Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for classification problems. Linear regression predicted continuous values whereas logistic regression predicts the probability that an input belongs to a specific class. It is used for binary classification where the output can be one of two possible categories such as Yes/No, True/False or 0/1. It uses sigmoid function to convert inputs into a probability value between 0 and 1.

Mathematical Equations

Logistic regression is designed for binary classification (e.g., Purchased = 0 or 1). It models the probability of the positive class (class 1) using the sigmoid function:

$$P(y = 1 \mid x) = \sigma(z) = 1 / (1 + e^{(-z)})$$

where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$. The sigmoid maps z to **[0, 1]**, representing the probability of class 1. For predictions, a threshold (typically 0.5) is used:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{if } \sigma(z) < 0.5 \end{cases}$$

The model optimizes the binary cross-entropy (log loss) function:

$$J(\beta) = -(1/m) \sum [\, y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i)) \,]$$

Coefficients β are updated via gradient descent:

$$J(\beta) = -(1/m) \sum [\, y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i)) \,]$$

where α is the learning rate.

Assumptions

1. **Independent observations:** Each data point is assumed to be independent of the others means there should be no correlation or dependence between the input samples.
2. **Binary dependent variables:** It takes the assumption that the dependent variable must be binary, means it can take only two values. For more than two categories SoftMax functions are used.
3. **Linearity relationship between independent variables and log odds:** The model assumes a linear relationship between the independent variables and the log odds of the dependent variable which means the predictors affect the log odds in a linear way.
4. **No outliers:** The dataset should not contain extreme outliers as they can distort the estimation of the logistic regression coefficients.

Evaluation Metrics

1. **Accuracy:** Proportion of correct predictions.

   Accuracy = (TP + TN) / (TP + TN + FP + FN)

2. **Precision:** Proportion of predicted positives that are correct.

   Precision = TP / (TP + FP)

3. **Recall:** Proportion of actual positives identified.

   Recall = TP / (TP + FN)

4. **F1-Score:** Harmonic mean of precision and recall.

   F1 = 2 * (Precision * Recall) / (Precision + Recall)

5. **ROC-AUC:** Area under the ROC curve, measuring class separation.\

   ROC-AUC = ∫ TPR(FPR) d(FPR)

   Where:

   TP = True Positives

   TN = True Negatives

   FP = False Positives

   FN = False Negatives

   TPR = True Positive Rate = Recall

   FPR = False Positive Rate = FP / (FP + TN)

Dataset Description

Source: I sourced a dataset from Kaggle. Its link is:

https://www.kaggle.com/datasets/dragonheir/logistic-regression

The Social Network Ads dataset (400 samples) includes:

- Features: Gender (categorical: Male/Female), Age (integer), EstimatedSalary (integer).
- Target: Purchased (binary: 0 = not purchased, 1 = purchased).

Results

Evaluation Metrics

- **Accuracy**: 0.89
- **Precision**: 0.91
- **Recall**: 0.75
- **F1-Score**: 0.82
- **ROC-AUC**: 0.97

Classification Report:

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.88	0.96	0.92	52
1	0.91	0.75	0.82	28
Accuracy			0.89	80
Macro Avg	0.90	0.86	0.87	80
Weighted Avg	0.89	0.89	0.88	80

Confusion Matrix:

[[50 2]

[7 21]]