DSC 40A - Homework 5

Due: Wednesday, Feb 21 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.

Problem 1. 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 2. Lloyd's Algorithm converges to a local optimum

Consider the following one dimensional dataset:

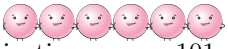
$$D = [-101, -99, 0, 29, 31]$$

We would like to cluster these points into k = 3 groups. The cost function we use will be squared distance:

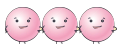
$$L(D, \mu) = \sum_i (D_i - \mu_{j^*})^2$$

where $j^* = \arg\min_j (\mu_j - D_i)^2$ is the closest center to D_i

a)  What is the optimal k-means solution by your observation? Give the locations of the centers as well as the k-means cost.

b)  Suppose we call Lloyd's k-means algorithm on this data, with $k = 3$ and with initialization $\mu_1 = -101, \mu_2 = -99, \mu_3 = 0$. By performing a few iterations by hand, show the final set of cluster centers and the final cost obtained by the algorithm. Compare this cost with the optimal cost you obtained in (a) and explain.

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. Texas Hold'em Poker and Combinatorics

In No Limit Texas Hold'em, a popular poker game, each player is dealt 2 cards from a deck of 52 (without Jokers), and there are 5 communal cards for all players. Your friend Tom Dwan is curious about the mathematics, specifically combinatorics, behind the game. He's asked for your help to calculate some probabilities to guide his strategy.

a) How many different combinations of 2-card starting hands can be dealt from a 52-card deck?

b) A pocket pair is when both cards have the same face value (for example $A\clubsuit$ and $A\spadesuit$). How many different pocket pair combinations are there? Suppose Tom decides to only play pocket pairs, what is the probability that Tom is dealt a pocket pair?

c) Suited cards are two cards of the same suit (for example $K\clubsuit$ and $9\clubsuit$). How many different combinations of suited cards are there? What is the probability that Tom is dealt suited cards?

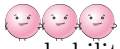
d)  If Tom decides to play hands that are either pocket pairs or suited cards, how many combinations does this strategy cover? What percentage of all possible hands does this strategy include?

Problem 5. 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 $1, 2, \ldots, 141$. 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 6. 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.

a)  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?

b)  What is the probability that today's Stringle string or yesterday's Stringle string starts with a vowel?

c)  What is the probability that today's Stringle string includes no vowels?

d)  What is the probability that today's Stringle string includes all vowels?

e)  What is the probability that today's Stringle string includes the letter J?

f)  What is the probability that today's Stringle string is exactly the same as yesterday's Stringle string?