DSC 40A - Homework 5

Due: Monday, November 8 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. Make sure to correctly assign pages to Gradescope when submitting.

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 45 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Note: Problems 1 and 2 refer to a supplemental Jupyter Notebook, which can be found **at this link**.

Problem 1. Transformation Tuesday

The logistic function, also known as the "sigmoid" function, is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The logistic function σ (has nothing to do with standard deviation) is used in a variety of fields. Pertinently, it is used to model the growth of populations and spread of diseases. You'll also see it later on in your data science career when you learn about logistic regression.

a) Show that the inverse of the logistic function is given by

$$\sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$

where log represents the natural logarithm with base e.

Hint: Recall, one strategy to find the inverse of a function $y = f(x)$ is to write $x = f(y)$ and solve for y.

b) *Note*: Parts (b), (c), and (d) of this question should not take very much time; you've already done the heavy lifting in part (a).

Suppose we have a dataset $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ and want to use least squares to fit a prediction rule

$$H(x) = \sigma(w_0 + w_1 x)$$

This is **not** linear in our parameters, w_0 and w_1. However, through a transformation, we can frame it as a linear prediction rule.

Using the process from Lecture 10, transform $H(x)$ into a prediction rule that is linear in terms of the parameters w_0 and w_1. Specify a design matrix X and observation vector $\vec{z}$ such that the optimal w_0^* and w_1^* are given by the solution to the normal equations $X^T X \vec{w}^* = X^T \vec{z}$. Your answers for X and $\vec{z}$ may involve x_i's, y_i's, $\sigma(\cdot)$, and/or $\sigma^{-1}(\cdot)$.

c) 🥑🥑 In the supplemental Jupyter Notebook, linked **here**, use the provided code and dataset to define the design matrix and observation vector you specified in the previous part and to find w_0^* and w_1^* for the prediction rule $H(x) = \sigma(w_0 + w_1 x)$. In your PDF writeup, provide a screenshot of the code you wrote as well as of the resulting visualization.

d) 🥑🥑 As you saw in the supplemental Jupyter Notebook in the previous part, our prediction rule was a good fit to our data.

What issue would arise using this technique if there were points in our dataset such that $y_i = 0$ or $y_i = 1$?

Problem 2. What do you k-mean?

a) 🥑🥑🥑🥑 Consider the five data points given below, $\vec{x}_1$ through $\vec{x}_5$.

$$\vec{x}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix}, \ \vec{x}_2 = \begin{bmatrix} 5 \\ 76 \end{bmatrix}, \ \vec{x}_3 = \begin{bmatrix} 1 \\ 8 \end{bmatrix}, \ \vec{x}_4 = \begin{bmatrix} 2 \\ 9 \end{bmatrix}, \ \vec{x}_5 = \begin{bmatrix} 3 \\ 78 \end{bmatrix}$$

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 clustering algorithm by hand. What are the two centroids and what are the two clusters found after this first iteration?

b) 🥑🥑🥑🥑 In the supplemental Jupyter Notebook, linked **here**, you will find a walkthrough of using k-Means Clustering on 209-dimensional data involving countries around the world. At the bottom of that notebook you will find two questions; write the answers to those questions here.

Problem 3. License Plates

In this problem, we will examine license plates from the Canadian province of Ontario, home to Billy the avocado farmer. In Ontario, license plates consist of 4 letters followed by three numbers. All letters are uppercase, and repeated characters are allowed.

BRAX-959 is an example of an Ontario license plate.

a) 🥑🥑🥑 What is the probability that two randomly generated license plates match? You may leave your answer as a product of powers of fractions.

b) 🥑🥑 What is the probability that a randomly generated license plate begins with a vowel?

c) 🥑🥑🥑 What is the probability that a randomly generated license plate begins with a vowel or ends in an odd number? Simplify your answer.

Problem 4. Nine Lives

In this question, we will consider two fair 9-sided dice, each with faces numbered 1, 2, 3, ..., 9.

a) 🥑🥑 Suppose you roll the two dice and look at just one of them. You see that it's an 8. What is the probability that the sum of the two die rolls is 14?

b) 🥑🥑🥑 Suppose you roll the two dice and look at both of them. You see that at least one of them is an 8. What is the probability that the sum of the two die rolls is 14?

c) 🥑🥑🥑 Suppose you roll two the two dice and look at both of them. You see that exactly one of them is an 8. What is the probability that the sum of the two dice rolls is 14?

Hint: it is not your answer to part (a) or part (b).

Problem 5. 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: If E and F are two events, $E \cap F$ is also an event. Also, intersections/"and"s are "associative", meaning that $E \cap F \cap G = (E \cap F) \cap G = E \cap (F \cap G)$; the same applies for unions/"or"s.

b) 🥑🥑🥑🥑 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:

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

Some hints and guidance:

- *While it's a great idea to draw Venn diagrams to reason to yourself why this property holds true, we are looking for an algebraic proof here, not a visual derivation.*

- *At some point, you may need to use the fact that if E, F, and G are events, then $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$. 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) 🥑🥑🥑 To identify what students find most important in DSC 10, we want to administer a survey to the students in DSC 20, DSC 30, and DSC 40A. Consider the following information:

- There are 300 students taking at least one of DSC 20, DSC 30, or DSC 40A right now.

- 200 students are taking DSC 20 right now, and 50 students taking DSC 30 right now. There are no students taking both DSC 20 and DSC 30 right now.

- 50 students are taking both DSC 20 and DSC 40A right now, and 30 students are taking both DSC 30 and DSC 40A right now.

Suppose I choose a single student uniformly at random from the population of students taking at least one of DSC 20, DSC 30, and DSC 40A. What is the probability that they are enrolled in DSC 40A? Simplify your answer.

Hint: Use the result in part (b).