Nameera Faisal Akhtar

Data 100, Fall 2020

Homework 1

Due Date: Thursday, September 3, 11:59PM

Total Points: 24

Submission Instructions

You must submit this assignment to Gradescope by **Thursday, September 3rd, at 11:59 PM**. While Gradescope accepts late submissions, you will not receive **any** credit for a late submission if you do not have prior accommodations (e.g. DSP).

You can work on this assignment in any way you like.

- One way is to download this PDF, print it out, and write directly on these pages (we've provided enough space for you to do so). Alternatively, if you have a tablet, you could save this PDF and write directly on it.
- Another way is to use some form of LaTeX. Overleaf is a great tool.
- You could also write your answers on a blank sheet of paper.

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like options 1 and 3 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must** assign pages to each question correctly (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our tutors. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on Piazza.

Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your submission.

Preliminary: Sums

Here's a recap of some basic algebra written in sigma notation. The facts are all just applications of the ordinary associative and distributive properties of addition and multiplication, written compactly and without the possibly ambiguous "...". But if you are ever unsure of whether you're working correctly with a sum, you can always try writing $\sum_{i=1}^{n} a_i$ as $a_1 + a_2 + \cdots + a_n$ and see if that helps.

- You can use any reasonable notation for the index over which you are summing, just as in Python you can use any reasonable name in 'for name in list'. Thus $\sum_{i=1}^{n} a_i = \sum_{k=1}^{n} a_k$.
- $\sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$
- $\sum_{i=1}^{n} d = nd$
- $\sum_{i=1}^{n}(ca_i + d) = c\sum_{i=1}^{n} a_i + nd$

We commonly use sigma notation to compactly write the definition of the arithmetic mean (commonly known as the average): $\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Summations

1. (6 points) For each of the statements below, either prove that it is true by using the definitions above, or show that it is false by providing a counterexample. For our purposes, each a_i and x_i is a real number. *Hint: One way to prove something is to start with one side of the equation, and manipulate it through a valid series of steps until it looks like the other side of the equation.*

   (a) $\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i} = \sum_{i=1}^{n} x_i$ (Assume $\sum_{i=1}^{n} a_i \neq 0$)

   Counterexample:
   Let $n = 3$
   $a_1 = 1 \quad x_1 = 4$
   $a_2 = 2 \quad x_2 = 5$
   $a_3 = 3 \quad x_3 = 6$

   $\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i} \stackrel{?}{=} \sum_{i=1}^{n} x_i$

   $\frac{a_1x_1 + a_2x_2 + a_3x_3}{a_1 + a_2 + a_3} \stackrel{?}{=} x_1 + x_2 + x_3$

   $\frac{4 + 10 + 18}{1 + 2 + 3} \stackrel{?}{=} 4 + 5 + 6$

   $\frac{32}{6} \neq 15$

   Thus, this counterexample works.

   $\therefore$ FALSE.

   (b) $\sum_{i=1}^{n} a_3 x_i = n a_3 \bar{x}$

   RHS: $n a_3 \bar{x}$

   $= n \cdot a_3 \cdot \frac{1}{n}(x_1 + x_2 + \ldots x_n)$

   $= a_3(x_1 + x_2 + \cdots + x_n)$

   $= a_3 \sum_{i=1}^{n} x_i$

   $= \sum_{i=1}^{n} a_3 x_i$: LHS

   $\therefore$ TRUE

   (c) $\sum_{i=1}^{n} a_i x_i = n\bar{a}\bar{x}$

   Counterexample:
   Let $n = 3$
   $a_1 = 1 \quad x_1 = 4$
   $a_2 = 2 \quad x_2 = 5$
   $a_3 = 3 \quad x_3 = 6$

   $\bar{a} = \frac{1+2+3}{3} = 2$

   $\bar{x} = \frac{4+5+6}{3} = 5$

   $\sum_{i=1}^{n} a_i x_i \stackrel{?}{=} n\,\bar{a}\,\bar{x}$

   $a_1x_1 + a_2x_2 + a_3x_3 \stackrel{?}{=} 3 \cdot 2 \cdot 5$

   $4 + 10 + 18 \stackrel{!}{=} 3 \cdot 2 \cdot 5$

   $32 \neq 30$

   Thus, this counterexample works.

   $\therefore$ FALSE

Calculus

2. (4 points) Let $\sigma(x) = \dfrac{1}{1+e^{-x}}$.

   (a) Show that $\sigma(-x) = 1 - \sigma(x)$.

RHS: $1 - \sigma(x)$

$= 1 - \dfrac{1}{1+e^{-x}}$

$= \dfrac{1+e^{-x}-1}{1+e^{-x}}$

$= \dfrac{e^{-x}}{1+e^{-x}}$

To get rid of negative exponents, multiply top and bottom by e^x.

$= \dfrac{e^{-x} \cdot e^{x}}{(1+e^{-x})e^{x}}$

$= \dfrac{1}{1+e^{x}} = \sigma(-x)$: LHS

Thus $\sigma(-x) = 1 - \sigma(x)$. ✓

   (b) Show that the derivative can be written as:

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$$

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

Using chain rule: $\dfrac{d}{dx}\sigma(x) = -(1+e^{-x})^{-2} \cdot -e^{-x}$

$= \dfrac{e^{-x}}{(1+e^{-x})^2}$

RHS: $\sigma(x)(1-\sigma(x))$

$= \sigma(x)\sigma(-x)$ Using (a)

$= \dfrac{1}{1+e^{-x}} \cdot \dfrac{1}{1+e^{x}}$

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

$= \dfrac{1}{1+e^{x}+e^{-x}+1}$

Multiplying top and bottom by e^{-x}

$= \dfrac{1 \cdot e^{-x}}{(1+e^{x}+e^{-x}+1)e^{-x}}$

$= \dfrac{e^{-x}}{e^{-x}+1+e^{-2x}+e^{-x}}$

$= \dfrac{e^{-x}}{1+2e^{-x}+e^{-2x}}$

$= \dfrac{e^{-x}}{(1+e^{-x})^2}$

$= \dfrac{d}{dx}\sigma(x)$: LHS

Thus $\dfrac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$ ✓.

Minimization

3. (3 points) Consider the function $f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$. In this scenario, suppose that our data points $x_1, x_2, ..., x_n$ are fixed, and that c is the only variable.

   Using calculus, determine the value of c that minimizes $f(c)$. You must justify that this is indeed a minimum, and not a maximum.

To find the max/min, we need $f'(c) = 0$.

$f'(c) = \dfrac{1}{n}\sum_{i=1}^{n} 2(x_i - c) \cdot -1$

$= -\dfrac{2}{n}(x_1 - c + x_2 - c + x_3 - c + \dots + x_n - c)$

$-\dfrac{2}{n}(-nc + x_1 + x_2 + x_3 + \dots + x_n) = 0$

$-nc = -x_1 - x_2 - x_3 - \dots - x_n$

$c = \dfrac{x_1 + x_2 + x_3 + \dots + x_n}{n}$

To verify that this is indeed a min and not a max, perform 2nd derivative test.

$f'(c) = \dfrac{2nc}{n} - x_1 - x_2 - \dots - x_n$

$f''(c) = 2 > 0$

Hence, $c = \dfrac{x_1 + \dots x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ is a minimum. ✓

Probability and Statistics

4. (4 points) Much of data analysis involves interpreting proportions – lots and lots of related proportions. So let's recall the basics. It might help to start by reviewing the main rules from Data 8, with particular attention to what's being multiplied in the multiplication rule.

   (a) The Pew Research Foundation publishes the results of numerous surveys, one of which is about the trust that Americans have in groups such as the military, scientists, and elected officials to act in the public interest. A table in the article summarizes the results.

   Pick one of the options (1) or (2) to answer the question below; if you pick (1), tell us what p is. Then, explain your choice.

   The percent of surveyed U.S. adults who had a great deal of confidence in both scientists and religious leaders

   1. is equal to p%.
   2. cannot be found with the information in the article.

It cannot be found with the information in the article. Let A and B be the following events:

A = "surveyed U.S. adult has great deal of confidence in scientists"

B = "surveyed U.S. adult has great deal of confidence in religious leaders"

By the multiplication rule from Data 8, $P(A \text{ and } B) = P(A) \times P(B|A) = P(A) \times \frac{P(B \cap A)}{P(A)}$.

Since we do not know the overlap between U.S adults who had a great deal of confidence in religious leaders and U.S. adults who had a great deal of confidence in scientists i.e $P(A \cap B)$ b/c we only know individual probabilities, we cannot compute/determine this percent.

   (b) Toyota is one of most commonly owned makes of cars in our county (Alameda). A car heading from Berkeley to San Francisco is pulled over on the freeway for speeding. Suppose I tell you that the car is either a Toyota or a Lamborghini, and you have to guess which of the two is more likely.

   What would you guess, and why? Make some reasonable assumptions and explain them (data scientists often have to do this), and justify your answer.

Assume that the kind of cars found in Alameda county is a good representation of the kind of cars found when heading from Berkeley to San Francisco (close proximity).

Now, consider Bayes' Rule:

Let F = "car is Toyota" and E = "car gets pulled over", then

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\bar{F})P(\bar{F})}$$ is the probability that the car is a Toyota given that the car gets pulled over. Since there are more Toyotas than Lamborghinis, $P(F) >> P(\bar{F})$, and hence $P(F|E)$ is guaranteed to be a relatively large number, and hence the car more likely to be pulled over is a Toyota.

5. (3 points) Consider the following scenario:

   Only 1% of 40-year-old women who participate in a routine mammography test have breast cancer. 80% of women who have breast cancer will test positive, but 9.6% of women who don't have breast cancer will also get positive tests.

   Suppose we know that a woman of this age tested positive in a routine screening. What is the probability that she actually has breast cancer? (Note: You must show all of your work, and also simplify your final answer to 3 decimal places.)

Use Bayes Theorem.

Let:

F be the event "a 40-year old woman participating in a routine screening has breast cancer"

E be the event "a 40-year old woman tested positive in a routine screening"

$P(F) = 0.01$

$P(\bar{F}) = 1 - 0.01 = 0.99$

$P(E|F) = 0.8$

$P(E|\bar{F}) = 0.096$

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\bar{F})P(\bar{F})}$$

$$= \frac{(0.8)(0.01)}{(0.8)(0.01) + (0.096)(0.99)}$$

$$= \frac{0.008}{0.008 + 0.09504}$$

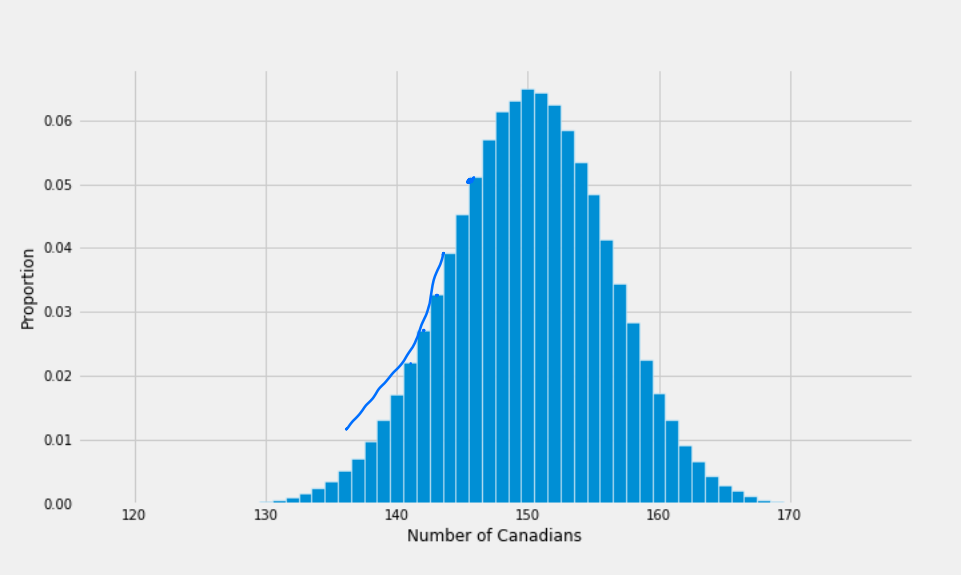
$$\approx 0.078$$

Thus, the probability that a 40 year old woman has breast cancer given that she tested positive is roughly 7.8%.

6. (2 points) Suppose we collected a sample of 200 students at UC Berkeley, and 150 of them happened to be Canadian (so, if we were to select a student uniformly at random from our sample, there is a 0.75 chance that they are Canadian).

   For inferential purposes, we choose to bootstrap this sample 500,000 times. That is, we simulate the act of re-sampling (with replacement) 200 students from our observed sample, and each time we record the number of Canadians in our re-sample.

   We provide a histogram of the sampling distribution below.



   What is the standard deviation of the sampling distribution shown above? Select the closest option below, and **explain your answer**.

   A. 1.5

   B. 6.1

   C. 12.4

   D. 10.1

   Hint: While it is possible to calculate the answer, the histogram has all of the information you need.

Let σ = standard deviation.

Since the inflection point of the normal curve is exactly 1 σ away from the mean,

σ = mean − inflection point

= 150 − 144

= 6

Since this is closest to 6.1, the answer is B.

Welcome Survey

7. (2 points) In order for the teaching staff to best ensure you have a stellar Data 100 experience, we've put together a short welcome survey for us to get to know more about you. **When you have finished the survey, you will receive a codeword. Please write this codeword as your answer to question 7.**

codeword : "central beef overview"