Assignment 2

**Due Date: 04/03/2025**

**Total Points: 150**

In this assignment, you will perform exploratory data analysis (EDA) on the given dataset and implement Logistic Regression ***from scratch*** using a programming language of your choice. **Please make sure to avoid using toolbox from R, MATLAB, Python, or any other programming language.** As part of your implementation of logistic regression (LR), you will implement the Gradient Descent Algorithm that we have discussed in class to find out the parameters for Θ. One way to verify gradient descent is working as expected is to monitor the value of J(Θ) whether it is decreasing with each training 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 on the code clearly explaining what you did and why you did what you did. In your report, include a README that states how your code is supposed to be run to obtain the expected results. 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.

We will use the same breast cancer dataset which was used in assignment 1 for this assignment.

**1.** Exploratory Data Analysis **(40 points)**

1. Consider the following numeric variables in the dataset: mean\_radius, mean\_texture, mean\_perimeter, mean\_area, mean\_smoothness, mean\_compactness, mean\_concavity and mean\_concave\_points. Summarize the statistics of these variables into count, mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, and maximum.

1. Consider the categorical variable “outcome” in the dataset. Summarize the statistics of variable into count, unique value, top value, and frequency of top value.

1. Is there a way to encode outcome variable from categorical to numerical data type? If so, how would you do that?

1. Do you think there are any redundant features present in the dataset? If so, explain how removing it won't impact the analysis. Also, based on the experiments so far, were there any interesting observations with respect to the variables?

1. What is the correlation between mean\_perimeter and se\_perimeter? **2.** Logistic Regression with One Variable **(20 points)**

1. Can you map the likelihood of breast cancer recurrence (outcome) based on "mean\_area” feature from the dataset?

1. Evaluate performance using a metric discussed in class (such as confusion matrix). You may also use graphs to explain your observations.

1. Logistic Regression with Multiple Variables **(50 points)**

* 1. Design a Logistic Regression model to predict breast cancer recurrence (outcome) using the following 12 variables from the dataset as input features:

Features: mean\_radius, mean\_texture, mean\_perimeter, mean\_area, mean\_smoothness, mean\_compactness, mean\_concavity, mean\_concave\_points, mean\_fractal\_dimension, se\_perimeter, se\_texture, se\_area

* 1. Design a Logistic Regression model to predict breast cancer recurrence (outcome) using forward selection to select the most significant variables in the dataset as input features. Which subset of features gave you the best performance? What are your thoughts on these features getting selected? (Use 12 features from 3a as Input features)

* 1. Compare the performance of the full model built using all the features in (3a) with the resultant accuracies of the full model using the selected features (3b). Which set of features performed better?

**Evaluate performance of your models using a metric discussed in class (such as the confusion matrix). You may also use graphs to explain your observations.**

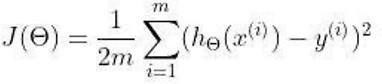
1. Experimenting with regularization and Cost function. **(40 points)**

* 1. Regularization and Feature Scaling: **(20 points)** 
     1. For the best performing model in Q.3 (Model from 3c), does regularization improve the performance? II. Does Feature Scaling improve the performance for the model in Q 3c?

**Evaluate performance for each case using a metric discussed in class (such as confusion matrix). You may also use graphs to explain your observations.**

* 1. Cost Function: **(20 points)**

* + 1. Keeping the best model after the experiments from Q.4a, design a Logistic Regression model to predict breast cancer recurrence (outcome) by changing the cost function to the following (Mean Squared Error).



* + 1. Compare the performance of both the models (4.b.i and 4.a.). Do they give the same solution with a difference in cost function?

**Evaluate performance for each case using a metric discussed in class (such as confusion matrix). You may also use graphs to explain your observations.**

**Please do not use scikit-learn. Please make sure to submit a zipped file in Dropbox on Pilot. The file should be named "Assignment-2\_YourName" and should contain a dataset, a code file (please use relative paths when reading/importing the dataset), a PDF-format report, and a README.txt. Ensure that your code is error- free; there should be no errors or warnings when running it. 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**](http://www.wright.edu/students/judicial/conduct.html)) **in particular the portions Pertaining to Academic Integrity** [**http://www.wright.edu/students/judicial/integrity.html) at**](http://www.wright.edu/students/judicial/integrity.html)) **all times.**