SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)
(Established under section 3 of the UGC Act 1956)
Re - accredited by NAAC with 'A' Grade
Founder: Prof.Dr. S. B. Mujumdar, M.Sc.,Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Name: Naivedya Rai
PRN: 20070122083

Machine Learning Lab Assignment 7

Aim: Apply Logistic Regression algorithm on a sample case study and data set. Evaluate results

Theory:

Logistic Regression:

Theory:

Logistic regression is a statistical model used for binary classification, which means it's primarily used to predict one of two possible outcomes based on one or more predictor variables. Despite its name, logistic regression is a classification algorithm, not a regression algorithm like linear regression. It's used when the dependent variable is categorical and represents one of two classes, typically coded as 0 and 1 (or "negative" and "positive").

Here's how logistic regression works:

1. Sigmoid Function:
Logistic regression uses the sigmoid (logistic) function to model the probability that a given input belongs to the positive class (class 1). The sigmoid function is defined as:

$S(z) = 1 / (1 + e^{(-z)})$

In this equation, "z" represents a linear combination of the predictor variables:

$z = \beta0 + \beta1X1 + \beta2X2 + ... + \beta n*Xn$

β0 is the intercept.
β1, β2, ..., βn are the coefficients associated with the predictor variables X1, X2, ..., Xn.

2. Probability Calculation:
The sigmoid function converts the linear combination "z" into a value between 0 and 1. This value represents the estimated probability of the positive class.

3. Thresholding:
By default, if the estimated probability is greater than or equal to 0.5, the model predicts the positive class (1); otherwise, it predicts the negative class (0). You can adjust the threshold for classification based on the specific problem's requirements.

4. Model Training:
To train a logistic regression model, you need a labeled dataset where the dependent variable represents the class labels (0 or 1), and the independent variables are the features used for prediction.
The model's parameters (coefficients β0, β1, β2, ..., βn) are estimated during training using optimization techniques like maximum likelihood estimation (MLE). The goal is to find the values of these parameters that maximize the likelihood of the observed data given the model.

5. Model Evaluation:
Logistic regression models are typically evaluated using metrics such as accuracy, precision, recall, F1-score, ROC curve, and AUC-ROC to assess their performance in classifying new or unseen data.

Metrics for Evaluation of Logistic Regression Model:

1. Confusion Matrix
   A confusion matrix is a table that offers a detailed summary of the performance of a classification model. It assists in determining the accuracy of a model's predictions by comparing projected and actual results.
   True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) make up a confusion matrix (FN).
2. ROC Curve and AUC Score
   The Receiver Operating Characteristic (ROC) curve is a graphical depiction of the performance of a binary classifier model. It depicts the trade−off between the true positive rate (TPR) and the false positive rate (FPR) for various categorization criteria.

A typical statistic for assessing the effectiveness of a binary classification model is the area under the ROC curve (AUC). Better model performance is indicated by a higher AUC value.

3. Residual Plot
   A residual plot is a graph that depicts how the expected value and the residual relate to one another (that is, the difference between the predicted value and the actual value). A residual plot is a graphical tool for assessing the effectiveness of a regression model. It displays the discrepancies between the expected and actual values of the dependent variable on the y−axis and the independent variable on the x−axis.

Example:
To predict whether a loan applicant with default on their loan or not

Sample Dataset:

Applicant	Credit Score	Annual Income ($)	Credit Lines	Defaulted
1	650	50,000	3	0
2	720	75,000	5	0
3	580	25,000	2	1
4	710	80,000	6	0
5	660	60,000	4	0

Logistic Regression model uses the following function:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3)}}$$

Where P(Y = 1) is the probability of defaulting
b0,b1,b2,b3 are all weights of the model
x1,x2,x3 are features (independent variables)

Based on the values of b0,b1,b2,b3 obtained, we can get the probability of defaulting by substituting in the above equation. Depending on the threshold value and the value obtained, the applicant can be accepted or rejected.

Logistic Regression Code:

https://github.com/Naivedya-Rai/ML-Algo-Implement/blob/main/Lab7_logisticRegression.ipynb