## **LOGISTIC REGRESSION FROM SCRATCH**

### Introduction

Logistic regression is a fundamental statistical method used for binary classification problems. In this report, we will implement logistic regression from scratch using the Iris dataset, a popular dataset in machine learning that contains measurements of iris flowers and their corresponding species. Our objective is to classify iris species based on their features.

### Dataset Description

The Iris dataset consists of 150 instances with the following attributes:

1. x\_1
2. x\_2
3. y

### Data Preprocessing

Before applying logistic regression, we need to preprocess the data and split the dataset into training and testing sets.

1. Selecting Relevant attributes: We will check for null data.
2. Splitting the Dataset: We will split the dataset into a training set (80%) and a testing set (20%) to evaluate our model's performance.

### Implementation of Logistic Regression from Scratch

Logistic regression models the probability that a given input belongs to a particular class. The logistic function (sigmoid) is used to map predicted values to probabilities:

where:

* ŷ is the predicted value.
* *X* is the matrix of input features.
* *w* is the vector of weights.
* ​*b* is the bias term.

An activation function is applied on this to predict the probability which in this case is the *Sigmoid Activation Function.*

​The binary cross-entropy loss function is used to measure the performance of the model:

where

* 𝑛 is the number of instances
* is the predicted probability
* is the actual label.

#### Gradient Descent Algorithm

We will use gradient descent to minimize the MSE. The updates for weights and bias are given by:

Where

* is the cost function.
* 𝛼 is the learning rate.
* The partial derivatives are:

### Results and Discussion

After training the logistic regression model from scratch, we obtain the following accuracies:

* Training Accuracy: This value indicates how well the model fits the training data.
* Testing Accuracy: This value indicates how well the model generalizes to new, unseen data.

Higher accuracy values indicate better model performance. In practice, we aim to achieve high accuracy on both training and testing sets to ensure that our model generalizes well.

### Conclusion

In this report, I have implemented logistic regression from scratch using the Iris dataset. I have preprocessed the data and used gradient descent to minimize the binary cross-entropy loss. The model's performance was evaluated using training and testing accuracy. This exercise demonstrates the fundamental concepts of logistic regression and the importance of data preprocessing and evaluation metrics.

### Future Work

Future improvements could include:

1. Regularization: Applying techniques like L1 or L2 regularization to prevent overfitting.
2. Advanced Models: Exploring more complex models like decision trees or neural networks to improve predictions.