Assignment 1

Due Date:02/18/2025

Total Points: 100

In this exercise, you will implement linear regression ***from scratch*** using the programming language of your choice. **Please make sure to avoid using toolbox from R, MATLAB, Python, or any other programming language.** You will implement the Gradient Descent Algorithm that we have discussed in class to find out the parameters for $\ominus$. One way to verify that gradient descent is working correctly is to look at the value of J ($\ominus$) and check that it is decreasing with each iteration. For implementing some of the principles of programming, try to modularize the code as much as possible and for improved code readability, please make sure to thoroughly comment the code clearly explaining what you did and why you did what you did. In this assignment, we will use breast cancer dataset for experiments. Assignment 1 contains four sections. Analysis is a crucial aspect of the assignment, so for each subpart try to answer the question in more detail to receive full credit and justify what you did in your implementation as well as the results you obtained. Also, please divide the data into training and test data and use the **test data** to evaluate performance.

1. Linear Regression with One Variables: **(15 points)**

    a. Implement linear regression to predict tumor_size (Tumor Size) using "mean_texture" feature from dataset. "tumor_size" will be the target variable.

    b. Evaluate performance using metrics (such as Mean Squared Error (MSE), R-squared (R^2), and Adjusted R-squared (Adjusted R^2). You may also use graphs for explaining your observations.

2. Linear Regression with Two Variables: **(15 points)**

    a. Add 'lymph_node_status' as an additional input feature to the previous linear regression model. Does adding this feature improve the performance of the model? Compare the performance of the models in Question 1 and Question 2 using evaluation metrics (such as Mean Squared Error (MSE), R-squared (R^2), and Adjusted R-squared (Adjusted R^2). You may also use graphs for explaining your observations.

3. Stepwise Linear Regression: **(50 points)**

In this section, you will explore stepwise linear regression to determine the most relevant features for predicting tumor size.

a. Select any 5 features out of the 10 provided below. Implement forward stepwise linear regression with the chosen features. The process involves iteratively adding one feature at a time from your selection. After adding each feature, evaluate the model's performance using metrics such as Mean Squared Error, R-squared, Adjusted R-squared, or BIC-Bayesian Information Criterion (preferred). Choose the feature that contributes the most to improving the model's performance and add it to the model. Continue this iterative process for a total of 5 iterations. Explain your selection criteria for adding or removing features.

   Features: 'mean_radius', 'mean_perimeter', 'mean_area', 'mean_smoothness', 'mean_symmetry', 'mean_fractal_dimension', 'worst_radius', 'worst_area', 'worst_symmetry', 'lymph_node_status'

b. For this task, you will be given a list of 10 features as below. Implement backward stepwise linear regression, beginning with a model that includes all 10 features. Remove one feature at a time from the model and evaluate the model's performance after each removal. Remove the feature that has the minimal adverse effect on the model's performance. Continue this iterative process for a total of 5 iterations. Please provide a detailed explanation of your criteria for selecting which features to remove or retain during this process.

   Features: 'mean_radius', 'mean_perimeter', 'mean_area', 'mean_smoothness', 'mean_symmetry', 'mean_fractal_dimension', 'worst_radius', 'worst_area', 'worst_symmetry', 'lymph_node_status'

c. Compare the final model obtained from forward stepwise regression with the final model obtained from backward stepwise regression. Which one is better? Discuss the differences in terms of the selected features, model performance.

d. Compare the performance of the model built using the features from Q.2 (a) with the resultant accuracies of the model built using the selected features Q.3(c). Which set of features performed better?

Evaluate model performance using metrics like Mean Squared Error, R-squared, Adjusted R-squared or BIC-Bayesian Information Criterion (preferred). You may also use graphs for explaining your observation.

4. Regularization and Feature Scaling: **(20 points)**

a. For the best performing model in Q.3 (Model from Q.3(d)), does regularization improve the performance?

b. Does Feature Scaling improve the performance for the model in Question 3(d)?

Evaluate performance using metrics like Mean Squared Error, R-squared, Adjusted Rsquared. You may also use graphs for explaining your observation.

Please do not use any in-built library such as scikit-learn etc.

Please make sure to submit a zipped file in Dropbox on Pilot. The file should be named "YourName_Assignment-1" and should contain a PDF format report, a code file.

Academic Integrity

Discussion of course contents with other students is an important part of the academic process and is encouraged. However, it is expected that course programming assignments, homework assignments, and other course assignments will be completed on an individual basis (unless specified otherwise). Students may discuss general concepts with one another, but may not, under any circumstances, work together on the actual implementation of any course assignment. If you work with other students on "general concepts" be certain to acknowledge the collaboration and its extent in the assignment. Unacknowledged collaboration will be considered dishonest. "Code sharing" (including code from previous quarters) is strictly disallowed. "Copying" or significant collaboration on any graded assignments will be considered a violation of the university guidelines for academic honesty.

If the same work is turned in by two or more students, all parties involved will be held equally accountable for violation of academic integrity. You are responsible for ensuring that other students do not have access to your work: do not give another student access to your account, do not leave printouts in the recycling bin, pick up your printouts promptly, do not leave your workstation unattended, etc. If you suspect that your work has been compromised notify me immediately. If you have any questions about collaboration or any other issues related to academic integrity, please see me immediately for clarification. In addition to the policy stated in this syllabus, students are expected to comply with the Wright State University Code of Student Conduct (http://www.wright.edu/students/judicial/conduct.html) and in particular the portions Pertaining to Academic Integrity http://www.wright.edu/students/judicial/integrity.html) at all times.