School of Computer Science and Engineering, VIT Chennai.

BCSE209P Machine Learning

Lab-3 Logistic Regression

Faculty : Dr. R. Jothi

Due Date : 23/01/2024

Submit your python code (Jupyter notebook): with output for all the questions.

Q1. Design a binary classifier for classifying the following emails as spam (1) or ham (0). Use email.csv for training. Each column x_i in the csv file represents certain keyword used for spam prediction. The first column x_0 is a bias. If $x_i = 1$ means, keyword x_i is present in the email.

- Print all the parameter values learnt after training
- Show the accuracy on the test set.

Test emails

Assume first four are spam and the last four are not spam.

$x_1 = 0$	$x_2 = 1$	$x_3 = 0$	$x_4 = 0$	$x_5 = 0$
$x_1 = 1$	$x_2 = 1$	$x_3 = 1$	$x_4 = 0$	$x_5 = 1$
$x_1 = 0$	$x_2 = 1$	$x_3 = 1$	$x_4 = 0$	$x_5 = 0$
$x_1 = 1$	$x_2 = 0$	$x_3 = 1$	$x_4 = 0$	$x_5 = 0$
$x_1 = 0$	$x_2 = 1$	$x_3 = 0$	$x_4 = 0$	$x_5 = 1$
$x_1 = 0$	$x_2 = 0$	$x_3 = 0$	$x_4 = 1$	$x_5 = 1$
$x_1 = 0$	$x_2 = 1$	$x_3 = 0$	$x_4 = 1$	$x_5 = 1$
$x_1 = 0$	$x_2 = 0$	$x_3 = 0$	$x_4 = 0$	$x_5 = 1$

Q2. Use appropriate Scikit Library function to apply logistic regression on the same dataset and compare the results with your implementation.

Q3. Also use logistic regression model to predict the risk of having heart disease using the given dataset (heart.csv).

You need to show complete pre-processing steps (identifying null or missing values, normalization, etc.)