DS-100 Final Exam

Fall 2017

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Instructions:

- This final exam must be completed in the **3 hour time** period ending at **11:00AM**.
- Note that some questions have 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**.
- When selecting your choices, you must **shade** in the box/circle. Check marks will likely be mis-graded.
- You may use a two page (two-sided) study guide.
- Work quickly through each question. There are a total of 127 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

"\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

XPath

An XPath expression is made up of location steps separated by forward slashes. Each location step has three parts: an axis, which gives the direction to look; a node test which indicates the node name or text(); and an optional predicate to filter the matching nodes:

axis::node[predicate]

We have used shortcut names for the axis: "." refers to self, "//" refers to self or descendants, ".." refers to to parent, and child is the default axis and can be dropped. The node of the XPath expression is either an element name or `text()` for text content or @attribute for an attribute.

The predicate contains an expression that evaluates to true or false. Only those nodes that evaluate to true are kept. To check whether an attribute is present in a node, we use, e.g., `[@time]` (this evaluates to true if the node has a time attribute). Similarly, `[foo]` evaluates to true if the node has a child node named foo. The value of an attribute can be checked with, e.g., `[@time = "2017"]`.

Variance and Expected Value Calculations

The expected value of X is

$$\mathbf{E}[X] = \sum_{j=1}^{m} x_j p_j$$

The variance of X is

$$\mathbf{Var}[X] = \sum_{j=1}^{m} (x_j - \mathbf{E}[X])^2 p_j = \sum_{j=1}^{m} x_j^2 p_j - \mathbf{E}[X]^2 = \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2$$

The standard deviation of X is $\mathbf{SD}[X] = \sqrt{\mathbf{Var}[X]}$.

For $X_1, \ldots X_n$,

$$\mathbf{E}[a_1 X_1 + a_2 X_2 + \cdots + a_n X_n] = a_1 \mathbf{E}[X_1] + \cdots + a_n \mathbf{E}[X_n] \; =$$

If the X_i are independent, then

$$\mathbf{Var}[a_1 X_1 + a_2 X_2 + \cdots + a_n X_n] = a_1^2 \mathbf{Var}[X_1] + \cdots + a_n^2 \mathbf{Var}[X_n]$$

In the special case where $\mathbf{E}[X_i] = \mu$, $\mathbf{Var}[X_i] = \sigma^2$, $a_i = 1/n$ and the X_i are independent, then we have

$$\mathbf{E}\left[\bar{X}\right] = \mu \qquad \mathbf{Var}\left[\bar{X}\right] = \sigma^2/n \qquad \mathbf{SE}\left[\bar{X}\right] = \sigma/\sqrt{n}$$

PySpark

`sc.textFile(filename)` Creates an RDD from the `filename` with each line in the file as a separate record.

`rdd.collect()` Takes an `rdd` and returns a Python list.

`rdd.filter(f)` Applies the function `f` to each record in `rdd` and keeps all the records that evaluate to `True`.

`rdd.map(f)` Applies the function `f` to each record in `rdd` producing a new RDD containing the outputs of `f`.

`rdd.mapValues(f)` Takes an `rdd` of key-value pairs (lists) and applies the function `f` to each record in `rdd` producing a new RDD containing the outputs of `f`.

`rdd.reduceByKey(f)` Takes an `rdd` of key-value pairs (lists). It then groups the values by the key and applys the reduce function `f` to combine (e.g., sum) all the values returning an RDD of `[key,sum(values)]` lists

`s.split()` Splits a string on whitespace.

`np.array(list)` Constructs a vector from a list of elements.

Data Cleaning, Regular Expressions, and XPath

1. Consider the following text data describing purchases of financial products:

Id	Date	Product	Company
0	99/99/99	Debt collection	California Accounts Service
1	06/15/10	Credit reporting	EXPERIAN INFORMATION SOLUTIONS INC
3	10/21/14	MORTGAGE	OCWEN LOAN SERVICING LLC
5	03/30/15		The CBE Group Inc
6	02/03/16	Debt collection	The CBE Group, Inc.
7	01/07/17	Credit reporting	Experian Information Solutions Inc.
8	03/15/17	Credit card	FIRST NATIONAL BANK OF OMAHA

(1) [2 Pts] Select all the true statements from the following list.

- √ **Some of the product values appear to be missing.**
- √ **Some of the date values appear to be missing.**
- ☐ The file is comma delimited
- √ **The file is fixed width formatted.**
- √ **To analyze the companies we will need to correct for variation in capitalization and punctuation.**
- ☐ None of the above statements are true.

(2) [2 Pts] Select all of the following regular expressions that properly match the dates.

- ☐ `\d?/\d?/\d?`
- √ **`\d+/\d+/\d+`**
- √ **`\d*/\d*/\d*`**
- √ **`\d\d/\d\d/\d\d`**
- ☐ None of the above regular expressions match.

(3) [2 Pts] which of the following regular expressions exactly matches the entry `FIRST NATIONAL BANK OF OMAHA`? Select all that matches.

- ☐ `[A-Z]*`
- √ **`FIR[A-Z,\s]* OMAHA`**
- √ **`F[A-Z, \s]+A`**
- ☐ `F[A-Z]*`
- ☐ None of the above regular expressions match.

Solution: The last one is a bit tricky. The expression `F[A-Z, \s]*` also matches the 'F' in "Student Loan Finance Corporation".

2. Consider the following HTML document:

```
<html>
<head></head>
<body>
<h1>Hello!</h1>
<p>
my story is <a href="www.xxx">here</a> and it's <em>silly</em>.
</p>
<table id="sym">
<tr><th>Name</th><th>Instrument</th> </tr>
<tr><td><a href="www.yyy">Abe</a></td><td>violin</td> </tr>
<tr><td>Amy</td><td>violin</td> </tr>
<tr><td>Dan</td><td>viola</td> </tr>
<tr><td><a href="www.ccc">Cal</a></td><td>trumpet</td> </tr>
</table>
<table id="xyz">
<tr><th>Name</th><th>Instrument</th> </tr>
<tr><td>Sally</td><td>bass</td> </tr>
<tr><td><a href="www.ter">Terry</a></td><td>guitar</td> </tr>
<tr><td>Cassie</td><td>drums</td> </tr>
<tr><td>Tobie</td><td>piano</td> </tr>
</table>
<p>The End!</p>
</body>
</html>
```

(1) [2 Pts] Which of the following XPath queries locates the p-elements in the document? Select **all** that apply.

√ **//p** √ **//table/../p** √ **//body//p** √ **./body/p**

(2) [2 Pts] What will be the result of XPath query: `./body/table/tr/td/a/text()`

◯ www.yyy ◯ Abe √ **[Abe,Cal,Terry]** ◯ [www.yyy,www.ccc,www.ter]

(3) [2 Pts] Which of the following XPath queries locates the names of all musicians in the second table (i.e., Sally, Cassie, and Tobie)? Select **all** that apply.

☐ `//table[@id]//td/text()`
☐ `./body/table[2]/text()`
√ **`//table[@id="xyz"]/tr/td[1]/text()`**
☐ `//tr/td[1]/text()`
☐ None

The query from the previous page is repeated below for quick reference.

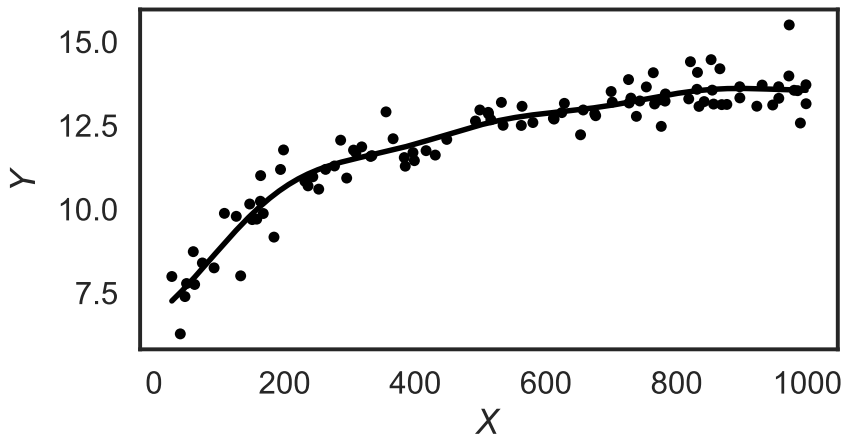
```
<html>
<head></head>
<body>
<h1>Hello!</h1>
<p>
my story is <a href="www.xxx">here</a> and it's <em>silly</em>.
</p>
<table id="sym">
<tr><th>Name</th><th>Instrument</th> </tr>
<tr><td><a href="www.yyy">Abe</a></td><td>violin</td> </tr>
<tr><td>Amy</td><td>violin</td> </tr>
<tr><td>Dan</td><td>viola</td> </tr>
<tr><td><a href="www.ccc">Cal</a></td><td>trumpet</td> </tr>
</table>
<table id="xyz">
<tr><th>Name</th><th>Instrument</th> </tr>
<tr><td>Sally</td><td>bass</td> </tr>
<tr><td><a href="www.ter">Terry</a></td><td>guitar</td> </tr>
<tr><td>Cassie</td><td>drums</td> </tr>
<tr><td>Tobie</td><td>piano</td> </tr>
</table>
<p>The End!</p>
</body>
</html>
```

(4) [3 Pts] Which of the following XPath queries locates the instruments of all musicians with Web pages? (A musician has a Web page if there is an a-tag associated with their name. Select **all** that apply.

- √ **`//td/a/../../td[2]/text()`**
- ☐ `//a/ancestor-or-self::table/tr/td[2]/text()`
- √ **`//table/tr/td[a]/../td[2]/text()`**
- ☐ `//tr/td[a]/text()`
- ☐ None

Visualization

3. [2 Pts] Which of the following transformations could help make linear the relationship shown in the plot below? **Select all that apply**:



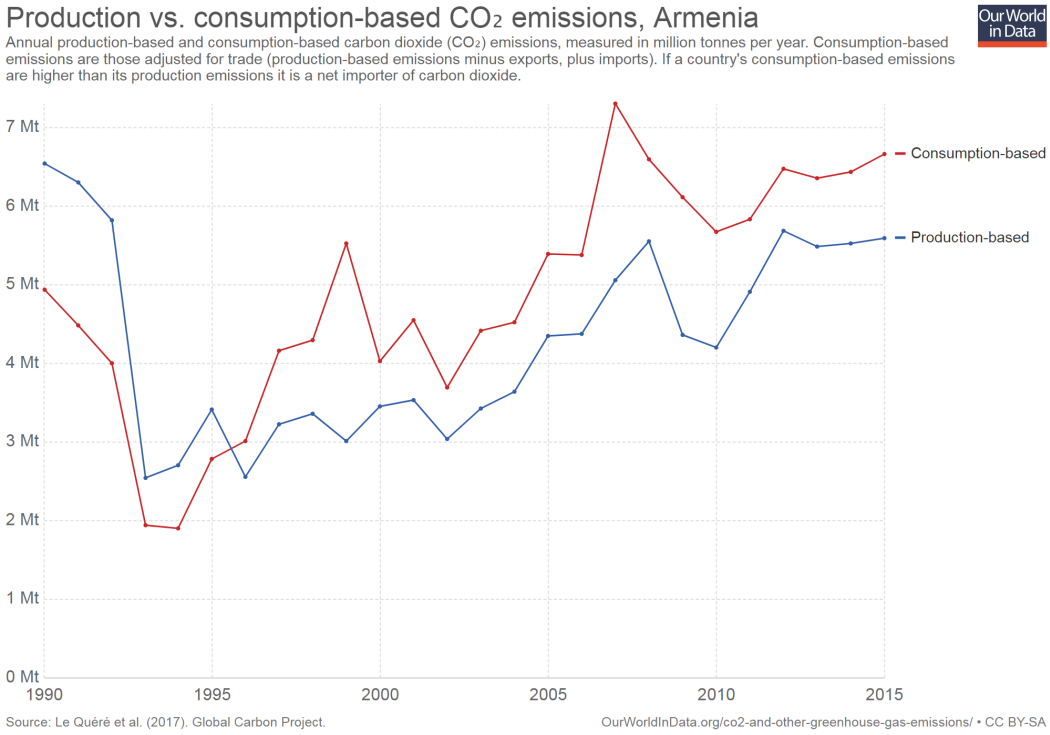
☐ $\log(y)$ ☐ x^2 √ $\sqrt{x}$ √ $\log(x)$ √ y^2 ☐ None

4. [2 Pts] Which graphing techniques can be used to address problems with over-plotting? Check all that apply.

☐ jiggling √ **transparency** √ **smoothing** √ **faceting**

☐ banking to 45 degrees √ **contour plotting** ☐ linearizing

5. The following line plot compares the annual production-based and consumption-based carbon dioxide emissions (million tons) in Armenia.



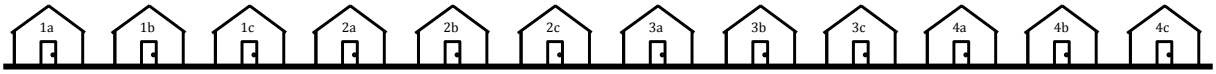
(1) [2 Pts] This plot best conveys:

- ◯ The relative increase in CO_2 emissions since 1990.
- √ **The overall trend in CO_2 emissions broken down by source.**
- ◯ The relative breakdown of CO_2 emissions sources over time.
- ◯ The cumulative CO_2 emissions.

(2) [2 Pts] What kind of plot would facilitate the relative comparison of the these two sources of emissions over time?

- ◯ stacked barchart
- ◯ side-by-side boxplots
- √ **line plot of annual differences**
- ◯ scatter plot of production-based emissions against consumption-based emissions

Sampling



6. Kalie wants to measure interest for a party on her street. She assigns numbers and letters to each house on her street as illustrated above. She picks a letter "a", "b", or "c" at random and then surveys every household on the street ending in that letter.

(1) [1 Pt] What kind of sample has Kalie collected?

◯ Quota Sample √ **Cluster Sample**
◯ Simple Random Sample ◯ Stratified Sample

Solution: Each group of houses, grouped by the last letter of the address, is a cluster.

(2) [1 Pt] What is the chance that two houses next door to each other are both in the sample?

◯ $\frac{1}{3}$ ◯ $\frac{1}{9}$ ◯ $\frac{1}{6}$ √ 0

Solution: None of the adjacent houses end in the same letter, so the chance is zero.

For the remaining parts of this question, suppose that $\frac{1}{2}$ of the houses ending in "a" favor the party; $\frac{3}{4}$ of the houses ending in "b" favor the party; and all of the houses ending in "c" favor the party. Hence, overall, $p = \frac{3}{4}$ of the houses favor the party.

(3) [4 Pts] If Kalie estimates how favorable the party is using the proportion $\widehat{p}$ of households in her survey favoring the party, what is the expected value of her estimator $\mathbb{E}[\widehat{p}]$? Show your work in the space below.

◯ $\frac{1}{2}$ ◯ $\frac{2}{3}$ √ $\frac{3}{4}$ ◯ 1

Solution: The expected value is an average, weighted by the probability of each value. Since each cluster is equally likely, $\mathbb{E}[\widehat{p}] = \frac{1}{3}\frac{1}{2} + \frac{1}{3}\frac{3}{4} + \frac{1}{3}1 = \frac{3}{4}$

(4) [6 Pts] If, as before, Kalie estimates how favorable the party is using the proportion $\widehat{p}$ of households in her survey favoring the party, what is the variance of her estimator $\text{Var}[\widehat{p}]$? Show your work in the space below.

◯ $\frac{1}{9}$ ◯ $\frac{2}{27}$ ◯ $\frac{1}{6}$ √ $\frac{1}{24}$

Solution: $\text{Var}[\widehat{p}] = \mathbb{E}\left[(\widehat{p} - p)^2\right] = \frac{1}{3}\left(\frac{1}{16} + 0 + \frac{1}{16}\right) = \frac{1}{3 \cdot 8} = \frac{1}{24}$.

SQL

7. [2 Pts] From the following list, select all the statements that are true:

☐ A *database* is a *system* that stores data.

> **Solution:** A database is a collection of data. A database management system (DBMS) is the system that manages access to the database.

√ **SQL is a declarative language that specifies what to produce but not how to compute it.**

> **Solution: SQL is declarative programming language which specifies *what* the user wants to accomplish allowing the system to determine *how* to accomplish it.**

☐ To do large scale data analysis it is usually faster to extract all the data from the database and use Pandas to execute joins and compute aggregates.

> **Solution:** Doing analysis directly in the database can often be much more efficient as database management systems are designed to accelerate data access and aggregation over very large datasets.

☐ The schema of a table consists of the data stored in the table.

> **Solution:** The schema of a table consists of the column names, their types, and any constraints on those columns. The instance of a database is the data stored in the database.

√ **The primary key of a relation is the column or set of columns that determine the values of the remaining column.**

☐ None of the above statements are true.

8. [4 Pts] The following relational schema represents a large table describing Olympic medalists.

```
MedalAwards(year, athlete_name, medal,
            event, num_competitors,
            country, population, GDP)
```

If we allow athletes to compete for different countries on different years and in multiple events, which of the following *normalized* representations most reduces data redundancy while encoding the same information.

◯
```
MedalAwards(year, athlete_name, medal, event)
Athlete(year, athlete_name, country, event,
        num_competitors, population, GDP)
```

◯
```
MedalAwards(year, athlete_name, medal, event)
Athlete(year, athlete_name, country, event,
        num_competitors)
CountryInfo(year, country, population, GDP)
```

√
```
MedalAwards(year, athlete_name, medal, event)
Events(year, event, num_competitors)
Athlete(year, athlete_name, country)
CountryInfo(year, country, population, GDP)
```

◯
```
MedalAwards(year, athlete_name, medal, event)
Events(event, num_competitors)
Athlete(athlete_name, country)
CountryInfo(country, population, GDP)
```

9. For this question you will use the the following database consisting of three tables:

```
CREATE TABLE medalist(
        name TEXT PRIMARY KEY,
        country TEXT,
        birthday DATE);

CREATE TABLE games(
        year INT PRIMARY KEY,
        city TEXT,
        country TEXT
        );

-- medaltype column takes three values:
-- 'G' for gold, 'S' for silver,
-- and 'B' for bronze
CREATE TABLE medals(
    name TEXT,
    year INT,
    FOREIGN KEY name REFERENCES medalist,
    FOREIGN KEY year REFERENCES games,
    category TEXT,
    medaltype CHAR);
```

(1) [1 Pt] Which of the following queries returns 5 rows from the `medalist` table (select all that apply):

☐ `SELECT * FROM medalist WHERE LEN(*) < 5;`

√ `SELECT * FROM medalist LIMIT 5;`

☐ `SELECT * FROM medalist HAVING LEN(*) < 5;`

☐ `FROM medalist SELECT * WHERE COUNT(*) < 5;`

(2) [1 Pt] Which of the following queries returns the names of all the German `medalist` (select all that apply):

√ `SELECT name FROM medalist WHERE country = 'Germany';`

☐ `FROM medalist SELECT name WHERE country = 'Germany';`

☐ `SELECT name FROM medalist HAVING country == 'Germany';`

☐ `FROM medalist SELECT name HAVING country IS 'Germany';`

Solution: The second solution has a wrong order of `FROM` and `SELECT`; the third and fourth use `HAVING` which is to be used only after `GROUP BY` clause.

Summarizing the schema on the pervious page for quick refence:

```
medalist(name, country, birthday);
games(year, city, country);
medals(name, year, category, medaltype);
```

(3) [3 Pts] Which of the following queries returns the total number of medals broken down by type (gold, silver, and bronze) for each country in the `'vault'` competition. (Select all that apply.)

√
```
SELECT  medalists.country,
        medals.medaltype,
        COUNT(*) AS medal_count
FROM medals, medalists
WHERE medalists.name = medals.name
AND medals.category = 'vault'
GROUP BY medalists.country, medals.medaltype
```

☐
```
SELECT  games.country,
    medals.medaltype,
    COUNT(medals.medaltype) AS medal_count
FROM medals, games
AND games.year = medals.year
HAVING medals.category = 'vault'
GROUP BY games.country, medals.medaltype
```

√
```
SELECT  medalists.country,
        medals.medaltype,
        COUNT(*) AS medal_count
FROM medals, medalists
WHERE medalists.name = medals.name
GROUP BY medalists.country, medals.medaltype, medals.category
HAVING category = 'vault'
```

☐
```
FROM medals, games
SELECT games.country,
    medals.medaltype,
    COUNT(medals.medaltype) AS medal_count
AND games.year = medals.year
AND medals.category = 'vault'
GROUP BY games.country, medals.medaltype
```

Solution: Both first and third solutions will technically return the desired result (the first solution might be faster as the filtering is performed before grouping). The second

solution applies `HAVING` in a wrong place (before the `GROUP BY` statement). The last solution has wrong order of `FROM` and `SELECT`

Summarizing the schema on the pervious page for quick refence:

```
medalist(name, country, birthday);
games(year, city, country);
medals(name, year, category, medaltype);
```

(4) [5 Pts] What does the following query compute?

```
WITH
    country_medal_count(country, count) AS (
        SELECT medalists.country, count(*) AS
        FROM medalists JOIN medals
        ON medalists.name = medals.name
        GROUP BY country
    ),
    annual_medal_count(country, year, count) AS (
        SELECT medalists.country, medals.year, count(*)
        FROM medalists JOIN medals
        ON medalists.name = medals.name
        GROUP BY medalists.country, year
    )
SELECT cmc.country, amc.year, amc.count / cmc.count
FROM country_medal_count AS cmc, annual_medal_count AS amc
WHERE cmc.country = amc.country
GROUP BY cm.country
```

◯ The average number of medals earned for each country in each year.

√ **The conditional distribution of medals over the years given the country.**

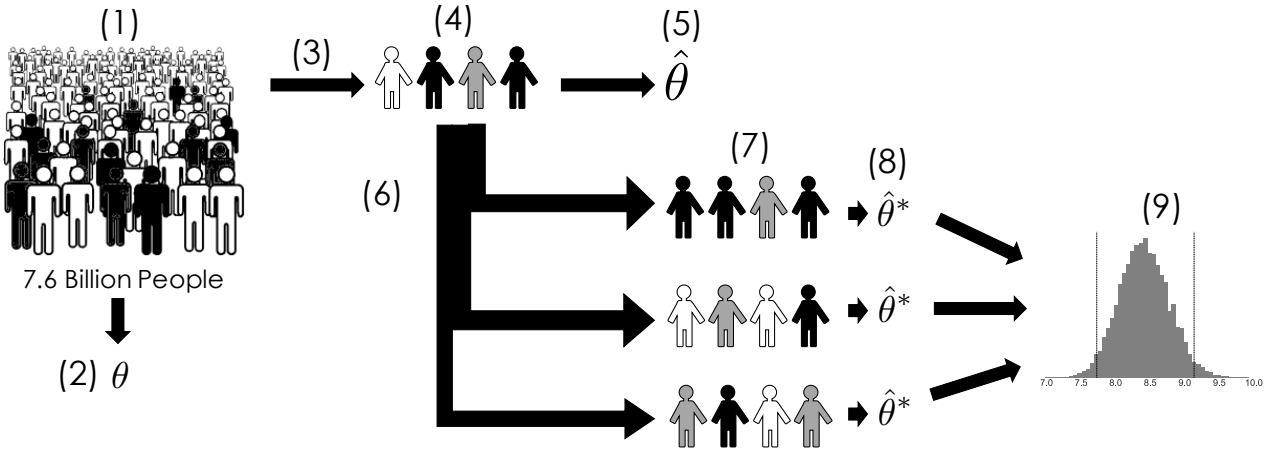
◯ The conditional distribution of medals over countries given the year.

◯ The joint distribution of medals over countries and years.

Bootstrap Confidence Intervals

10. [6 Pts] Consider the following diagram of the bootstrap process. Fill in 9 blanks on the diagram using the phrases below:

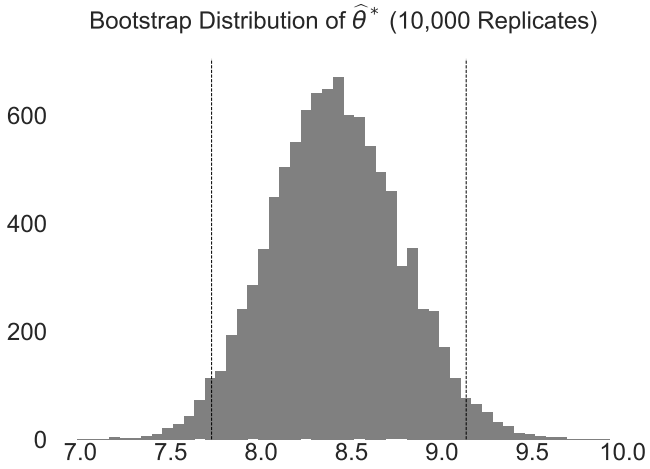
(A) Population
(B) Bootstrap population
(C) Observed sample
(D) Expected sample
(E) Bootstrap sample
(F) Sampling distribution
(G) Sampling
(H) Bootstrapping
(I) Bootstrap sampling distribution
(J) Empirical distribution
(K) True distribution
(L) Population parameter
(M) Sample Statistic
(N) Bootstrap Statistic



1. (A)
2. (L)
3. (G)
4. (C)
5. (M)
6. (H)
7. (E)
8. (N)
9. (I)

11. A fast food chain collects a sample of $n = 100$ service times from their restaurants, and finds a sample average of $\widehat{\theta} = 8.4$ minutes and a sample standard deviation of 2 minutes. They wish to construct a confidence interval for the population mean service time, denoted by θ.

(1) [2 Pts] The 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution for the mean $\widehat{\theta}^*$ below are located at 7.7 and 9.1, respectively. Which of the following constitutes a valid 95% **bootstrap confidence interval** for θ?



◯ $\left(8.4 - 1.96 \cdot \frac{2}{10}, 8.4 + 1.96 \cdot \frac{2}{10}\right)$
√ $(7.7, 9.1)$
◯ $\left(7.7 - 1.96 \cdot \frac{2}{10}, 9.1 + 1.96 \cdot \frac{2}{10}\right)$
◯ $(8.4 - 7.1, 8.4 + 9.1)$

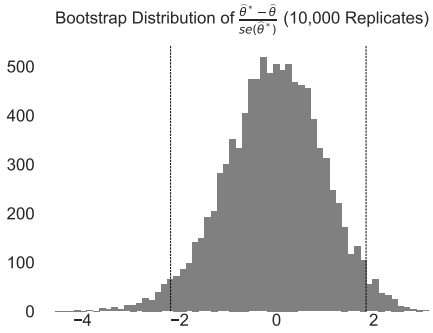
Explain your reasoning in the box below.

Solution: Use the percentiles of the bootstrap distribution directly: $(7.7, 9.1)$

(2) [4 Pts] The 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution for the *studentized mean*

$$\frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\mathbf{SE}}\left[\widehat{\theta}^*\right]}$$

(depicted below) are located at -2.2 and 1.9, respectively. Which of the following constitutes a valid 95% **bootstrap confidence interval** for θ? Recall that $\widehat{\theta} = 8.4$ and $\widehat{\mathbf{SE}}\left[\widehat{\theta}\right] = \frac{2}{\sqrt{100}}$.



√ $\left(8.4 - 1.9\frac{2}{\sqrt{100}}, 8.4 + 2.2\frac{2}{\sqrt{100}}\right)$ ○ $\left(8.4 - 2.2\frac{2}{\sqrt{100}}, 8.4 + 1.9\frac{2}{\sqrt{100}}\right)$

○ $(-1.9, 2.2)$ ○ $\left(7.7 - 2.2\frac{2}{\sqrt{100}}, 9.1 + 1.9\frac{2}{\sqrt{100}}\right)$

Explain your reasoning in the box below.

Solution: Let $q^*_{.025}$ and $q^*_{.975}$ denote the quantiles of the bootstrap distribution for the studentized mean. Then using $q^*_{.025}$ and $q^*_{.975}$ as estimates of the quantiles of $\frac{\widehat{\theta} - \theta}{\mathbf{SE}\left[\widehat{\theta}\right]}$, we have

$$\begin{aligned} 0.95 &\approx \mathbb{P}\left(q^*_{.025} \leq \frac{\widehat{\theta} - \theta}{\text{se}(\widehat{\theta})} \leq q^*_{.975}\right) \\ &= \mathbb{P}\left(\widehat{\theta} - q^*_{.975}\text{se}(\widehat{\theta}) \leq \theta \leq \widehat{\theta} - q^*_{.025}\text{se}(\widehat{\theta})\right) \end{aligned}$$

Hence the appropriate interval is

$$\left(\widehat{\theta} - q^*_{.975}\text{se}(\widehat{\theta}), \widehat{\theta} - q^*_{.025}\text{se}(\widehat{\theta})\right) = \left(8.4 - 1.9\frac{2}{\sqrt{100}}, 8.4 + 2.2\frac{2}{\sqrt{100}}\right) = (8.02, 8.84).$$

Map Reduce, Spark, and Big Data

12. [2 Pts] From the following list, select all the statements that are true:

☐ Schema on *read* means that the organization of data is determined when it is *loaded* into the data warehouse.

> **Solution:** In *schema on load* data is organized as it is loaded into the data warehouse. In *schema on read* data is organized as it is read during data analysis.

☐ In a star schema the primary keys are stored in the fact table and the foreign keys are stored in the dimension table.

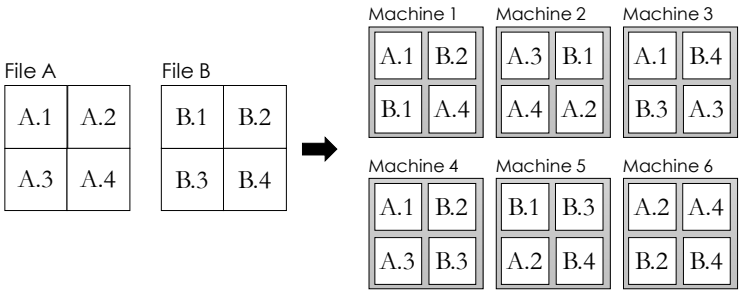
> **Solution:** In a start schema the fact table contains foreign key reference to each of the dimension tables.

√ **Data stored in a data lake will typically require more data cleaning than the data stored in the data warehouse.**

> **Solution: Data in the data warehouse is typically cleaned during the ETL process while data in the data lake is captured in its raw form and may required substantially data cleaning and transformation.**

☐ None of the above statements are true.

13. Consider the following layout of the files `A` and `B` onto a distributed file-system of 6 machines.



Assume that all blocks have the same file size and computation takes the same amount of time.

(1) [1 Pt] If we wanted to load file `A` in parallel which of the following sets of machines would give the best load performance:

◯ $\{M1, M2\}$ ◯ $\{M1, M2, M3\}$ √ $\{M2, M4, M5, M6\}$

Solution: While all choices would be able to load the file, only $\{M2, M4, M5, M6\}$ could load the file in parallel.

(2) [1 Pt] If we were to lose machines $M1$, $M2$, and $M3$ which of the following file or files would we lose (select all that apply).

☐ File `A` ☐ File `B` √ **We would still be able to load both files.**

(3) [1 Pt] If each of the six machines fail with probability p, what is the probability that we will lose block $B.1$ of file `B`.?

◯ $3p$ √ p^3 ◯ $(1-p)^3$ ◯ $1-p^3$

14. [4 Pts] Suppose you are given the following `raw.txt` containing the income for set of individuals:

```
State     Age     Income
VA          28      45000
CA          33      72000
VA          24      50000
CA          32      100000
TX          45      53000
ca          42      89000
ca          70      8000
TX          35      41000
TX          48      71000
VA          92      3000
```

What does the following query compute?

```
(sc.textFile("raw.txt")
   .map(lambda x: x.split())
   .filter(lambda x: x[0] != "State")
   .map(lambda x: [x[0].upper(), float(x[1]), float(x[2])])
   .filter(lambda x: x[1] <= 65.0)
   .map(lambda x: [x[0], np.array([1.0, x[2], x[2]**2])] )
   .reduceByKey(lambda a, b: a + b)
   .mapValues(lambda x: np.sqrt(x[2]/x[0] - (x[1]/x[0])**2))
).collect()
```

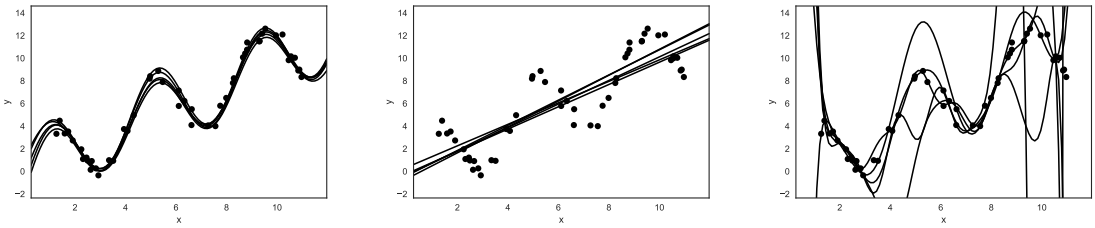
- ◯ The variance in income for each state.
- ◯ The standard deviation in income for each state.
- √ **The standard deviation of the income for each state excluding individuals who are older than 65.0**
- ◯ The standard deviation of the income excluding individuals who are older than 65.

15. [2 Pts] Select all of the following aggregation operations that will produce the same result regardless of the ordering of the data.

- √ `lambda a, b: max(a, b)`
- √ `lambda a, b: a + b`
- ◯ `lambda a, b: a - b`
- ◯ `lambda a, b: (a-b)**2`

Bias Variance Trade-off and Regularized Loss Minimization

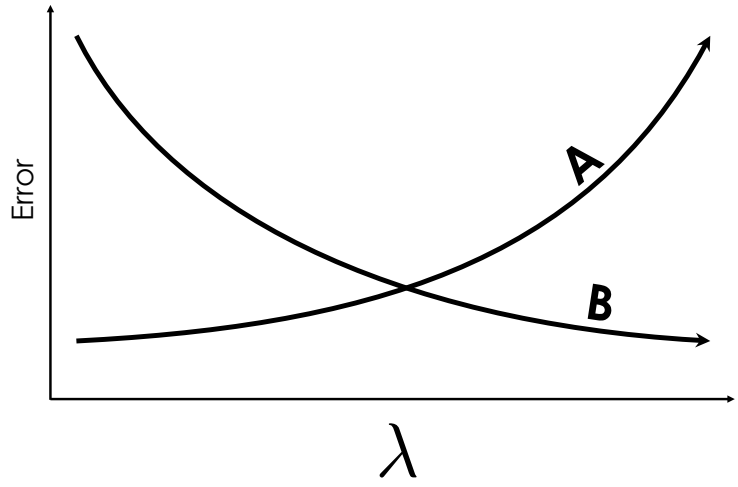
16. [1 Pt] Which of the following plots depicts models with the highest model variance?



(a) (b) (c)

◯ (a) ◯ (b) √ **(c)**

17. [3 Pts] Assuming a regularization penalty of the form $\lambda R(\theta)$. Complete the following illustration. Note that the x-axis is the regularization parameter λ and not the model complexity.



◯ (A) is the **Test Error** and (B) is the error due to **(Bias)**^{2}.

◯ (A) is the error due to **Model Variance** and (B) is the **Training Error**

◯ (A) is the error due to **Model Variance** and (B) is the error due to **(Bias)**^{2}.

√ **(A) is the error due to (Bias)**^{2} **and (B) is the error due to Model Variance.**

18. 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. You use f_θ to model the data where θ is the model parameter. You choose to use the following regularized loss:

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

(1) [1 Pt] This regularized loss is best described as:

- ◯ Average absolute loss with L^2 regularization.
- ◯ Average squared loss with L^1 regularization.
- √ **Average squared loss with L^2 regularization.**
- ◯ Average Huber loss with λ regularization.

(2) [6 Pts] Suppose you choose the model $f_\theta(x_i) = \theta x_i^3$. Using the above objective derive and circle the loss minimizing estimate for θ.

Solution:

Step 1: Take the derivative of the loss function.

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

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

Step 2: Set derivative equal to zero and solve for θ.

$$0 = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \theta x_i^3\right)x_i^3 + 2\lambda\theta \tag{4}$$

$$\theta = \frac{1}{n\lambda}\sum_{i=1}^{n}\left(y_i - \theta x_i^3\right)x_i^3 \tag{5}$$

$$\theta = \frac{1}{n\lambda}\sum_{i=1}^{n}y_i x_i^3 - \theta\frac{1}{n\lambda}\sum_{i=1}^{n}x_i^6 \tag{6}$$

$$\theta\left(1 + \frac{1}{n\lambda}\sum_{i=1}^{n}x_i^6\right) = \frac{1}{n\lambda}\sum_{i=1}^{n}y_i x_i^3 \tag{7}$$

(8)

Thus we obtain the final answer:

$$\boxed{\hat{\theta} = \frac{\frac{1}{n}\sum_{i=1}^{n}y_i x_i^3}{\left(\lambda + \frac{1}{n}\sum_{i=1}^{n}x_i^6\right)}} \tag{9}$$

Least Squares Regression

19. Given a full-rank $n \times p$ design matrix X, and the corresponding response vector $y \in \mathbb{R}^n$, the Least Squares estimator is

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Let $e = y - \hat{y}$ denote the $n \times 1$ vector of residuals, where $\hat{y} = X\hat{\beta}$. (Illustrated below)

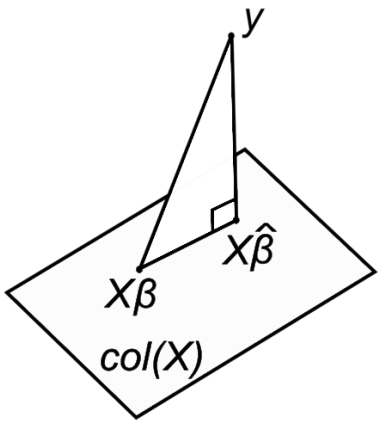


Figure 2: Geometric interpretation of Least Squares, courtesy of Wikipedia.

(1) [1 Pt] There exists a set of weights β such that $\sum_{i=1}^{n} (y_i - (X\beta)_i)^2 < \sum_{i=1}^{n} (e_i)^2$.

◯ True √ **False**

Solution: The least squares estimator $\hat{\beta}$ minimizes the RSS, so

$$\sum_{i=1}^{n} (y - X\hat{\beta})_i^2 \leq \sum_{i=1}^{n} (y - X\beta)_i^2$$

for every β.

(2) [1 Pt] We always have that $e \perp \hat{y}$ (i.e. $e^{\mathsf{T}}\hat{y} = 0$).

√ **True** ◯ False

(3) [1 Pt] For any set of weights β, we always have that $e \perp X(\hat{\beta} - \beta)$.

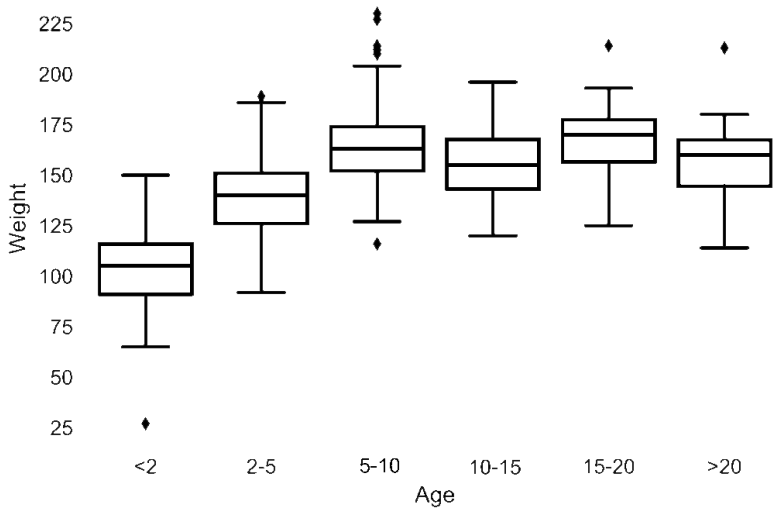
√ **True** ◯ False

Provide a short argument:

Solution:

- The residual is orthogonal to the columns space of X.
- It is true by the picture.

20. [2 Pts] When developing a model for a donkeys weight, we considered the following box plots of weight by age category.



This plot suggests:

◯ age is not needed in the model

√ **some of the age categories can be combined**

◯ age could be treated as a numeric variable

◯ none of the above

21. [8 Pts] Suppose that we try to predict a donkey's weight, y_i from its sex alone. (Recall that the sex variable has values: gelding, stallion, and female). In class, we studied the following model consisting of dummy variables:

$$y_i = \theta_F D_{F,i} + \theta_G D_{G,i} + \theta_S D_{S,i}$$

where the dummy variable $D_{F,i} = 1$ if the i^{th} donkey is female and $D_{F,i} = 0$ otherwise. The dummy variables D_G and D_S are dummies for geldings and stallions, respectively.

Prove that if we using the following loss function:

$$L(\theta_F, \theta_G, \theta_S) = \sum_{i=1}^{n} \left(y_i - (\theta_F D_{F,i} + \theta_G D_{G,i} + \theta_S D_{S,i})\right)^2$$

then the loss minimizing value $\hat{\theta}_F = \bar{y}_F$ where $\bar{y}_F$ is the average weight of the female donkeys.

Solution: The summation that we are minimizing can be split into three separate sums because only one of the dummy variables is 1 for any observation. That is, when $D_{F,i} = 1$ then $D_{G,i} = 0$ and $D_{S,i} = 0$.

$$\min_{\theta_F, \theta_G, \theta_S} \sum_{i=1}^{n} (y_i - (\theta_F D_{F,i} + \theta_G D_{G,i} + \theta_S D_{S,i}))^2$$
$$= \sum_{F} (y_i - \theta_F)^2 + \sum_{G} (y_i - \theta_G)^2 + \sum_{S} (y_i - \theta_S)^2$$

This implies that we can minimize over θ_F separately, i.e.,

$$\min_{\theta_F} \sum_{F} (y_i - \theta_F)^2$$

We can differentiate with respect to θ_F to get

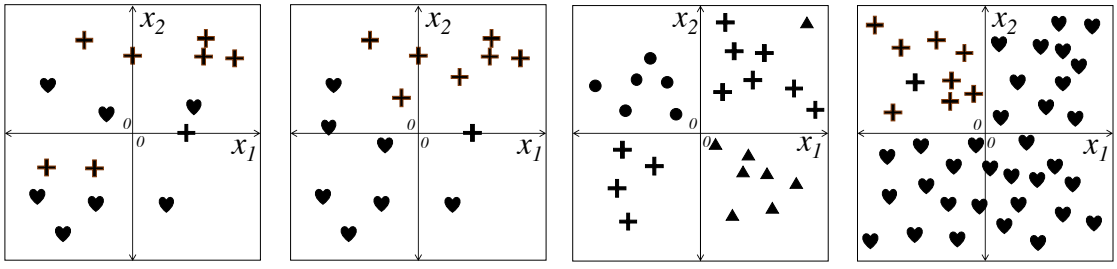
$$\sum_{F} -2(y_i - \theta_F)$$

Set this to 0 and solve for θ_F

$$\frac{1}{\#F} \sum_{F} y_i = \hat{\theta}_F$$

Classification and Logistic Regression

22. Consider the following figures of different shapes plotted in a two dimensional feature space. Suppose we are interested in classifying the type of shape based on the location.



(A) (B) (C) (D)

(1) [1 Pt] Which figure best illustrates substantial class imbalance?

◯ (A) ◯ (B) ◯ (C) √ **(D)**

(2) [1 Pt] Which figure is linearly separable.

◯ (A) √ **(B)** ◯ (C) ◯ (D)

(3) [1 Pt] Which figure corresponds to a multi-class classification problem.

◯ (A) ◯ (B) √ **(C)** ◯ (D)

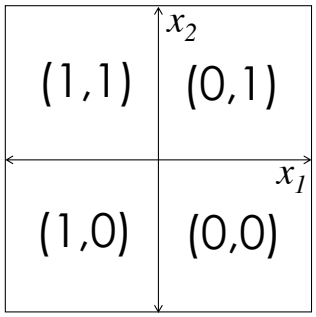
(4) [3 Pts] Assuming we applied the following feature transformation:

$$\phi(x) = [\ \mathbb{I}(x_1 < 0),\ \ \mathbb{I}(x_2 > 0),\ \ 1.0\ ]$$

where $\mathbb{I}(z)$ is the indicator which is 1.0 if the expression z is true and 0 otherwise. Which of the above plots is linearly separable in the transformed space (select all that apply).

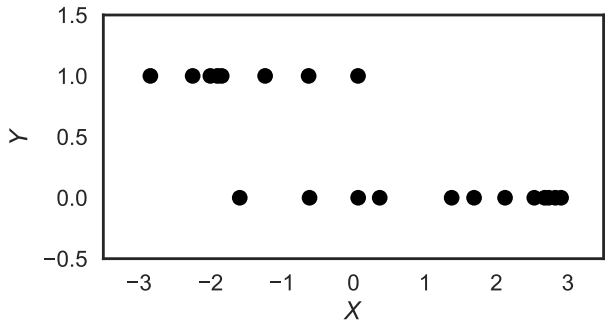
◯ (A) ◯ (B) ◯ (C) √ **(D)** ◯ None of the plots.

Solution: This question is a bit tricky. The feature transformation maps each quadrant to the feature values in the following picture (bias term not included):



We see that in this case (D) is clearly linearlly separable. While (C) is almost linearly separable there is a triangle in the 1st quadrant that would not be separable from the crosses.

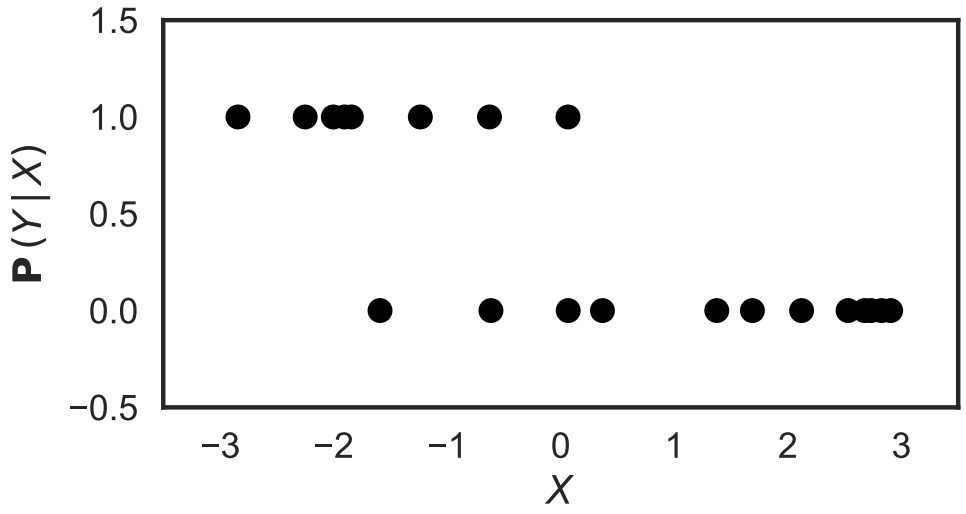
23. Suppose you are given the following dataset $\{(x_i, y_i)\}_{i=1}^n$ consisting of x and y pairs where the covariate $x_i \in \mathbb{R}$ and the response $y_i \in \{0, 1\}$.



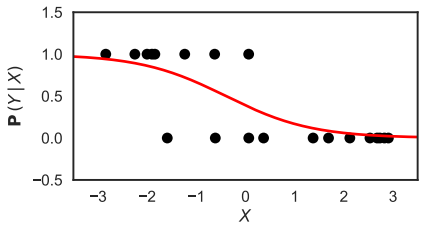
(1) [1 Pt] Given this data, the value $\mathbf{P}(Y = 1 \mid x = 3)$ is likely closest to:

◯ 0.95 ◯ 0.50 √ **0.05** ◯ -0.95

(2) [2 Pts] Roughly sketch the predictions made by the logistic regression model for $\mathbf{P}(Y = 1 \mid X)$.



Solution:



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

```
1  def stochastic_grad_descent(
2              X, Y, theta0, grad_function,
3              max_iter = 1000000, batch_size=100):
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     returns the optimal theta
12     """
13     theta = theta0
14     for t in range(1, max_iter+1):
15
16         (xbatch, ybatch) = (X[1:batch_size, :], Y[1:batch_size])
17
18         grad = grad_function(theta0, xbatch, ybatch)
19
20         theta = theta - t * grad
21
22     return theta
```

(1) [4 Pts] Select all the issues with this Python implementation

- √ **Line 16 does not adequately sample all the data.**
- √ **Line 18 should be evaluated at `theta` and not `theta0`.**
- ☐ **Line 18** should take the negative of the gradient.
- ☐ **Line 20** should be evaluated at `theta0` and not `theta`.
- √ **Line 20, `t` should be replaced with `1/t`.**

(2) [2 Pts] Supposed we wanted to add L^2 regularization with parameter `lam`. Which of the following rewrites of **Line 18** would achieve this goal:

- ◯ `grad = (grad_function(theta, xbatch, ybatch) + theta.dot(theta) * lam)`
- ◯ `grad = (grad_function(theta, xbatch, ybatch) - theta.dot(theta) * lam)`
- √ **`grad = (grad_function(theta, xbatch, ybatch) + 2*theta*lam)`**
- ◯ `grad = (grad_function(theta, xbatch, ybatch) - 2*theta*lam)`

P-Hacking

25. [2 Pts] An analysis of tweets the day after hurricane Sandy reported a *surprising* finding – that nightlife picked up the day after the storm. It was supposed that after several days of being stuck at home cabin fever struck. However, later someone pointed out that most tweets were from Manhattan and that those tweeting were not suffering from an extended black out. The earlier study's conclusions are an example of:

    ◯ Texas sharpshooter bias

    √ **sampling bias**

    ◯ confirmation bias

    ◯ Simpson's paradox

26. [2 Pts] Suppose that everyone of the 275 students in Data 100 is administered a clairvoyance test as part of the final exam and two of the students "pass" the test and are declared to be clairvoyant. What kind of mistake have the professors in Data 100 have made in their testing:

    ◯ post-hoc ergo procter-hoc

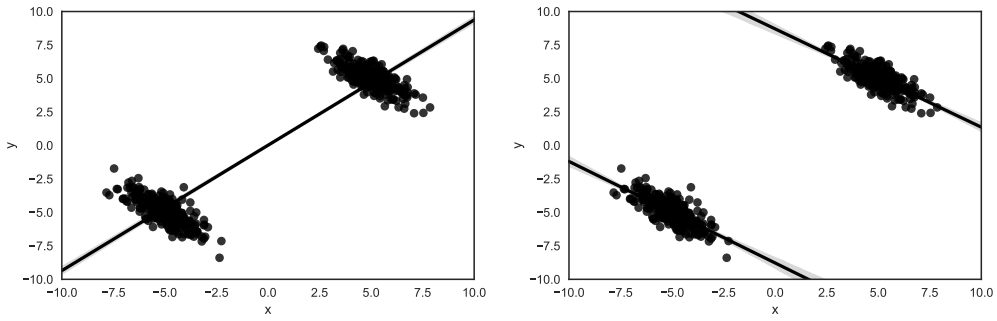
    ◯ gambler's fallacy

    ◯ early stopping

    √ **multiple testing**

    ◯ Simpson's paradox

27. [2 Pts] The following plot illustrates a reversal in trends observed when conditioning a model on subgroups.



    This is an example of:

    ◯ post-hoc ergo procter-hoc

    ◯ sampling bias

    ◯ selection bias

    √ **Simpson's paradox**

Feature Engineering, Over-fitting, and Cross Validation

28. [2 Pts] Select **all** statements that are true.

- ☐ If there are two identical features in the data, the L^2-regularization will force the coefficient of one redundant feature to be 0.
- ☐ We **cannot** use linear regression to find the coefficients for θ in $y = \theta_1 x^3 + \theta_2 x^2 + \theta_3 x + \theta_4$ since the relationship between y and x is non-linear.
- √ **Introducing more features increases the model complexity and may cause over-fitting.**
- ☐ None of the above statements are true.

29. [2 Pts] Bag-of-words encodings have the disadvantage that they drop semantic information associated with word ordering. Which of the following techniques is able to retain some of the semantic information in the word ordering? Select **all** that apply.

- ☐ Remove all the stop words
- √ **Use N-gram features.**
- ☐ Give more weights if one word occurs multiple times in the document. (Similar to the TF-IDF)
- √ **Create special features for common expressions or short phrases.**
- ☐ None of the above.

30. Suppose you are fitting a model parameterized by θ using a regularized loss with regularization parameter λ. Indicate which error you should use to complete each of the following tasks.

(1) [1 Pt] To optimize θ you should use the:

√ **Training Error** ◯ Cross-Validation Error ◯ Test Error

(2) [1 Pt] To determine the best value for λ you should use the:

◯ Training Error √ **Cross-Validation Error** ◯ Test Error

(3) [1 Pt] To evaluate the degree of polynomial features you should use the:

◯ Training Error √ **Cross-Validation Error** ◯ Test Error

(4) [1 Pt] To evaluate the quality of your final model you should use the:

◯ Training Error ◯ Cross-Validation Error √ **Test Error**

End of Exam