Assignment 1 – Data Analysis & Linear Regression

Part A:

Write a python program that performs analysis on the "gene_table.txt" dataset attached in order to:

1. Compute the total number of genes and compute the number of different gene biotypes.
2. Compute the minimum, maximum, average and median number of known isoforms.
3. Compute, for each chromosome, the number of genes it contains by plotting a bar chart. Also, print the chromosomes with the corresponding number of genes in increasing order.
4. Computes, for each chromosome, the percentage of genes located on the + strand.
5. Compute, for each biotype, the average number of transcripts associated to genes belonging to the biotype.

Part B:

The attached dataset ***"diabetic_kidney_disease.csv"*** contains 110 records of patients with diabetic kidney disease. We need to examine the relation between fasting blood glucose (FBG) and urinary albumin creatinine ratio (UACR) because kidney disease is a common complication of diabetes.

Write a python program that uses linear regression with gradient descent to predict the value of UACR based on the FBG of a patient since this value can be used for early detection of kidney disease in diabetic patients.

Note: You will need to normalize the data before applying linear regression. You can use minmax normalization where z is the normalized value and z = (x – min) / (max – min).

So, given the hypothesis function Y = C_1 + C_2 X;
Y (target variable) = UACR, X (predictor) = FBG, C_1 and C_2 are the parameters of the function:

1. Split the data into 2 parts: training and testing. Choose the value of learning rate and the number of iterations.
2. Implement gradient descent to optimize the parameters of the function (C1 and C2).
3. Calculate the error of the hypothesis function to see how it changes with every iteration. (Hint: You will need to calculate the error in every iteration.)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

4. Use optimized hypothesis function to make predictions on new data.
5. Try different values of the learning rate and the iterations to see how this changes the accuracy of the model.
6. Plot the initial line that you started with and at the end, plot the line produced from linear regression (the line that best fits the data).

Important Notes:

- You can only use “pandas”, “numpy” and “matplotlib” libraries. ***(Don’t use “sklearn”)***
- The maximum number of students in a team is 3 and the minimum is 2.
- No late submission is allowed.
- Cheating students will take negative grades and no excuses will be accepted.

Good Luck