भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 306: Machine Learning Lab
Practice Assignment 3

Instructions: This is only for practice. Complete it by 12:00 PM today. Your completion will be reviewed by the Teaching Assistants.

1. Download The salary dataset regarding the prediction of salary from the year of experience. Write a program to do the followings:
   (a) Read the dataset.
   (b) Randomly split the dataset in percentage of training samples (NTs): Percentage of test samples ($NTes$) ratio .
   Here, the values of NTs and $NTes$ are as follows ($NTs + NTes$ = 100% always):
   NTs = 10:10:90 (i.e., initial: 10%, increment by 10%, maximum is 90%)
   $NTes$ = 90:10:10 (i.e., initial: 90%, decrement by 10%, minimum is 10%)
   (c) Design different hypothesis ($\hat{y}$) to predict the salary from the year of experience using Gradient Descent (GD) method, **without using in-built Python packages /libraries** for estimation of the parameters, considering the generated training-testing splits from 1(b). Consider theta (θ) is initialized to zero and the learning rate (α) is set to 0.001. Separately save the model parameters for all the training splits.
   (d) Report the followings in form of a folder for plots/graphs (store directly from code), excel sheet/ CSV files for results (store directly from code), and the word document for result analysis/ interpretation:
   i. Store the plots (in the designated folder directly from code) of different estimated lines corresponding to different hypotheses, obtained in 1(c) over the various training splits.
   ii. Calculate and store (in the designated excel sheet for results directly from code) the prediction of salary for the test samples over different test splits considering the respective hypothesis.
   iii. Calculate and store (in the designated excel sheet for results directly from code) the coefficient of determination (R^2) and mean of sum of squared residuals (mean-RSS) for different training-testing splits separately.
   iv. Plot the obtained results of mean-RSS values in a graph between percentage of training samples in X-axis and values of mean-RSS for respective training and testing sets in Y-axis.
   v. Plot the obtained results of R^2 scores in a graph between percentage of training samples in X-axis and values of R^2 for respective training and testing sets in Y-axis.
   vi. Select the best training-testing split (in terms of (R^2) or mean-RSS). Using the selected split, calculate the model parameters and design hypotheses, considering various θ-initializations as zero, random values in the range [0,1] and [0,100]. Store the obtained model parameters, (R^2) and mean-RSS scores for all the three cases.

vii. Using the selected best training-testing split, calculate the model parameters and design hypotheses, considering various θ-initializations as zero, random values in the range [0,1] and [0,100], and learning rates (α) as 0.0001, 0.05, 0.1, 1, 10, 100, and 1000. Store the obtained model parameters, (R^2) and mean-RSS scores for all the possible combinations of θ-initializations and learning rates.
viii. Write your own results analysis in the word document from the results and plots/-graphs.

2. Execute the above assignment (in Q1) using in-built package/library for Gradient Descent.

3. Compare the various results obtained in Q1 (from i-v), with the ones obtained using OLS and interpret your analysis.