DSC 40A Final Exam

March 16, 2019

- This exam has 14 questions and 64 points. You have 3 hours.
- No books, notes, cell phones, computers.
- Only the boxed areas will be graded.
- Just provide the answer. You do not need to justify answers or show work unless otherwise instructed.
- You can leave all answers unsimplified unless otherwise instructed. You are welcome to use notation such as $C(6, 3)$ and $P(6, 3)$ in your answers.
- **Do not start** until you are instructed to do so.
- Write your name and PID on every page.

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam.

Signature: _____

Name (Printed): _____

PID: _____

Version 1

Name:_____ PID: _____

1. (4 points) For a data set $y_1 \leq y_2 \leq \cdots \leq y_n$, define the loss function

$$L_2(h) = \sum_{i=1}^{n} (h - y_i)^2.$$

Give an example of a data set of size $n = 5$ with $0 \leq y_1 \leq y_2 \leq y_3 \leq y_4 \leq y_5 \leq 10$ for which, at any value of $h < y_4$, $L_2(h)$ is decreasing.

2. (4 points) Give an example of a data set of size $n = 4$ where the mean equals the median, and the mean absolute deviation from the median is greater than the variance.

Name:_____ PID: _____

3. (6 points) Suppose you have a data set $y_1, y_2, \ldots, y_n$ for which

   - $n = 26$
   - minimum $= 3$
   - mode $= 7$
   - midrange $= 19$
   - median $= 13$
   - mean $= 10$

   Then we add a data value of 37 to this data set to create a new data set. Answer the questions below for the new data set. Give an **exact simplified answer** or circle "cannot be determined."

   (a) mode = [ ] or Cannot Be Determined

   (b) midrange = [ ] or Cannot Be Determined

   (c) median = [ ] or Cannot Be Determined

   (d) mean = [ ] or Cannot Be Determined

   (e) maximum = [ ] or Cannot Be Determined

   (f) second largest value = [ ] or Cannot Be Determined

Name:_____ PID: _____

4. (6 points) Suppose we fit a line to the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to minimize mean square error, and this fitted line has intercept b_0 and slope b_1. Consider what happens when move the points up or down and re-fit a line to minimize mean square error. Let $z_i = y_i + \delta_i$ for $i = 1, 2, \ldots, n$. Let c_0 and c_1 be the intercept and slope of the line fit to the points $(x_1, z_1), (x_2, z_2), \ldots, (x_n, z_n)$. Suppose that $n = 5$ and

$$x_1 = 1, \quad x_2 = 3, \quad x_3 = 7, \quad x_4 = 9, \quad x_5 = 10.$$

(a) If $\delta_1 = 2$, give values of $\delta_2, \delta_3, \delta_4, \delta_5$ so that $c_1 = b_1$. In other words, determine how to move the other points to keep the slope unchanged.

$\delta_2 =$ ☐ $\delta_3 =$ ☐ $\delta_4 =$ ☐ $\delta_5 =$ ☐

(b) Answer the same question as part (a), but with a different set of values $\delta_2, \delta_3, \delta_4, \delta_5$.

$\delta_2 =$ ☐ $\delta_3 =$ ☐ $\delta_4 =$ ☐ $\delta_5 =$ ☐

(c) If $\delta_4 = -6$, give values of $\delta_1, \delta_2, \delta_3, \delta_5$ so that $c_1 = b_1$ and exactly one of $\delta_1, \delta_2, \delta_3, \delta_5$ is nonzero.

$\delta_1 =$ ☐ $\delta_2 =$ ☐ $\delta_3 =$ ☐ $\delta_5 =$ ☐

Make sure that three of the four blanks in part (c) are filled with zero.

Name:_____ PID: _____

5. (4 points) Consider a data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of n points which **do not all fall on the same line**. Let

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Let $\vec{b}$ be the vector that satisfies $X^T X\vec{b} = X^T\vec{y}$. For each statement below, fill in the circle if the statement must be true.

◯ Since $\vec{b}$ satisfies $X^T X\vec{b} = X^T\vec{y}$, $\vec{b}$ also satisfies $X\vec{b} = \vec{y}$.

◯ $\vec{b}$ is the orthogonal projection of $\vec{y}$ onto $\text{Col}(X)$, the column space of X.

◯ $X^T\vec{y}$ is in $\text{Col}(X)$.

◯ $\vec{y}$ is in $\text{Col}(X)$.

◯ $X\vec{b} - \vec{y}$ is in $\text{Col}(X)$.

◯ $X\vec{b} - \vec{y}$ is in $\text{Nul}(X^T)$, the nullspace of X^T.

◯ $\vec{y} \perp \text{Col}(X)$.

◯ $X\vec{b} - \vec{y} \perp \text{Col}(X)$.

Name:_____ PID: _____

6. (5 points) For each statement below, fill in the circle if the statement is true.

   ◯ $\log_2 n^2 \in O(\log_2 n)$

   ◯ $(3n)^6 \in \Theta(3^6)$

   ◯ $\sqrt{4n^4 + 2n} \in \Omega(n)$

   ◯ If f and g are functions from the natural numbers to the non-negative real numbers with $f(n) \in \Omega(g(n))$, then $f(n) + n \in \Omega(g(n))$.

   ◯ If f and g are functions from the natural numbers to the non-negative real numbers with $f(n) \in O(g(n))$, then $g(n) \in \Theta(f(n))$.

7. (2 points) Any function that can be written in the form $f(n) = c_1 n + c_0$ is in $\Theta(n)$. For example $3n + 7$ and $14n - 2$ are in $\Theta(n)$. Give an example of a function $f(n)$ that **cannot** be written in the form $c_1 n + c_0$ such that $f(n)$ is in $\Theta(n)$.

   $f(n) =$ [ ]

Name:_____ PID: _____

8. (6 points) You want to plant an herb garden, so you go to a garden store that has 50 different herbs: 20 are culinary herbs, 15 are medicinal herbs, and 15 are aromatic herbs. You select 5 herbs for your herb garden by taking a random sample **without replacement** from the 50 available herbs.

   (a) If you consider the herbs you select as a sequence where the order in which you select each herb matters, how many sequences of 5 herbs are possible?

   (b) If you consider the herbs you select as a sequence where the order in which you select each herb matters, how many sequences of 5 herbs include 2 culinary herbs and 3 aromatic herbs?

   (c) If you consider the herbs you select as a set where the order in which you select each herb does not matter, how many sets of 5 herbs are possible?

   (d) If you consider the herbs you select as a set where the order in which you select each herb does not matter, how many sets of 5 herbs include 2 culinary herbs and 3 aromatic herbs?

(e) Which expression equals the probability that you choose 2 culinary herbs and 3 aromatic herbs for your herb garden? Fill in one circle for this question.

- ◯ $\frac{\text{your answer to part (b)}}{\text{your answer to part (a)}}$
- ◯ $\frac{\text{your answer to part (d)}}{\text{your answer to part (c)}}$
- ◯ both of the above
- ◯ none of the above

(f) Suppose the first two herbs you select are aromatic herbs. Based on this, what is the probability that you wind up with 2 culinary herbs and 3 aromatic herbs?

Name:_____ PID: _____

9. (4 points) Every day for lunch, you randomly decide whether to eat a salad, a small sandwich, or a gigantic sandwich. You are equally likely to choose any of these options. On $\frac{3}{4}$ of the days where you eat salad, you are hungry in the afternoon. On $\frac{1}{2}$ of the days where you eat a small sandwich, you are hungry in the afternoon. On $\frac{1}{4}$ of the days where you eat a gigantic sandwich, you are hungry in the afternoon.

   You find yourself hungry one afternoon. Based on this information, what is the probability that you ate a gigantic sandwich for lunch? **Show your work.**

Name:_____ PID: _____

10. (6 points) At an animal shelter, there are 18 crates lined up in a single row. Each animal that arrives to the shelter has a $\frac{1}{3}$ probability of being a dog, and a $\frac{2}{3}$ probability of being a cat. The crates are originally empty, and as each animal arrives, it is placed in the next available crate from left to right, until there are 18 animals in the shelter.

    (a) What is the probability that there is at least one dog in the animal shelter?

    (b) What is the expected number of dogs in the animal shelter?

    (c) What is the expected number of times a dog and cat have adjacent crates? (For example, if there were only 5 crates: Cat, Dog, Cat, Cat, Dog would be 3.)

    (d) What is the expected number of dogs that are adjacent to only cats? (For example, if there were only 5 crates: Cat, Dog, Cat, Cat, Dog would be 2.)

Name:_____ PID: _____

11. (4 points) Consider the following training data set about customer reviews of a pineapple slicer. For each customer review, the number of exclamation points and periods in the review were counted and compared to see which was more common. The number of times the word *"not"* appeared in the review was counted. Finally, each review was determined to be positive or negative in its opinion about the pineapple slicer.

more common punctuation	*"not"* word count	opinion
period	1	negative
period	1	negative
exclamation point	2	negative
period	0	negative
exclamation point	1	negative
period	2	negative
period	1	negative
period	0	negative
exclamation point	1	positive
period	0	positive
period	1	positive
exclamation point	0	positive
exclamation point	2	positive

Use the data above and the naive Bayes algorithm to predict whether the following review is negative or positive. **Show your work.**

"This thing is **not** *really a necessity, but it does its job well. I cut a pineapple in a fraction of the time I used to spend cutting a pineapple. I'll probably eat more pineapple now that it's* **not** *such a pain to cut. Pineapple is pretty tasty."*

Prediction according to naive Bayes: | negative | positive |

Show your work leading to this prediction on the next page.

Name:_____ PID: _____

Work for naive Bayes question:

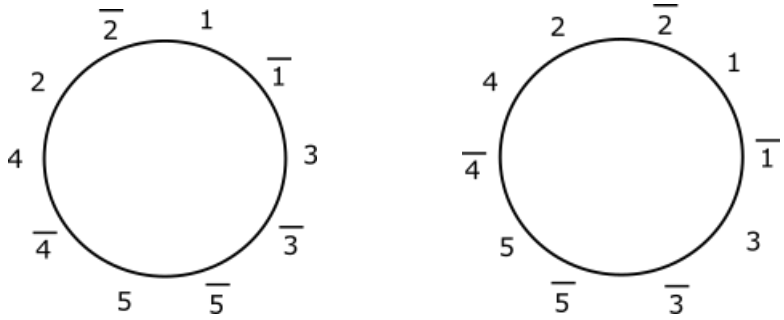
Name:_____ PID: _____

12. (4 points) At a party, there are 10 people who arrive in 5 couples. We will label the people

$$1, \overline{1}, 2, \overline{2}, 3, \overline{3}, 4, \overline{4}, 5, \overline{5}$$

where 1 and $\overline{1}$ are a couple, 2 and $\overline{2}$ are a couple, and so on. All 10 people will be seated at a single round table.

When we count seating arrangments, we will consider two seating arrangements to be the same only when each person is sitting in the exact same chair in both seating arrangements. In particular, when we rotate the seats around the table, this is considered a different seating arrangement. For example, the two seating arrangements shown below are considered different and should be counted separately.



(a) How many seating arrangements are possible?

(b) **Extra Credit (+2):** How many seating arrangements have each person sitting next to the person they came with?

(c) In a randomly chosen seating arrangement, what is the probability that person 1 is next to person $\overline{1}$?

(d) In a randomly chosen seating arrangement, what is the expected number of people who are sitting next to the person they came with?

13. (4 points) A population consists of n people. Two samples of size k are drawn from this population, each **without replacement**.

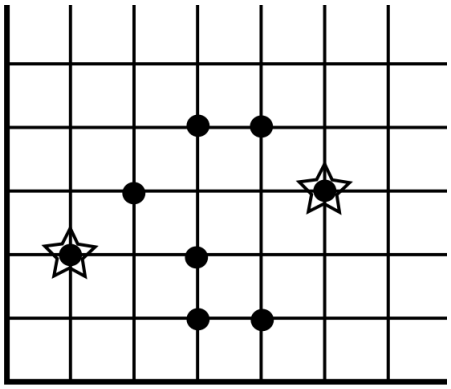
(a) What is the probability that the two samples contain the same set of individuals (possibly in a different order)?

(b) What is the probability that the two samples contain the same individuals in the same order?

Name:_____ PID: _____

14. (5 points) We are given the following data and want to find $k = 2$ clusters.

$$x^{(1)} = (3, 4), \quad x^{(2)} = (4, 4), \quad x^{(3)} = (2, 3), \quad x^{(4)} = (5, 3)$$
$$x^{(5)} = (1, 2), \quad x^{(6)} = (3, 2), \quad x^{(7)} = (3, 1), \quad x^{(8)} = (4, 1)$$



Suppose we randomly select $x^{(4)}$ and $x^{(5)}$ as cluster centers (starred above).

(a) What is the initial cost determined by these randomly chosen cluster centers? Give a **fully simplified** answer.

Cost = $\boxed{}$

(b) After one iteration of Lloyd's algorithm for k-means clustering (one time of moving the centroids), where will the centroids be located?

$\boxed{\left(, \right) \text{ and } \left(, \right)}$

(c) On the graph below, circle each of the clusters determined by this first iteration of the algorithm.