
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.

Solution: The optimal three centers are $-100, 0, 30$. The corresponding optimal cost is: $1^2 + 1^2 + 0 + 1^2 + 1^2 = 4$

- 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.

Solution:

- First iteration: centers: $[\mu_1, \mu_2, \mu_3] = [-101, -99, 0]$, assignments: $[\mu_1, \mu_2, \mu_3, \mu_3, \mu_3]$, update centers: $[-101, -99, (0 + 29 + 31)/3] = [-101, -99, 20]$
- Second iteration: centers: $[-101, -99, 20]$, assignments: $[\mu_1, \mu_2, \mu_3, \mu_3, \mu_3]$, algorithm stops

The final cost is $0^2 + 0^2 + 20^2 + 9^2 + 11^2 = 602 \gg 4$. Lloyd's k-means algorithm is initialization-sensitive, there is no guarantee that the algorithm converges to the global optimum.

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 $(A \cap B) \cap C$.

Solution: We use the multiplication rule for two events twice, along with the hint:

$$\begin{aligned} P(A \cap B \cap C) &= P((A \cap B) \cap C) \\ &= P(A \cap B)P(C|(A \cap B)) \\ &= P(A)P(B|A)P(C|(A \cap B)) \end{aligned}$$

- 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.

Solution: Remember that events are subsets of the sample space. For an outcome to be in $(E \cup F) \cap G$, this means it's either in E or F , plus it is also in G . Put another way, that outcome is either (1) in E and G , or (2) in F and G . That's the same as saying it's in $(E \cap G) \cup (F \cap G)$.

- 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:

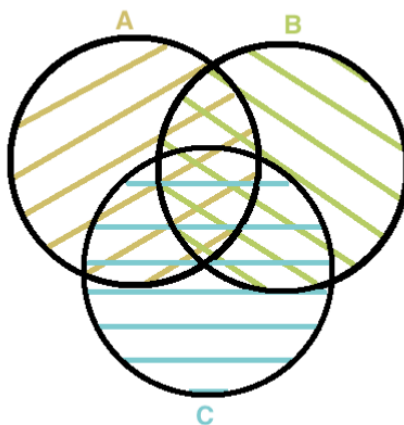
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

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

Solution: Similar to (a), we obtain that

$$\begin{aligned}
 P(A \cup B \cup C) &= P((A \cup B) \cup C) \\
 &= P(A \cup B) + P(C) - P((A \cup B) \cap C) \\
 &= P(A \cup B) + P(C) - P((A \cap C) \cup (B \cap C)) \\
 &= P(A \cup B) + P(C) - P(A \cap C) - P(B \cap C) + P((A \cap C) \cap (B \cap C)) \\
 &= P(A) + P(B) - P(A \cap B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \\
 &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)
 \end{aligned}$$

An intuitive explanation of why this property holds true:



When we add up the probabilities of A , B , and C separately, we count the outcomes in the regions of overlap multiple times. Those in the intersection of two sets get counted two times, and we can correct this by subtracting each intersection of two sets. In the very middle of the Venn diagram is the intersection of all three sets. Outcomes here get added three times, once for each of A , B , C , then get subtracted three times, once for each of $A \cap B$, $A \cap C$, $B \cap C$. So we need to add back in the intersection of all three sets one more time, which is accomplished by the last term in the general addition rule for three events.

There's a helpful visualization of this rule [at this link](#).

- 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?

Solution: $\frac{27}{50}$.

Let MV , CL , and SP be the events that a randomly selected survey participant predicted Max Verstappen, Charles Leclerc and Sergio Perez, respectively.

We need to find $P(CL)$. The information we are given is:

- $P(MV \cup CL \cup SP) = 1 - \frac{20}{500} = \frac{480}{500}$
- $P(MV) = \frac{350}{500}$
- $P(MV \cap SP) = \frac{240}{500}$
- $P(MV \cap CL) = \frac{150}{500}$
- $P(SP) = \frac{300}{500}$
- $P(SP \cap CL) = \frac{140}{500}$
- $P(MV \cap CL \cap SP) = \frac{90}{500}$

Then, by the general addition rule you proved in part (c), we have

$$\begin{aligned}
 P(MV \cup CL \cup SP) &= P(MV) + P(CL) + P(SP) - P(MV \cap SP) \\
 &\quad - P(MV \cap CL) - P(SP \cap CL) + P(MV \cap CL \cap SP) \\
 \frac{480}{500} &= \frac{350}{500} + P(CL) + \frac{300}{500} - \frac{240}{500} - \frac{150}{500} - \frac{140}{500} + \frac{90}{500} \\
 \implies P(CL) &= \frac{270}{500} = \frac{27}{50}
 \end{aligned}$$

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?

Solution: The number of possible combinations of two cards from a deck of 52 can be calculated using the formula for combinations:

$$\text{Combinations} = \frac{n!}{r!(n-r)!}$$

where

- n is the total number of items (52 cards in the deck), and
- r is the number of items to choose (2 cards for the hand)

Using this formula, the number of possible starting hands in Texas Hold'em is:

$$\begin{aligned}
 \text{Combinations} &= \frac{52!}{2!(52-2)!} \\
 &= \frac{52!}{2! \times 50!} \\
 &= 1326
 \end{aligned}$$

Thus, the total number of possible starting hands in Texas Hold'em poker is found to be 1,326.

- 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?

Solution: There are 13 different face values in poker, 1, 2, 3, ..., J, Q, K . For each face value, there are 4 cards of different suits. The total number of combinations is the number of combinations to get a pair of an arbitrary face value times 13. Within each suit, we're choosing 2 cards from 4 total possible candidates. Using this idea, the number of combinations of pocket pairs is:

$$\begin{aligned}\text{Combinations} &= 13 \times \binom{4}{2} \\ &= 13 \times \frac{4!}{2!(4-2)!} \\ &= 78\end{aligned}$$

Thus, the total number of pocket pairs is 78.

To calculate the probability of being dealt a pocket pair, we simply use divide the number of pocket pairs by the total number of possible starting hands, which gives us:

$$\begin{aligned}\text{Probability} &= \frac{78}{1326} \\ &= \frac{1}{17}\end{aligned}$$

Thus, the probability of being dealt a pocket pair is $\frac{1}{17}$.

- 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?

Solution: We can use the same concept as the last problem. The total number of suited cards is the number of combinations for each suit multiplied by the number of suits. Using this idea, the number of combinations of suited cards is:

$$\begin{aligned}\text{Combinations} &= 4 \times \binom{13}{2} \\ &= 4 \times \frac{13!}{2!(13-2)!} \\ &= 312\end{aligned}$$

Thus, the total number of pocket pairs is 312.

To calculate the probability of being dealt suited cards, we simply use divide the number of suited cards by the total number of possible starting hands, which gives us:

$$\begin{aligned}\text{Probability} &= \frac{312}{1326} \\ &= \frac{4}{17}\end{aligned}$$

Thus, the probability of being dealt suited cards is $\frac{4}{17}$.

- 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?

Solution: Note that having pocket pairs and having suited cards is mutually exclusive, e.g. you cannot have a suited pocket pair. This is because in a deck of cards, no two cards are of the same face value and suit at the same time. Therefore, the number of combinations and the probability Tom's strategy entails is simply the two combinations/probabilities added up.

$$\text{Combinations} = 78 + 312 = 390$$

$$\text{Probability} = \frac{1}{17} + \frac{4}{17} = \frac{5}{17}$$

Thus, there are 390 unique combinations of cards in Tom's strategy, which constitutes $\frac{5}{17}$ of the total number of cards.

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 $1, 2, \dots, 141$. 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?

Solution: $\frac{1}{11}$.

There are 11 tickets that start with 4, and only one of them is 46. These 11 are

$$4, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49.$$

- 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?

Solution: $\frac{1}{25}$.

There are 25 tickets that contain a 4 somewhere, and only one of them has the number 46. These 25 are

14 tickets end with 4:

$$4, 14, 24, 34, \mathbf{44}, 54, 64, 74, 84, 94, 104, 114, 124, 134$$

12 tickets with 4 as the tens digit:

$$40, 41, 42, 43, \mathbf{44}, 45, 46, 47, 48, 49, 140, 141$$

Minus 1 ticket that is counted twice:

44

- 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?

Solution: $\frac{1}{24}$.

There are 24 tickets that contain exactly one 4, and only one of them has the number 46. These 24 are all the ones listed in part (b), except for 44:

4, 14, 24, 34, 40, 41, 42, 43, 45, 46, 47, 48, 49, 54, 64, 74, 84, 94, 104, 114, 124, 134, 140, 141.

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?

Solution: $\left(\frac{6}{26}\right)^2$.

Let's define two events.

- A: yesterday's Stringle string starts with a vowel.
- B: today's Stringle string starts with a vowel.

We are asked to find $P(A \cap B)$.

There are 26 possibilities for the first letter, each of which are equally likely. 6 of these are vowels. Thus, for any given string, the chance that it starts with a vowel is $\frac{6}{26}$. This says $P(A) = P(B) = \frac{6}{26}$.

Using the multiplication rule, the probability of both yesterday's and today's string starting with a vowel is the probability of yesterday's string starting with a vowel ($\frac{6}{26}$) times the probability of today's string starting with a vowel, assuming that yesterday's did. Since each string is equally likely to be chosen each day, the probability of today's string starting with a vowel is unaffected by the assumption that yesterday's string also started with a vowel, so it's still $\frac{6}{26}$. Therefore, the probability that both strings start with a vowel is

$$\begin{aligned} P(A \cap B) &= P(A) * P(B|A) \\ &= P(A) * P(B) \quad \text{since } A \text{ and } B \text{ are independent in this case} \\ &= \frac{6}{26} * \frac{6}{26} \\ &= \left(\frac{6}{26}\right)^2. \end{aligned}$$

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

Solution: $\frac{6}{26} + \frac{6}{26} - \left(\frac{6}{26}\right)^2$.

Using the same events A and B as defined in the previous part, now we are asked to find $P(A \cup B)$. First, we should recognize that events A and B are not mutually exclusive, meaning that it's possible for both yesterday's and today's strings to start with a vowel. As we calculated in the previous part, $P(A \cap B) = \left(\frac{6}{26}\right)^2$. Then,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{6}{26} + \frac{6}{26} - \left(\frac{6}{26}\right)^2 \end{aligned}$$

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

Solution: $\frac{20}{26} * \frac{19}{25} * \frac{18}{24} * \frac{17}{23} * \frac{16}{22} * \frac{15}{21}$.

We can solve this using the multiplication rule with many events:

- E_1 : the first letter is not a vowel
- E_2 : the second letter is not a vowel
- E_3 : the third letter is not a vowel
- E_4 : the fourth letter is not a vowel
- E_5 : the fifth letter is not a vowel
- E_6 : the sixth letter is not a vowel

We need to find $P(E_1 \cap E_2 \cap \dots \cap E_6)$. The multiplication rule says we can write this as a product of the probability of each event occurring, assuming that all prior events have occurred.

For example, $P(E_1) = \frac{20}{26}$ because 20 out of 26 letters are not vowels. Then, assuming the first letter is not a vowel, the probability of E_2 is $\frac{19}{25}$ because 19 out of the 25 remaining letters are not vowels. Continuing on in this way gives

$$\frac{20}{26} * \frac{19}{25} * \frac{18}{24} * \frac{17}{23} * \frac{16}{22} * \frac{15}{21}.$$

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

Solution: $\frac{6}{26} * \frac{5}{25} * \frac{4}{24} * \frac{3}{23} * \frac{2}{22} * \frac{1}{21}$.

Similarly to how we solved the previous part, we can define events

- E_1 : the first letter is a vowel
- E_2 : the second letter is a vowel
- E_3 : the third letter is a vowel
- E_4 : the fourth letter is a vowel
- E_5 : the fifth letter is a vowel
- E_6 : the sixth letter is a vowel

We need to find $P(E_1 \cap E_2 \cap \dots \cap E_6)$. The probability of E_1 is $\frac{6}{26}$. Then, assuming the first letter is a vowel, there are 5 vowels remaining out of 25 letters remaining, so $P(E_2|E_1) = \frac{5}{25}$.

Continuing in this fashion gives

$$\frac{6}{26} * \frac{5}{25} * \frac{4}{24} * \frac{3}{23} * \frac{2}{22} * \frac{1}{21}.$$

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

Solution: $\frac{6}{26}$.

We'll use the complement rule, as it's easier to calculate the probability that the word has no J. To not have a J, we need to make sure each letter is not a J. This is similar to making sure each letter is not a vowel, as we did in a previous part. Since 25 letters are not J, the probability that today's string has no J is

$$\frac{25}{26} * \frac{24}{25} * \frac{23}{24} * \frac{22}{23} * \frac{21}{22} * \frac{20}{21}.$$

Interestingly, this simplifies to $\frac{20}{26}$, which means the probability that today's string includes J is $1 - \frac{20}{26} = \frac{6}{26}$.

A direct interpretation is that of the 26 letters, 6 distinct letters are chosen to participate in today's string. J, just like any other, has a $\frac{6}{26}$ chance of being selected.

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

Solution: $\frac{1}{26} * \frac{1}{25} * \frac{1}{24} * \frac{1}{23} * \frac{1}{22} * \frac{1}{21}$

Think of yesterday's Stringle string as fixed, and we just need to find the probability that each letter of today's string matches the corresponding letter of yesterday's string.

The first letter matches with probability $\frac{1}{26}$ since there are 26 letters and only one of them works. Assuming this is a match, the probability of the second letter matching is $\frac{1}{25}$ because there are no repeated letters, so the second letter must be one of 25 remaining choices. Carrying on in this way, the probability that today's string matches yesterday's is given by

$$\frac{1}{26} * \frac{1}{25} * \frac{1}{24} * \frac{1}{23} * \frac{1}{22} * \frac{1}{21}$$