Data 100, Summer 2022

Homework 1

Due Date: Monday, June 27, 11:59PM

Submission Instructions

You must submit this assignment to Gradescope by **Monday, June 27, 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) or beyond the use of 5 slip days.

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.

Calculus

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

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

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

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

Minimization

2. (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.

Probability and Statistics

3. (2 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.

   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.

4. (3 points) Consider the following scenario:

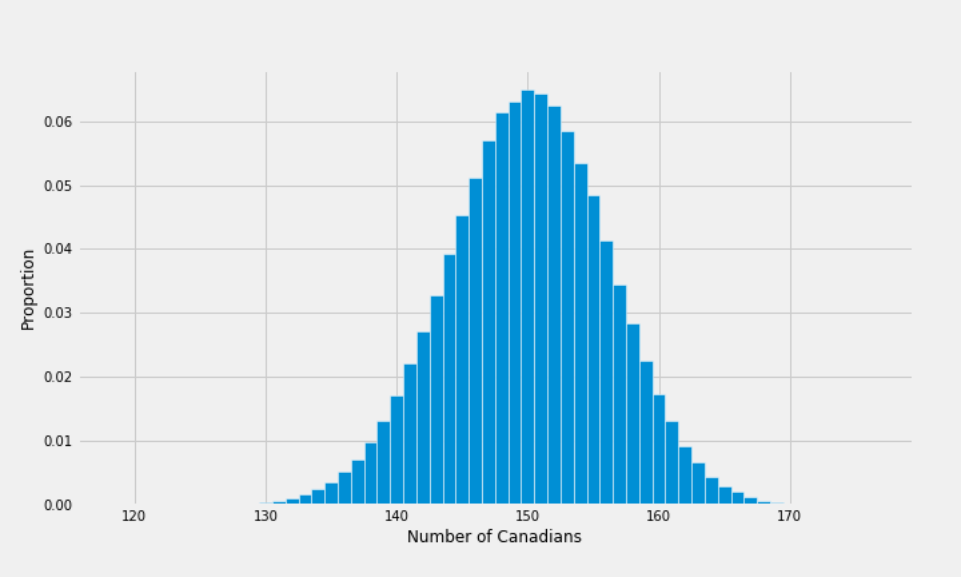
   Only 1% of 40-year-old women who participate in a routine mammography test have breast cancer. 80% of women who have breast cancer will test positive, but 9.6% of women who don't have breast cancer will also get positive tests.

   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.)

5. (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.

Linear Algebra

6. (6 points) A common representation of data uses matrices and vectors, so it is helpful to familiarize ourselves with linear algebra notation, as well as some simple operations.

   Define a vector $\vec{v}$ to be a column vector. Then, the following properties hold:

   - $c\vec{v}$ with c some constant, is equal to a new vector where every element in $c\vec{v}$ is equal to the corresponding element in $\vec{v}$ multiplied by c. For example, $2\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}2\\4\end{bmatrix}$.
   - $\vec{v}_1 + \vec{v}_2$ is equal to a new vector with elements equal to the elementwise addition of $\vec{v}_1$ and $\vec{v}_2$. For example, $\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}-3\\4\end{bmatrix} = \begin{bmatrix}-2\\6\end{bmatrix}$.

   The above properties form our definition for a **linear combination** of vectors. $\vec{v}_3$ is a linear combination of $\vec{v}_1$ and $\vec{v}_2$ if $\vec{v}_3 = a\vec{v}_1 + b\vec{v}_2$, where a and b are some constants.

   Oftentimes, we stack column vectors to form a matrix. Define the **rank** of a matrix A to be equal to the maximal number of linearly independent columns in A. A set of columns is **linearly independent** if no column can be written as a linear combination of any other column(s) within the set.

   For example, let A be a matrix with 4 columns. If three of these columns are linearly independent, but the fourth can be written as a linear combination of the other three, then $\text{rank}(A) = 3$.

   For each part below, you will be presented with a set of vectors, and a matrix consisting of those vectors stacked in columns. State the rank of the matrix, and whether or not the matrix is full rank. If the matrix is not full rank, state a linear relationship among the vectors—for example: $\vec{v}_1 = \vec{v}_2$.

   (a) $\vec{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix}1\\1\end{bmatrix}$, $A = \begin{bmatrix}| & |\\ \vec{v}_1 & \vec{v}_2\\ | & |\end{bmatrix}$

   (b) $\vec{v}_1 = \begin{bmatrix}3\\-4\end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix}0\\0\end{bmatrix}$, $B = \begin{bmatrix}| & |\\ \vec{v}_1 & \vec{v}_2\\ | & |\end{bmatrix}$

(c) $\vec{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$, $\vec{v}_3 = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$, $C = \begin{bmatrix} | & | & | \\ \vec{v}_1 & \vec{v}_2 & \vec{v}_3 \\ | & | & | \end{bmatrix}$

(d) $\vec{v}_1 = \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}$, $\vec{v}_2 = \begin{bmatrix} -2 \\ -2 \\ 5 \end{bmatrix}$, $\vec{v}_3 = \begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix}$, $D = \begin{bmatrix} | & | & | \\ \vec{v}_1 & \vec{v}_2 & \vec{v}_3 \\ | & | & | \end{bmatrix}$