

---

## DSC 40A - Homework 5

Due: Tuesday, May 16 at 11:59pm

---

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59pm on the due date. You can use a slip day to extend the deadline by 24 hours.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it.


For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

### Notes:

- For full credit, make sure to **assign pages to questions** when you upload your submission to Gradescope.
- For Problem 1, parts (b) and (c), code your answers in the [supplementary Jupyter notebook \(linked\)](#). You'll need to turn in your completed Python file to Gradescope separately from the rest of this homework, in a file called `hw5code.py`. Parts (b) and (c) of problem 1 will be autograded, so no explanation is needed.

### Problem 0. Reflection and Feedback Form



 Make sure to fill out this [Reflection and Feedback Form, linked here](#) for two points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

### Problem 1. k-Means Clustering


For parts (a) and (b) of this question, we'll use the five data points given below,  $\vec{x}_1$  through  $\vec{x}_5$ .

$$\vec{x}_1 = \begin{bmatrix} 7 \\ 60 \end{bmatrix}, \vec{x}_2 = \begin{bmatrix} 31 \\ 17 \end{bmatrix}, \vec{x}_3 = \begin{bmatrix} 1 \\ 64 \end{bmatrix}, \vec{x}_4 = \begin{bmatrix} 35 \\ 15 \end{bmatrix}, \vec{x}_5 = \begin{bmatrix} 4 \\ 71 \end{bmatrix}$$


Just by looking at the data, you should be able to roughly identify two clusters. Let's see how  $k$ -means clustering finds these clusters algorithmically.

-  Using  $\vec{x}_1$  and  $\vec{x}_2$  as initial centroids, trace through one iteration of the  $k$ -means algorithm by hand. What are the two centroids and what are the two clusters found after this first iteration?
-  In the [supplementary Jupyter notebook \(linked\)](#), implement a Python function that takes in a cluster and a centroid, and returns the cost of the cluster associated with that centroid.




Then, use the function you've written to compute the value of the cost function before the first iteration, as well as after the first iteration, for the five data points we used in part (a). You should see that the cost function has decreased with this first iteration.

- c)  For this part, you will implement the code for k-means clustering on a larger dataset. Follow the prompts in the [supplementary Jupyter notebook \(linked\)](#) to implement three functions: `initialize_centroids`, `find_closest_centroid`, and `initialize_centroids` and `k_means`.

## Problem 2. Avi's Lottery

Our adorable mascot Avi  has decided to launch a lottery among all 141 students taking DSC 40A this quarter. Each student will be randomly assigned a lottery ticket numbered  $1, 2, \dots, 141$ . Avi will then randomly generate a winning number, and the student with that same number on their lottery ticket will win their very own avocado plush toy.

Avi announces that the winning number is 46!

- a)  If you only look at the first (leftmost) digit of your lottery number and see that it's a 4, what is the probability that you've won the lottery?
- b)  If you glance at your lottery number and see that it contains a 4 somewhere, what is the probability that you've won the lottery?
- c)  If you glance at your lottery number and see that it contains exactly one 4, what is the probability that you've won the lottery?

## Problem 3. Probability Rules for Three Events

- a)  The multiplication rule for two events says

$$P(A \cap B) = P(A) \cdot P(B|A)$$

Use the multiplication rule for two events to prove the multiplication rule for three events:


$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|(A \cap B))$$

**Hint:** You can think of  $A \cap B \cap C$  as  $(A \cap B) \cap C$ .

- b)  Suppose  $E$ ,  $F$ , and  $G$  are events. Explain in words why

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G).$$

Intuitively, the relationship between  $\cap$  and  $\cup$  is similar to the relationship between multiplication and addition; if  $e, f, g$  are numbers, then  $(e + f) \cdot g = e \cdot g + f \cdot g$  as well.


- c)  The general addition rule for any two events says:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Use the general addition rule for two events to prove the general addition rule for three events:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

**Hint:** You will need to use the result of part (b).

- d)  A survey was administered to 500 Formula One (F1) Racing fans asking about their predictions for the 2023 F1 season. Each respondent named 3 drivers that they predicted would finish in the top 3. The survey revealed the following information:







- 20 respondents' predictions did not include any of Max Verstappen, Charles Leclerc, and Sergio Perez.
- 350 responses included Max Verstappen.
- Of the 350 respondents who said Max Verstappen, 240 also said Sergio Perez.
- Of the 350 respondents who said Max Verstappen, 150 also said Charles Leclerc
- 300 respondents said Sergio Perez.
- Of the 300 respondents who said Sergio Perez, 140 also said Charles Leclerc.
- 90 respondents predicted all three of Max Verstappen, Charles Leclerc, and Sergio Perez.

Suppose we randomly select one survey participant. What is the probability that they predicted that Charles Leclerc would be among the top 3 this year?

#### Problem 4. Stringle

In this problem, we will look at a made-up game called Stringle. Each day, a random six-letter string is chosen, and players have to try to guess what it is.




In Stringle, any six-letter string of uppercase letters is allowed, as long as it does not have any repeated letters. The string does not have to make sense as an English word. For example, the string of the day might be ZVODUP. Any valid string is equally likely to be chosen each day.

-  Consider A, E, I, O, U, and Y to be vowels. What is the probability that today's Stringle string and yesterday's Stringle string both start with a vowel?
-  What is the probability that today's Stringle string or yesterday's Stringle string starts with a vowel?
-  What is the probability that today's Stringle string includes no vowels?
-  What is the probability that today's Stringle string includes all vowels?
-  What is the probability that today's Stringle string includes the letter J?
-  What is the probability that today's Stringle string is exactly the same as yesterday's Stringle string?

#### Problem 5. Avi's Bootstraps

Recall from DSC 10 the process of bootstrap resampling. From a population of size  $n$ , we draw one random sample of size  $k$ , without replacement. Then, we create many bootstrap resamples by sampling  $k$  elements from the original sample, with replacement.

Suppose we have a population of 100 stuffed toys, one of which is Avi. From this population, we draw a sample of size 20, without replacement. From this original sample, we create 5 different bootstrap resamples.

-  What is the probability that Avi is included in the original sample?
-  What is the probability that Avi is included in the first resample?  
**Hint:** Don't make any assumptions about whether Avi was included in the original sample.
-  What is the probability that Avi is included in some resample?