MSSE Data C200, Spring 2021

Homework 0

Due Date: Wednesday, February 3, *11:59PM*

Total Points: 9

Submission Instructions

You must submit this assignment to Gradescope by **Wednesday, February 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 $\sum_{i=1}^{n} a_i \neq 0$)

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

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

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.