Data 100, Spring 2024

Homework 1B

Due Date: Thursday, January 25, 11:59 PM

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Thursday, January 25, 11:59 PM**. Please read the syllabus for **the grace period policy**. No late submissions beyond the grace period will be accepted. While course staff is happy to help you if you encounter difficulties with submission, we may not be able to respond to last-minute requests for assistance (TAs need to sleep, after all!). **We strongly encourage you to plan to submit your work to Gradescope several hours before the stated deadline.** This way, you will have ample time to contact staff for submission support.

There are four parts to this assignment listed on Gradescope:

- **Homework 1 Coding**: Submit your Jupyter Notebook zip file for Homework 1A, which can be generated and downloaded from DataHub using the grader.export() cell provided.

- **Homework 1A Written**: Gradescope will automatically submit the PDF from the zip file submitted earlier. You do not need to submit anything to this assignment yourself, but please check that the submission went through properly.

- **Homework 1B Written**: Submit a PDF to Gradescope that contains all your answers to all questions in Homework 1B.

- Syllabus Quiz: The assignment is multiple-choice style on Gradescope. You may change or update your answers anytime before the deadline.

To receive credit on this assignment, **you must submit both your coding and written portions to their respective Gradescope portals as well as the syllabus quiz**.

You can answer the below Homework 1B written questions in one of many ways:

1. Type your answers. We recommend LaTeX, the math typesetting language. Overleaf is a great tool to type in LaTeX.

2. Download this PDF, print it out, and write directly on these pages. If you have a tablet, you may save this PDF and write directly on it.

3. Write your answers on a blank sheet of physical or digital paper. Note: If you write your answers on physical paper, use a scanning application (e.g., CamScanner, Apple Notes) to generate a PDF.

Important: When submitting Homework 1B on Gradescope, you **must tag pages to each question correctly** (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our readers. Failure to do this may result in a score of 0 for untagged questions.

You are responsible for ensuring your submission follows our requirements and that the automatic submission for Homework 1A Written answers went through properly. We will not be granting regrade requests nor extensions to submissions that don't follow instructions. If you encounter any difficulties with submission, please don't hesitate to contact staff before the deadline.

Collaboration Policy

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 discuss the assignments with others, please include their names below.

N/A

Homework 1A Manually Graded Questions

0. This is not a question. This is a reminder to make sure that your Homework 1A manually graded questions (automatically submitted as a PDF by Gradescope when HW 1 Coding assignment is submitted) were properly generated and uploaded.

Calculus and Algebra

1. (3 points) Let $f(x) = \dfrac{4x^2 - 2x + 1}{x}$. Find the local minimum and maximum point(s) of $f(x)$, and write them in the form (a, b).

Answer. First, we note that a local minimum/maximum occurs when $f'(x) = 0$. Then, we observe that:

$$
\begin{aligned}
f(x) &= \frac{4x^2 - 2x + 1}{x} \\
&= 4x - 2 + \frac{1}{x} \\
&= 4x - 2 + x^{-1} \\
f'(x) &= 4 - x^{-2}
\end{aligned}
$$

Setting this to be equal to zero, we get:

$$
\begin{aligned}
4 - x^{-2} &= 0 \\
4 &= x^{-2} \\
x^2 &= \frac{1}{4} \\
x &= \pm\sqrt{\frac{1}{4}} \\
x_1 &= \frac{1}{2} \\
x_2 &= -\frac{1}{2}
\end{aligned}
$$

Then, to find if it's a minimum or maximum, we have to look at $f''(x)$:

$$
f''(x) = 2x^{-3}.
$$

We evaluate $f''(x)$ at the two extremas we found earlier and see:

$$f''(x_1) = 2x_1^{-3} = 2(8) = 16 > 0$$
$$f''(x_2) = 2x_2^{-3} = 2(-8) = -16 < 0$$

So, we see that $x_1 = \frac{1}{2}$ is a local maximum, and $x_2 - \frac{1}{2}$ is a local minimum.

Then, evaluating f at these points, we get:

$$f(x_1) = 4x_1 - 2 + x_1^{-1}$$
$$= 2 - 2 + 2$$
$$= 2$$
$$f(x_2) = 4x_2 - 2 + x_2^{-1}$$
$$= -2 - 2 - 2$$
$$= -6$$

Therefore, we have that the local maximum is $\left(\frac{1}{2}, 2\right)$ and the local minimum is $\left(-\frac{1}{2}, -6\right)$.

2. (3 points) Let $g(x, y, z) = 10z^4x^2 - e^{5x}ln(y) + \frac{z^4}{y^3} + \frac{z-ln(y)}{z-x}$.

   (a) Holding all other variables constant, take the partial derivative of $g(x, y, z)$ with respect to x, $\frac{\partial}{\partial x}g(x, y, z)$.

   Answer. We can proceed as follows:

   $$\frac{\partial}{\partial x}g(x, y, z) = \frac{\partial}{\partial x}\left(10z^4x^2 - e^{5x}\ln(y) + \frac{z^4}{y^3} + \frac{z-\ln(y)}{z-x}\right)$$

   $$= \frac{\partial}{\partial x}10z^4x^2 - \frac{\partial}{\partial x}e^{5x}\ln(y) + \frac{\partial}{\partial x}\frac{z^4}{y^3} + \frac{\partial}{\partial x}(z-\ln(y))(z-x)^{-1}$$

   $$= 20z^4x - 5e^{5x}\ln(y) + \frac{z-\ln(y)}{(z-x)^2}$$

   (b) Holding all other variables constant, take the partial derivative of $g(x, y, z)$ with respect to y, $\frac{\partial}{\partial y}g(x, y, z)$.

   Answer. We can proceed as follows:

   $$\frac{\partial}{\partial y}g(x, y, z) = \frac{\partial}{\partial y}\left(10z^4x^2 - e^{5x}\ln(y) + \frac{z^4}{y^3} + \frac{z-\ln(y)}{z-x}\right)$$

   $$= \frac{\partial}{\partial y}10z^4x^2 - \frac{\partial}{\partial y}e^{5x}\ln(y) + \frac{\partial}{\partial y}z^4y^{-3} + \frac{\partial}{\partial y}\frac{z}{z-x} - \frac{\partial}{\partial y}\frac{\ln(y)}{z-x}$$

   $$= -\frac{e^{5x}}{y} - \frac{3z^4}{y^4} - \frac{1}{y(z-x)}$$

Probability and Statistics

3. (3 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 (Chapter 9.5), with particular attention to what's being multiplied in the multiplication rule.

   For this question, assume we have a bag filled with 18 marbles: 7 blue marbles, 5 red marbles, and 6 yellow marbles. (Note: You can leave your solutions in equation form, and values can be left as proportions.)

   (a) Yash selects 2 marbles *without replacement*. What's the probability that the selected marbles have the same color?

   Answer. The probability for this occurring is:

   $$\frac{7}{18} \cdot \frac{6}{17} + \frac{5}{18} \cdot \frac{4}{17} + \frac{6}{18} \cdot \frac{5}{17}.$$

   (b) Yash selects 6 marbles *with replacement*. What's the probability that he sees at least 1 yellow marble?

   Answer. The probability that this occurs is:

   $$1 - \left(\frac{12}{18}\right)^6$$

Content Warning:
This question includes discussion about cancer. If you feel uncomfortable with this topic,
please contact your GSI or the instructors.

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. You can leave the answer
as either a percent or decimal value.)

Answer. We want to find the following:

$$\mathbb{P}(\text{Positive}|\text{Tested Positive})$$

Then, we can proceed as follows, using Bayes' Theorem:

$$\mathbb{P}(\text{Positive}|\text{Tested Positive}) = \frac{\mathbb{P}(\text{Tested Positive}|\text{Positive}) \cdot \mathbb{P}(\text{Positive})}{\mathbb{P}(\text{Tested Positive})}$$

$$= \frac{\mathbb{P}(\text{Tested Positive}|\text{Positive}) \cdot \mathbb{P}(\text{Positive})}{\mathbb{P}(\text{Tested Positive}|\text{Positive})\mathbb{P}(\text{Positive}) + \mathbb{P}(\text{Tested Positive}|\text{Not Positive})\mathbb{P}(\text{Not Positive})}$$

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

$$\approx 0.07764$$

$$= 7.764\%$$

Linear Algebra

5. (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 **column 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. If all columns in a matrix are linearly independent, it means that the matrix is **full column rank**.

   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 rank$(A) = 3$. Alternatively, if all four columns of A were linearly independent, rank$(A) = 4$., and A would be full column rank.

   For each of the following matrices, state the rank of the matrix and whether or not the matrix is full column rank. If the matrix is not full column rank, state that it is not full column rank and give a linear relationship among the vectors—for example: $\vec{v}_1 = \vec{v}_2$.

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

   > **Answer.** The rank of the matrix is 2, and thus is full column rank.

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

   > **Answer.** The rank of the matrix is 1, and is not full column rank. This is because we can write $v_2 = 0v_1$.

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

Answer. The rank of the matrix is 2, and thus is not full column rank. We see that we have $10v_1 + 2v_2 = v_3$.

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

Answer. The rank of the matrix is 2. We observe that $v_2 + v_3 = v_1$.