Machine Learning
CSE475

Lab Manual on

Logistic Regression using Python

Md Mohsin Uddin
Senior Lecturer
Department of Computer Science & Engineering
East West University

Objective:

In this lab, we will learn the differences between logistic and linear regression. After that, we will do a hands-on implementation of logistic regression in Python using Jupyter notebook and other related libraries such as Pandas, sklearn, NumPy, seaborn and matplotlib. In the end, we will generate a report for our produced model and measure it from a different aspect.

Tasks:

1. Import necessary libraries
2. Load dataset from CSV file
3. View data to find discrepancies
4. Analyze the data for dependency and resemblance
5. Filter and process the data table
    a. Look for null and NaN values
    b. Check the percentage of irrelevant and unwanted data
    c. Visualize null values with seaborn heatmap and other distribution histograms
    d. fill up the null values with mean values
    e. drop unnecessary and redundant columns
    f. check for non-numeric data columns
    g. Generate dummy variables [1] for them
    h. verify the dataset with a seaborn heatmap
6. Split the data set into training and testing parts in a moderate ratio (train_test_split)
7. Import logistic regression model from sklearn
8. Create a model and predict with the testing dataset
9. Create a model prediction report using the sklearn metrics library which will contain (precision [2] , recall [3] , f1-score[4], accuracy etc.)

***[1] Dummy variables** are numerical variables that represent the actual data.*

Gender	Dummy Variable
Male	0
Male	0
Male	0
Male	0
Female	1
Female	1
Female	1
Male	0

Dummy Variables representing male and female gender.

***[2] Precision** is the ratio of correctly predicted positive observations to the total predicted positive observations.*

***[3] Recall** is the ratio of correctly predicted positive observations to all observations in actual class*

***[4] F1 Score** is the weighted average of Precision and Recall.*