CAP 5610
Assignment #2 Solution

February 3, 2025

Arman Sayan

---

1 Types of Attributes [10 points]

Classify the following attributes as nominal, ordinal, interval, ratio. **Explain why**.

(a) Rating of an Amazon product by a person on a scale of 1 to 5

(b) The Internet Speed

(c) Number of customers in a store.

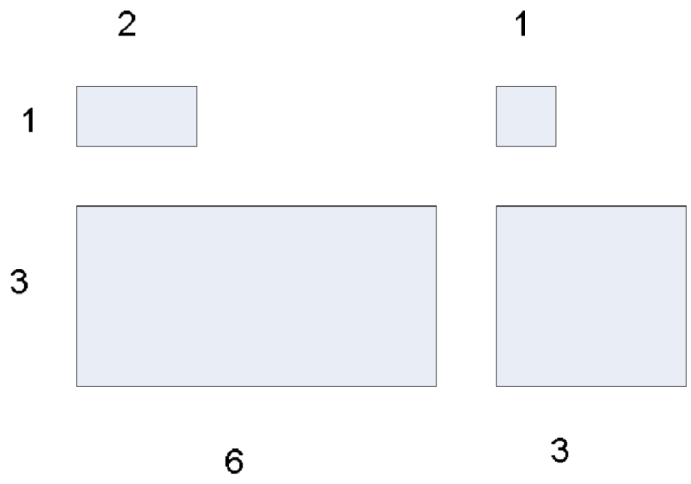
(d) UCF Student ID

(e) Letter grade (A, B, C, D)

Ans: Write answer

2 Distance/Similarity Measures [20 points]

Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. **Please explain your choice**.



1. [10 points] Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)?

   Ans: Write answer

2. [10 points] Which proximity measure would you use to group the boxes based on their size?

   Ans: Write answer

3 Coding Question [20 points]

Please write a Python code to calculate Cosine similarity, and Euclidean distance using NumPy. The input can be two randomly generated vectors or fixed vectors written by yourself. Note that: For Coding Questions, please **do not** directly call linear regression and non-linear regression built-in functions in existing library packages such as scikit-learn. You may call basic computation functions built in Numpy.

Ans: Write answer

4 Coding Question [25 points]

Please implement a Linear Regression to find the best linear model for the provided linear data. Please plot the result using "matplotlib.pyplot".

Note that

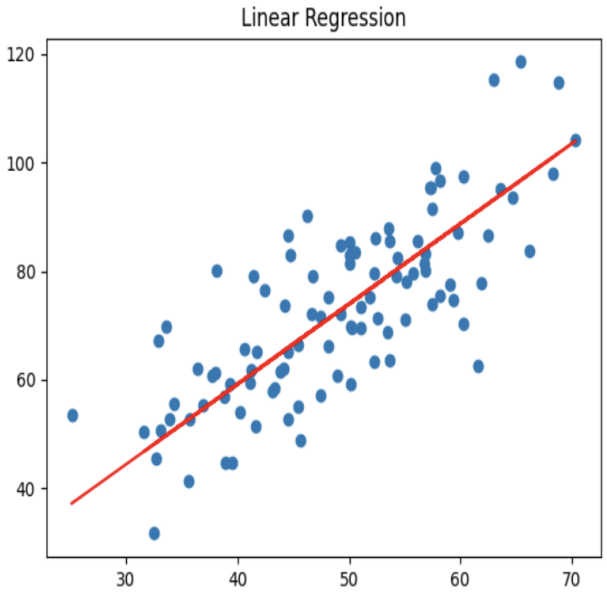
(1) The linear model is in the following format $Y = mX + c$

(2) Use MSE as the loss function

(3) You may use "pandas" to read the csv file and load the values into two vectors X and Y.

(4) Use Gradient Descent for the training. You may choose fixed learning rate (such as 0.0001) and epochs (such as 1000) without considering mini-batch.

(5) The result will look like the following image.



Ans: Write answer

5 Coding Question [25 points]

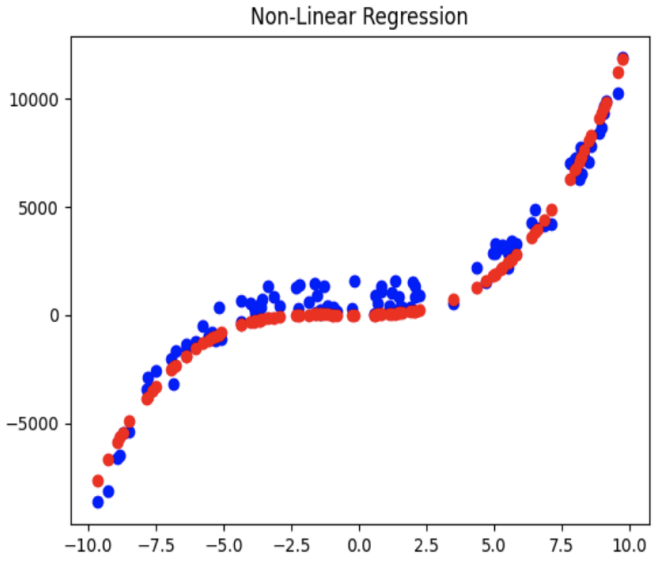
Please implement a non-linear regression to find the best cubic function model for the provided non-linear data. Please plot the result, too.

(1) The cubic function is in the following format: $Y = aX^3 + bX^2 + cX + d$

(2) Use MSE as the loss function.

(3) Use Gradient Descent for the training. You may choose fixed learning rate (such as 0.000001 (1e-6)) and epochs (such as 10000) without considering mini-batch. It may take 10-15 seconds to finish the running for 10000 steps. Please be patient.

(4) The result will look like the following



Ans: Write answer