Biostatistics PhD Placement Assessment – Data Science

July 30, 2024

- The purpose of this assessment is to help us understand your data science and computing background so that you, your academic advisor, and the Graduate Program leadership can develop a study plan for you that best supports your future success.

- You have 48 hours to work on the problems below. Your work must be your own and you are not allowed to talk to others.

- If you have questions about the assessment problems, please email Hongkai Ji (hji@jhu.edu) and Brian Caffo (bcaffoweb@jhu.edu).

- Please email your solutions with codes, documents, and github links to Hongkai Ji (hji@jhu.edu) as a zipped folder.

Question 1

Using R or Python scrape the wikipedia page on natural disasters:

https://en.wikipedia.org/wiki/List_of_natural_disasters_by_death_toll

for the tables of the 20th and 21st century all cause disasters into a data frame, tibble or pandas data frame.

Convert the death toll to numbers using the midpoints when a range is given and the bound when an upper or lower bound is given (example 20,000+ converts to 20000).

Merge the 20th and 21st century data frames and plot the death toll (vertical / y axis) by year (horizontal / x axis) color coded by kind of disaster.

Put your answer in a github repo. Include your web scraping and plotting code. In your readme, describe your plot to a layman.

You may use any web resource at your disposal to complete this task including LLMs. However, please do not converse with other students / live chat …

Question 2

Let x and y be vectors of length n. Consider minimizing the loss $L(b) = ||y - b\,x||^2$ over b where b is a scalar. (The solution is $b = \langle x, y \rangle / ||x||^2$.) Write a function in R or python that takes two

vectors or numpy vectors and iterates to solve for b using gradient descent. That is, the update is

Update(b) = Current value of b - e * Derivative of L with respect to b evaluated at the current value of b

Where e is a user-supplied real number usually called the learning rate or step size.

Test your function out on some randomly generated normal vectors where you know the value of b. How does the performance of the algorithm's depend on e? When does the algorithm fail and why?