भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 306: Machine Learning Lab
Evaluation Assignment 2

Instructions: This assignment is for evaluation and marks will be awarded. You need to complete execution by 12 PM. The saved results folder (consisting .csv files, plots, word file) has to be submitted through a Google form, which will be shared by Teaching Assistant.

1. Download The Pima Indian Diabetes dataset regarding the prediction of diabetes in patients based different given predictors. Write a program to do the followings:
   (a) (3 points) Randomly split the dataset into the following rations of training: validation: testing sets:
      1. 50:10:40
      2. 60:10:30
      3. 70:10:20
      4. 80:10:10
   (b) (6 points) Design different hypothesis ($\hat{y}$) to predict diabetes in the patients using Logistic regression with Batch Gradient Descent (GD) method, **without using in-built Python packages /libraries** for estimation of the parameters for the different training-testing splits obtained from 1(a). Consider model parameters (θ) or (ω) are to be initialized to zero. Apply hyper-parameter tuning for selecting the best model for the following hyper-parameters , considering the given values:
      1. learning rate, α: $\{0.0001, 0.1\}$
      2. ρ: $\{0.001, 0.01\}$ where, ρ signifies absolute error-difference in two consecutive epochs.
      3. epoch, (T): $\{50,100\}$
   (c) (3 points) Calculate training accuracy (after convergence) and test accuracy for different splits.
   (d) (5 points) Report the relevant graphs/ plots for displaying the results , and also write the result analysis.
2. (6 points) Re-implement $Q.1(a) - (d)$ using Stochastic GD (without using in-built package).