Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Jhanvi Parekh

60009210033

CSE(Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

Experiment 5

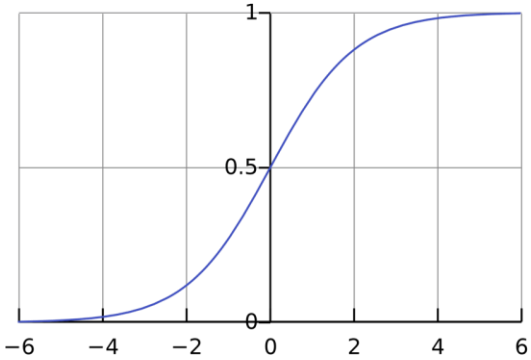
(Logistic Regression)

Aim: Implement Logistic Regression on a given Dataset with binary and multiclass labels.

Theory:

Logistic Regression is a statistical approach and a Machine Learning algorithm that is used for classification problems and is based on the concept of probability. It is used when the dependent variable (target) is categorical. It is widely used when the classification problem at hand is binary; true or false, yes or no, etc. For example, it can be used to predict whether an email is spam (1) or not (0). Logistics regression uses the sigmoid function to return the probability of a label.

Sigmoid Function is a mathematical function used to map the predicted values to probabilities. The function has the ability to map any real value into another value within a range of 0 and 1.



The rule is that the value of the logistic regression must be between 0 and 1. Due to the limitations of it not being able to go beyond the value 1, on a graph it forms a curve in the form of an "S". This is an easy way to identify the Sigmoid function or the logistic functionhttps://colab.research.google.com/drive/1Z8ujrZldNbKBNKGDfr76wNbG8lldoTK0?usp=sharing. In regards to Logistic Regression, the concept

used is the threshold value. The threshold values help to define the probability of either 0 or 1. For example, values above the threshold value tend to 1, and a value below the threshold value tends to 0.

Type of Logistic Regression

1. Binomial: This means that there can be only two possible types of the dependent variables, such as 0 or 1, Yes or No, etc.
2. Multinomial: This means that there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
3. Ordinal: This means that there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Binary Logistic Regression Major Assumptions

1. The dependent variable should be dichotomous in nature (e.g., presence vs. absent).
2. There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29.
3. There should be no high correlations (multicollinearity) among the predictors. This can be assessed by a correlation matrix among the predictors. Tabachnick and Fidell (2013) suggest that as long correlation coefficients among independent variables are less than 0.90 the assumption is met. He aim of training the logistic regression model is to figure out the best weights for our linear model within the logistic regression. In machine learning, we compute the optimal weights by optimizing the cost function. **Cost function:** The cost function J(Θ) is a formal representation of an objective that the algorithm is trying to achieve. In the case of logistic regression, the cost function is called LogLoss (or Cross-Entropy) and the goal is to minimize the following cost function equation:
4. $$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))\right]$$

Gradient descent is a method of changing weights based on the loss function for each data point. We calculate the LogLoss cost function at each input-output data point. We take a partial derivative of the weight and bias to get the slope of the cost function at each point. (No need to brush up on linear algebra and calculus right now. There are several matrix optimizations built into the Python library and Scikit-learn, which allow data science enthusiasts to unlock the power of advanced artificial intelligence without coding the answers themselves). Based on the slope, gradient descent updates the values for the bias and the set of weights, then reiterates the training loop over new values (moving a step closer to the desired goal). This iterative approach is repeated until a minimum error is reached, and gradient descent cannot minimize the cost function any further. We can change the speed at which we reach the optimal minimum by adjusting the learning rate. A high learning rate changes the weights more drastically, while a low learning rate changes them more slowly.

Lab Assignments to complete in this session:

Use the given dataset and perform the following tasks:
Dataset 1: Synthetic Dataset
Dataset 2: IRIS.csv
Dataset 3: Airlines_Passanger.csv

1. Perform required Logistic Regression from scratch on Dataset 1. Compare the F1 score of the LR model built from scratch and built using python library.
2. Perform Multimodal classification on Dataset 2 using python library.
3. Compare the results of Logistic Regression model with and without regularization.

With and without using libraries:

https://colab.research.google.com/drive/1Z8ujrZldNbKBNKGDfr76wNbG8lldoTK0?usp=sharing

For dataset 1

https://colab.research.google.com/drive/1nkJ3HmcXjk6d96JFrTWCUBTpuXP5dDbq?usp=sharing

For dataset 2

https://colab.research.google.com/drive/1eRuX4n-Rmn9PnXeQXOOtZR4wZVlnDMYR?usp=sharing