Logistic Regression: A Detailed Explanation with Formulas

1. Introduction to Logistic Regression

Logistic Regression is a fundamental classification algorithm in machine learning used to predict binary outcomes (0/1, True/False, Yes/No). Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities that map to discrete classes using the logistic function.

2. Mathematical Formulation of Logistic Regression

Logistic regression is based on the sigmoid (logistic) function, which is defined as:

sigma(z) = 1 / (1 + e^(-z))

where:

- z = w0 + w1x1 + w2x2 + ... + wnxn is a linear combination of input features with learned weights w.
- sigma(z) outputs values between 0 and 1, representing probabilities.
- If sigma(z) > 0.5, we classify the input as class 1; otherwise, class 0.

Log-Likelihood Function and Optimization

To estimate the parameters (weights w), logistic regression maximizes the log-likelihood function:

L(w) = Summation[y log(h_w(x)) + (1 - y) log(1 - h_w(x))]

where:

- y is the actual class label,

- h_w(x) is the predicted probability,

- m is the number of training examples.

The cost function for logistic regression is:

J(w) = - (1/m) Summation[y log h_w(x) + (1 - y) log (1 - h_w(x))]

Optimization is typically done using **Gradient Descent**, which iteratively updates the weights to minimize the cost function.

3. Why Use Logistic Regression?

Logistic Regression is widely used due to:

- **Simplicity & Interpretability**: Easy to understand and interpret results.

- **Probability Scores**: Unlike decision trees, logistic regression provides probability estimates.

- **Efficiency**: Works well when the dataset is linearly separable.

- **Baseline Classifier**: Often used as a benchmark before applying complex models.

4. Advantages of Logistic Regression

- **Efficient for Binary Classification**: Works well for problems requiring two class labels.

- **Handles Large Feature Sets**: Can manage high-dimensional data with appropriate regularization.

- **Interpretable Model**: Provides insight into how each feature affects the decision.

- **Fast Training**: Computationally inexpensive compared to deep learning models.

5. Disadvantages of Logistic Regression

- **Assumes Linearity in Features**: Poor performance if the relationship between input variables and output is non-linear.

- **Sensitive to Outliers**: Can be affected by extreme values in the dataset.

- **Limited to Binary Classification**: Extensions are needed for multi-class problems (One-vs-All, Softmax regression).

- **Feature Engineering Required**: Performance depends on how well features are selected and transformed.

6. Example of Logistic Regression

Dataset: Predicting whether a student will pass an exam based on study hours.

Study Hours	Pass (Y/N)
1	0
2	0
3	0
4	1
5	1
6	1

Using logistic regression, we fit a model:

sigma(w0 + w1 * StudyHours)

where w0 and w1 are learned using gradient descent.

If sigma(z) gives a probability p greater than 0.5, the student is predicted to pass; otherwise, fail.

7. Conclusion

Logistic Regression is a powerful, interpretable classification algorithm useful for many practical applications. While it has limitations, understanding its mathematical formulation and use cases helps in choosing the right model for classification problems.