Applied Linear Structures for Computing
Mini-project 2, Total Grade= 100 points+5Bonus Points
Instructor: M. Hadi Amini
TA: Leila Zahedi
For questions regarding homework please reach out to the TA (lzahe001@fiu.edu)

[1]**Part I - Regression**

Questions 1: (35 points)
Using the analytical solution you learned for univariate regressions, compute the parameters θ_0 and θ_1 for the following data:

$$x = [1, 3, 5, 7, 9]$$

$$y = [2, 6, 10, 14, 18]$$

ans: y= -360 +74x

Question 2:(15 points)
Goal: The goal of this part of your assignment is to take the theory you have learned on how linear algebra is behind regression models and to put it to work by programming your own multivariate regression.

First, write a function to compute the analytical solution to a multivariate regression, which is provided to you in your lecture slides. This function should use all but the last row of the data provided to you in the file *reg_dataset.json*. Then, you will use the solution from the first part to make a prediction for the last row of values of the data set. Print out the error value for the prediction of the last row.

Part II - Clustering

Question 1:(40 points) Given the following sequence, $S = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$, and two starting centroids, $c_1 = 6$, $c_2 = 15$, use the k-means algorithm to update the centroids until convergence. Show two update steps of the k-means in your solution.

Question 2:(15 points) Goal: For this clustering part you will attempt to use the k-means algorithm to make a prediction over the Iris data set.

The data set is provided to you in a json file titled *iris_dataset.json*. First of all it, contains a list with

---

[1]This assignment and its solution is designed and developed by Luiz Pereira and M. Hadi Amini. Please do not distribute publicly and do not share with other without instructor's permission.

the data, which is essentially a four column matrix with the following features: sepal length, sepal width, petal length, and petal width. Then, it has a target list, which is a vector containing the correct membership to which cluster each row belongs (these being the 3 categories of flowers). Your job is to do the following:

1. Read the json formatted data (for python, you can use this link for reference: https://www.programiz.com/python-programming/json)

2. Separate the entire data and the last row so you train your model over everything but the last row

3. Build the model (you may try to do it yourself or use sk-learn)

4. Make a prediction using the last row of data

Different models are more better and more useful for different situations. Regardless of whether your prediction was right or wrong, your goal is to make sure your code runs and builds the model so you understand how to take the theory you learned in class and put it into practice.

Submission Details

You will submit **one single pdf** file, including explanation for any details of your project (i.e. if you're using a compiled language, details on compilation) and also copy and paste your code in the submission file. Title your pdf file as *firstName_lastName* . Make sure to add comments to your code to clarify what is happening.