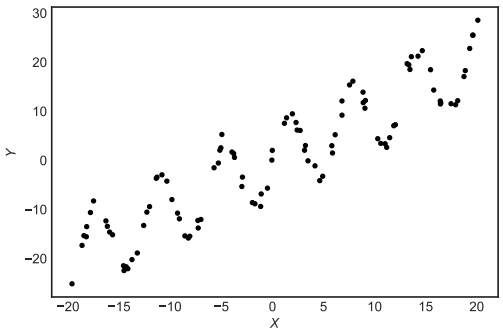
$$\ell'(\theta) = \theta^{-1} - \theta^{-2} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) \tag{10}$$

Setting this to zero yields $\widehat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right)$. Taking the expected value,

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}[X_1^2] = \text{Var}(X_1) + \mathbb{E}[X_1]^2 = \theta^* + (\mu^*)^2 \tag{11}$$

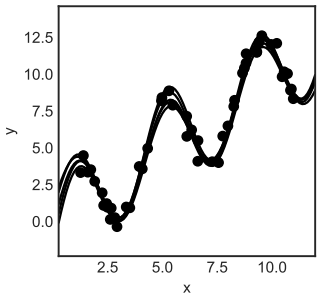
Regression

25. [2 Pts] Which model would be the most appropriate **linear model** for the following dataset?

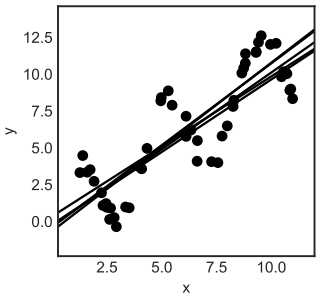


◯ $y = \theta_1 x + \theta_2$ ◯ $y = \sum_{k=1}^{d} \theta_k x^k$ √ $y = \theta_1 x + \theta_2 \sin(x)$ ◯ $y = \theta_1 x + \theta_2 \sin(\theta_3 x)$
◯ Since y is a non-linear function of x, the relationship can't be expressed by a linear model.

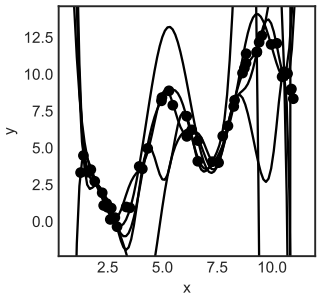
26. [2 Pts] Which of the following depicts models with the largest bias?



◯



√



◯

27. [4 Pts] Given a full rank feature matrix $\Phi \in \mathbb{R}^{nd}$ and response values $Y \in \mathbb{R}^n$ the following equation computes what quantity:

$$\Phi^T \left(Y - \Phi \left(\Phi^T \Phi \right)^{-1} \Phi^T Y \right) \tag{12}$$

√ **0** ◯ residuals ◯ squared residuals ◯ $\hat{Y}$ ◯ $\hat{\theta}$ ◯ squared error

28. [2 Pts] Which of the following loss functions is most sensitive to extreme outliers.

◯ L^1-Loss Function ◯ Squared Loss √ **Absolute Loss Function** ◯ Huber Loss

29. Given a dataset of 100 tweets where each tweet is no longer than 10 words. We apply a bag-of-words featurization to the tweets with a vocabulary of $10,000$ unique words plus an addition bias term. Our goal is to predict the number of retweets from the text of the tweet.

(a) [1 Pt] Because words are nominal this is a classification task.

◯ True √ **False**

(b) [2 Pts] The feature (covariate) matrix including bias term has 100 rows and how many columns (including zero valued columns)?

◯ 1 ◯ 10 ◯ 11 ◯ $10,000$ √ $10,001$ ◯ $10,010$ ◯ $10,011$

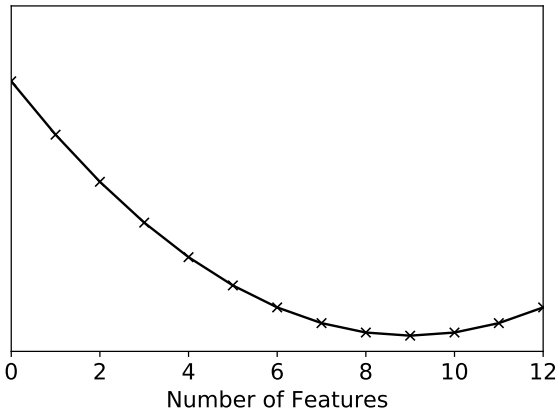
(c) [2 Pts] Because $n > d$ the solution to the normal equations is not well defined.

√ **True** ◯ False

(d) [2 Pts] By applying L^1 regularization we may be able to identify informative words.

√ **True** ◯ False

30. In the process of training linear models with different numbers of features you created the following plot but forgot to include the Y-axis label.



(a) [1 Pt] The Y-axis might represent the training error: ◯ True √ **False**

(b) [1 Pt] The Y-axis might represent the bias: ◯ True √ **False**

(c) [1 Pt] The Y-axis might represent the test error: √ **True** ◯ False

(d) [1 Pt] The Y-axis might represent the variance. ◯ True √ **False**

31. Consider the following model training script to estimate the training error:

```
1  X_train, X_test, y_train, y_test =
2      train_test_split(X, y, test_size=0.1)
3
4  model = lm.LinearRegression(fit_intercept=True)
5  model.fit(X_test, y_test)
6
7  y_fitted = model.predict(X_train)
8  y_predicted = model.predict(X_test)
9
10 training_error = rmse(y_fitted, y_predicted)
```

(a) [3 Pts] **Line 5** contains a serious mistake. Assuming our eventual goal is to compute the *training error*, which of the following corrects that mistake.

◯ `model.fit(X_train, y_test)`

√ **`model.fit(X_train, y_train)`**

◯ `model.fit(X, y)`

(b) [3 Pts] **Line 10** contains a serious mistake. Assuming we already have corrected the mistake in **Line 5** which of the following corrects the mistake on **Line 10**.

◯ `training_error = rmse(y_train, y_predicted)`

◯ `training_error = rmse(y_train, y_test)`

◯ `training_error = rmse(y_fitted, y_test)`

√ **`training_error = rmse(y_fitted, y_train)`**

32. [2 Pts] Which of the following techniques could be used to reduce over-fitting?

◯ Adding noise to the training data

√ **Cross-validation to remove features**

◯ Fitting the model on the test split

◯ Adding features to the training data

33. Suppose you are given a dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}$ is a one dimensional feature and $y_i \in \mathbb{R}$ is a real-valued response. To model this data, you choose a model characterized by the following loss function:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \theta_0 - x_i^3\theta_1\right)^2 + \lambda|\theta_1| \tag{13}$$

For the following statements, indicate whether it is True or False.

(a) [1 Pt] This model includes a bias/intercept term.

√ **True** ◯ False

(b) [1 Pt] As λ decreases to smaller values, the model will reduce to a constant θ_0

◯ True √ **False**

(c) [1 Pt] Larger λ values help reduce the chances of overfitting.

√ **True** ◯ False

(d) [1 Pt] Increasing λ decreases model variance.

√ **True** ◯ False

(e) [1 Pt] The training error should be used to determine the best value for λ.

◯ True √ **False**

Stochastic Gradient Descent (Going Downhill Quickly!)

34. Consider the following broken Python implementation of *stochastic* gradient descent.

```
1  def stochastic_grad_descent(
2              X, Y, theta0, grad_function,
3              max_iter = 10000, batch_size=2):
4      """
5      X: A 2D array, the feature matrix.
6      Y: A 1D array, the response vector.
7      theta0: A 1D array, the initial parameter vector.
8      grad_function: Maps a parameter vector, a feature matrix,
9          and a response vector to the gradient of some loss
10         function at the given parameter value.
11     batch_size: the number of data points to use in each
12         gradient estimate
13     returns the optimal theta
14     """
15     theta = theta0
16     ind = np.random.choice(len(Y), replace=False,
17                            size=batch_size)
18     for t in range(1, max_iter+1):
19         (xbatch, ybatch) = (X[ind, :], Y[ind])
20         grad = grad_function(theta, xbatch, ybatch)
21         theta = theta + t / grad
22     return theta
```

(a) [3 Pts] Which of the following best describes the bug in how data are sampled?

- ◯ The call to sample on **Line 16** should have been with replacement.
- √ **A new random sample of indices should be constructed on each loop iteration.**
- ◯ Like the bootstrap, each random sample should be the size of the original data set (i.e., `size = len(Y)` in **Line 17**)
- ◯ The `len(Y)` on **Line 16** should be `len(X)`.

(b) [3 Pts] Assuming that the stochastic gradient `grad` is computed correctly, what is the correct implementation of the gradient update on **Line 21**:

- ◯ `theta = theta - t / grad`
- ◯ `theta = theta - t * grad`
- ◯ `theta = theta + 1/t * grad`
- √ **`theta = theta - 1/t * grad`**

(c) [5 Pts] Suppose we wanted to add L^2 regularization with each dimension having a different regularization parameter:

$$R_\lambda(\theta) = \sum_{k=1}^{d} \lambda_k \theta_k^2 \tag{14}$$

where λ is now a vector of regularization parameters. Which of the following rewrites of **Line 20** would achieve this goal (assuming $\lambda =$ `lam`):

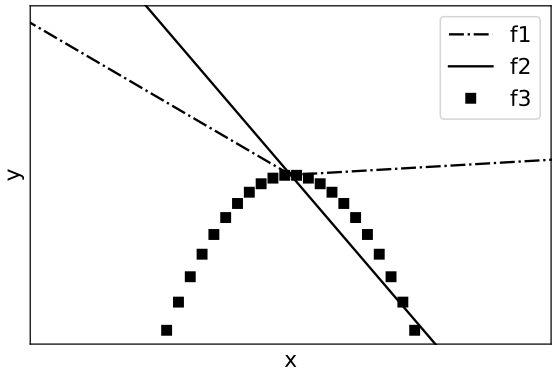
√ **`grad = (grad_function(theta, xbatch, ybatch) + 2*theta*lam)`**

◯ `grad = (grad_function(theta, xbatch, ybatch) + theta.dot(lam))`

◯ `grad = (grad_function(theta, xbatch, ybatch) - theta.dot(lam))`

◯ `grad = (grad_function(theta, xbatch, ybatch) - 2*theta*lam)`

35. Use the following plot to answer each of the following questions about convexity:



(a) [1 Pt] $f_1(x) = \max(0.01x, -x)$ is convex. √ **True** ◯ False

(b) [1 Pt] $f_2(x) = -2x$ is convex. √ **True** ◯ False

(c) [1 Pt] $f_3(x) = -x^2$ is convex. ◯ True √ **False**

(d) [1 Pt] $f_4(x) = f_1(x) + f_2(x)$ is convex. √ **True** ◯ False

Classification

36. [4 Pts] True or False.

(a) A binary (0/1) classifier that always predicts 1 can get 100% precision, and its recall will be the fraction of ones in the training set.

◯ True √ **False**

(b) If the training data is linearly separable we expect a logistic regression model to obtain 100% training accuracy.

√ **True** ◯ False

(c) We should use classification if the response variable is categorical.

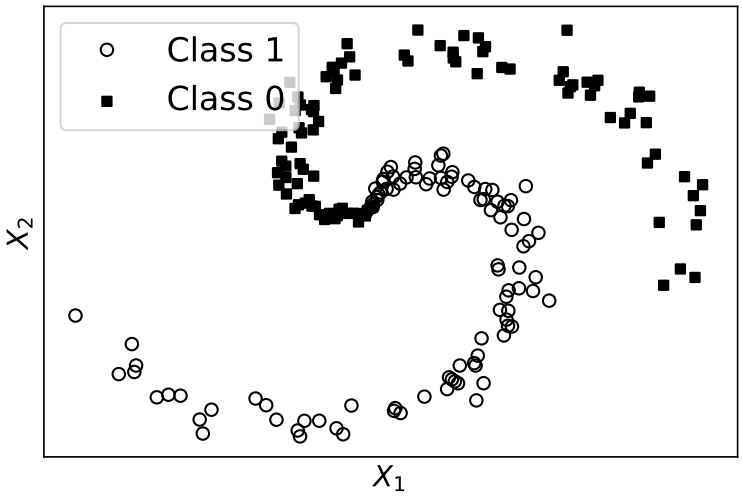
√ **True** ◯ False

(d) A binary classifier that only predicts class 1 may still achieve 99% accuracy on some prediction tasks.

√ **True** ◯ False

37. [2 Pts] The plot below is a scatter plot of a dataset with two dimensional features and binary labels (e.g., Class 0 and Class 1). Without additional feature transformations, is the this dataset linearly separable?

◯ Yes. √ **No.** ◯ We cannot tell that from this plot.



38. [4 Pts] We perform a 4-fold cross validation on 4 different hyper-parameters, the mean square error are shown in the table below. Which λ should we select?

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Max	Row Min	Row Avg
1	80.2	84.1	70.1	91.2	91.2	70.1	83.36
2	76.8	77.3	83.3	88.8	88.8	76.8	83
3	81.5	74.5	81.6	86.5	86.5	74.5	82.12
4	79.4	75.2	79.2	85.4	85.4	75.2	80.92
Col Avg	79.475	77.775	78.55	87.975			

◯ $\lambda = 0.1$ √ $\lambda = 0.2$ ◯ $\lambda = 0.3$ ◯ $\lambda = 0.4$

39. [4 Pts] Answer **true** or **false** for each of the following statements about logistic regression:

(a) If no regularization is used and the training data is linearly separable, the optimal model parameters will tend towards positive or negative infinity.

√ **True** ◯ False

(b) After using L^2 regularization, the optimal model parameter will be the mean of the data, since L^2 regularization is similar to the square loss.

◯ True √ **False**

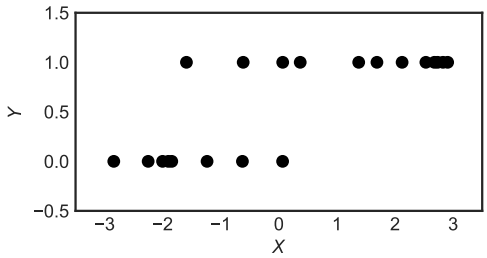
(c) L^1 regularization can help us select a subset of the features that are important.

√ **True** ◯ False

(d) After using the regularization, we expect the training accuracy to increase and the test accuracy to decrease.

◯ True √ **False**

40. [2 Pts] Suppose you are given the following dataset $\{(x_i, y_i)\}_{i=1}^n$ consisting of x and y pairs where the covariate $x_i \in \mathbb{R}$ and the response $y_i \in \{0, 1\}$.



Given this data, the value $\mathbf{P}(Y = 1 \mid x = -1)$ is likely closest to:

◯ 0.95 √ **0.50** ◯ 0.05 ◯ -0.95

Statistical Inference

41. [3 Pts] True or False.

(a) A 95% confidence interval is wider than a 90% confidence interval.

√ **True** ◯ False

(b) A p-value of 0.97 says that under the null model, there is a 97% chance of observing a test statistic at least as extreme as the one calculated from the data.

√ **True** ◯ False

(c) Suppose we have 100 samples drawn independently from a population. If we construct a separate 95% confidence interval for each sample, 95 of them will include the **population** mean.

◯ True √ **False**

42. A roulette wheel has 2 green slots, 18 red slots, and 18 black slots. Suppose you observe 760 games on a particular wheel and see that the red slot is chosen 380 times.

(a) [2 Pts] What is the expected number of times that the red slot is chosen?

◯ $\frac{760}{2}$ √ $\frac{18}{38} * 760$ ◯ $\left(1 - \frac{18}{38}\right) * 760$ ◯ $\frac{760}{18}$

(b) [2 Pts] You hypothesize that the wheel has been altered so that the frequency of landing on red is higher than by chance. Given the data generation model that landing on red is the outcome of a Bernoulli trail with probability p, which of the following would be the most appropriate null hypothesis?

◯ $p = 0$ √ $p = \frac{18}{38}$ ◯ $p = \frac{1}{2}$ ◯ $p = 1$

(c) [2 Pts] You decide to study this problem by running a simulation of $N = 10,000$ replications of a fair roulette wheel constructed as described above. The percentiles for the proportion of red is shown below:

Percentile	2.5%	5%	10%	50%	90%	95%	97.5%
Proportion	0.438	0.445	0.450	0.474	0.497	0.504	0.509

Using the 5% convention for statistical significance, is the null model consistent with your observations? √ **Yes** ◯ No ◯ Not enough information given to answer

Solution: Since the proportion of reds observed is 0.500 and the 95%ile is greater than this, the null model is consistent with our observations.

43. Two methods of memorizing words are to be compared. You deterministically pair 1,000 people in the study together, manually matching so that the two people in each pair have very similar education levels and ages. For each of the 500 pairs of people, you randomly assign one person to memorization method 1 and the other to method 2. After a week of training, the number of words recalled in a memory test is recorded. A portion of the data is shown below:

Participant ID	Pair ID	Memorization Method	Age Group	Education	Words Recalled
1	1	1	18-25	High School	25
2	1	2	18-25	High School	21
3	2	2	26-35	Undergraduate	20
4	2	1	26-35	Undergraduate	30
...	...	...	...	...	...
999	500	1	55-65	Masters	29
1000	500	2	55-65	Masters	17

(a) [2 Pts] The null hypothesis for this experiment is that there is no difference in the average number of words recalled

√ **across memorization methods.** ◯ across Pair IDs. ◯ across age groups.
◯ across education levels.

(b) [2 Pts] Which of the following describes a reasonable test statistic for this experiment? By "ungroup", we mean remove any lingering effect of the `group_by` operation.

√ **1. Group by pair ID.**
2. Subtract words recalled for memorization method 2 from method 1.
3. Ungroup.
4. Take the average of the differences from step 2.

◯ 1. Group by memorization method, age group, education.
2. Take the average of words recalled. Group by age group and education.
3. Subtract words recalled for memorization method 2 from method 1.
4. Ungroup.
5. Take the average of the differences from step 3.

(c) Instead you decide to use permutation test to analyze the data. What permutation is justified by the design of the experiment?

i. [1 Pt] Group by

√ **Pair ID** ◯ Age Group ◯ Education
◯ Words Recalled ◯ None of the above

ii. [1 Pt] and permute the values of

◯ Pair ID ◯ Age Group ◯ Education
√ **Memorization Method** ◯ None of the above

Probability

44. There are 32 participants in a randomized clinical trial: 8 are male and 24 are female. 16 are assigned to treatment and the others are put into the control group. What is the probability that none of the men are in the treatment group if:

(a) [3 Pts] the treatment was assigned using stratified random sampling, grouping by gender?

○ $\binom{32}{8}/\binom{32}{16}$ ○ $\left(1-\frac{8}{32}\right)^{16}$ ○ $\prod_{i=0}^{15}\frac{24-i}{32-i}$ √ **0**

(b) [4 Pts] the treatment was assigned using simple random sampling?

○ $\left(1-\frac{8}{32}\right)^{16}$ √ $\frac{24!16!}{8!32!}$ ○ $\prod_{i=1}^{16}\binom{16-i}{i}$ ○ $1/\binom{32}{16}$

(c) [4 Pts] the treatment was assigned using cluster random sampling of 2 groups of 8 using clusters as described below?

Cluster	Male	Female
A	0	8
B	3	5
C	5	3
D	0	8

○ 0 ○ $\frac{1}{2}$ √ $\frac{1}{6}$ ○ $\frac{1}{8}$

Ethics

45. [2 Pts] During the guest lecture on ethics, Josh Kroll presented a case study on facial recognition. In it, Joy Buolamwini, an MIT PhD student, argues that current facial recognition technology:

    ◯ Could lead to issues concerning the right to privacy

    ◯ Straddles a gray zone on data usage agreements

    √ **Does not perform well on certain subpopulations**

    ◯ Creates an asymmetry of power between those who have data and those who do not

46. [2 Pts] PredPol is a model that helps police determine where they should patrol to maximize their likelihood of spotting crimes. As presented in lecture, which of the following best explains why PredPol was consistently suggesting low-income minority neighborhoods as areas that needed more policing?

    √ **Past arrest rates in those neighborhoods are higher than in their higher-income, non-minority counterparts.**

    ◯ The contracted company developing the model introduced systematic biases into the model to advance a political agenda.

    ◯ Location data was missing for most of the police reports. Cases that had location data were often in lower-income minority neighborhoods.

    ◯ There was a flaw in the data cleaning that led to an aggregation of crime cases into particular low-income minority communities.

Last Thoughts

This page was intentionally left blank for you! Feel free to use it as scrap paper, draw a picture, write a song, or just tell us how you feel now that the semester is over.