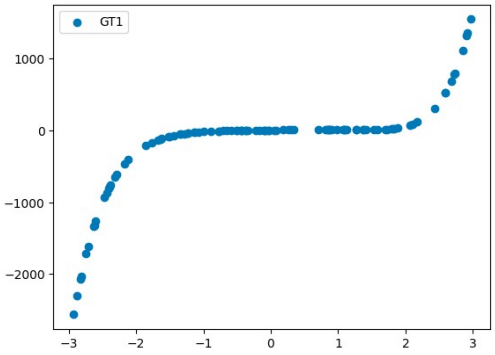
Assignment 1
CS 5630/6630
Due: 09/20/2024 11:59 pm

[Question 1] **[15 points]**

Suppose that you are conducting a scientific experiment where you are observing the effects of one variable (*x_train.npy* and *x_test.npy*) on the output (*y_train.npy* and *y_test.npy*). On visualizing the relationship between the variables, you see the following plot:



Your goal is to come up with a linear regression model that can take the training data (*x_train.npy* and *y_train.npy*) and model the relationship between the variables *x* and *y*. You should implement your own version of linear regression either using gradient descent or normal equations. **You SHOULD NOT use any pre-packaged library such as Sci-Kit Learn.**

Here are somethings to keep in mind for tackling this problem:

1. Try to plot this relationship on your own using matplotlib. You can also visualize the test data to see if it gives you any clues about the underlying relationship between the variables.
2. Based on the observations from your visualization, address the following:

a. What kind of relationship exists between the variables (linear, non-linear, or something else)?
b. Does the current model need improvements to capture the observed relationship effectively?
- If so, what modifications would you suggest to better capture the data pattern? ii. What types of transformations or additional features might improve the model's performance? After implementing these changes, how does each new feature or transformation impact the fit of the model? Plot and analyze them individually to assess the impact.

You will need to write a short report detailing your thought process, the code you wrote in Python to implement the linear regression model and the equation that models the relationship between *x* and *y* that you found. You should provide evidence that corroborates your final statement such as plots, prediction errors, etc.

[Question 2] **[15 points]**

Imagine that you are a realtor in Auburn. You have data points (See excel file. Last column is the target variable.) that correspond to the recent sales of different houses in and around Auburn. Your goal is to help estimate the prices of houses that one can use to sell or buy listings. Can you use your knowledge of linear regression to find the best regression model? Use your implementation from Question 1 (without any basis functions) to answer the following questions.

1. What is the average least squares error for the given data using your simple linear regression model?

2. Which feature has the greatest impact on the predicted house prices? How can you justify this finding? Could this feature alone serve as a predictor for house prices?
3. Which feature has the least impact on the predicted house prices? How can you verify this? What is the impact on model performance when this feature is removed from the dataset?

[Question 3] **[Total: 20 points]**

Suppose you are working on a binary classification problem where you are tasked with predicting whether a certain type of tumor is malignant (y = 1) or benign (y = 0) based on a set of features. You are given a dataset that includes these features and corresponding labels for training.
Your goal is to implement a logistic regression model from scratch (without using any pre-packaged libraries such as Sci-Kit Learn) to classify the tumors. You should use gradient descent to find the optimal parameters that minimize the cost function.
Here are some tasks to guide you through this problem**:** **[Code Implementation 11 points]**

1. **Implement the Sigmoid Function:** Write the code for the sigmoid function that will map your predictions to probabilities.
2. **Derive the Cost Function:** Formulate the cost function for logistic regression and explain why it is different from the cost function in linear regression.
3. **Apply Gradient Descent:** Implement the gradient descent algorithm to minimize the cost function and optimize the parameters of your model. Ensure that you include a stopping criterion based on either a fixed number of iterations or a threshold for changes in the cost function.
4. **Evaluate the Model:** Once the model is trained, evaluate its performance on a test dataset. Plot the decision boundary and explain how the logistic regression model separates the two classes.

Answer the following questions: [**9 points]**

1. **Why is the sigmoid function critical in logistic regression? How does it influence the interpretation of the output?**
2. **What are the key differences between the logistic regression model and the linear regression model, particularly in the context of the cost function and the type of problem each is used for?**
3. **How would you modify your logistic regression model to handle multi-class classification problems?**

<u>You SHOULD NOT use any pre-packaged library such as Sci-Kit Learn.</u>

Submission Requirements:
You will need to submit the following as a single IPYNB file:

1. Use the “text cells” on Colab to write your report.
2. Include a cell with README instructions at the top of your notebook to note on any dependencies that are required to run your code.

Note:

1. If your code does not run on Colab, you will not get any credit for the code segment. We will only grade what is in your report.
    a. This includes any syntax errors due to indentation, unnamed/unknown libraries that were not listed in the README file, etc.

2. Please submit code only in Python and in the IPython notebook format. You can write your answers as part of the notebook if you do not want a separate report file, but it must be comprehensive.
    a. Any code not in Python will not be graded at all.