भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 306: Machine Learning Lab
Evaluation Assignment 1

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 Mental Health and Digital Behavior dataset regarding the prediction of digital well-being based on different predictors. Write a program to do the followings:

   (a) (2 points) Read the dataset, to get the predictors (features) and the target (the column "digital well-being score").

   (b) (2 points) Normalize the features using Min-Max Normalization (if applicable).

   (c) (3 points) Randomly splits the dataset into Training, Validation and Testing sets, considering the following ratios:

      1. 60:20:20
      2. 70:10:20
      3. 80:10:10

2. Design different hypothesis ($\hat{y}$) to predict the digital well-being score considering all features (predictors), for estimation of the model parameters using Batch Gradient Descent (GD) method (without using in-built package):

   (a) (4 points) Perform hyper-parameter tuning, using Grid-search method (without any inbuilt packages), with respect to all the three splits, separately (from $Q.1$) 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, and during implementation of all questions, consider only ρ for stopping criteria.

   Hint: For hyper-parameter tuning, initially train the model using Batch GD, for all combinations of α and ρ, and then validate to obtain the optimal combination of the hyper-parameters.

   (b) (4 points) Re-train and test the model with the optimal hyper-parameters, and predict the hypothesis ($\hat{y}$), for all the three training: testing splits separately.

   (c) (2 points) Calculate the mean-RSS with respect to the hypothesis obtained for the three training: testing splits, separately.

   (d) (2 points) Plot the obtained results of mean-RSS considering percentage of training samples in X-axis and the obtained mean-RSS in Y-axis.

(e) (5 points) Re-implement $Q.2(a) - (d)$ using Stochastic GD (without using in-built package).

(f) (5 points) Re-implement $Q.2(a) - (d)$ using mini-batch GD (without using in-built package). The mini-batch size should be taken as user-input during run-time.

(g) (4 points) Plot the mean-RSS during training phase with respect to epochs for all three variants of GD, separately (considering best train:test split in terms of test mean-RSS for demonstration).

(h) (3 points) Write your own results analysis in the word document from the results and plots/graphs.