भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

CS 306: Machine Learning Lab

Practice Assignment 2

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

1. Please download The salary dataset regarding the prediction of salary from the year of experience. Write a program to do the followings without using in-built package for OLS:
   (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 ordinary least square method (OLS) for estimation of the parameters of simple linear regression model considering the training samples over the different training splits, separately. Save the model parameters for the the different training splits, separately.
   (d) Report the followings in form of a folder for plots/graphs (store directly from code), excel sheet for results (store directly from code), and the word document for result analysis:
   i. Store the plots (in the designated folder directly from code) of different lines corresponding to different hypothesis (having the samples also in the plots)
   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) coefficient of determination (R^2) and mean of sum of squared residual (RSS) for different train and test splits separately
   iv. Depict the result in a graph (percentage of training samples in x-axis and mean of sum of squared residual (RSS) in y-axis) to show percentage of training samples vs mean of sum of squared residual (RSS) over train and test samples. Store the graphs in the designated folder directly from code.
   v. Depict the result in a graph (percentage of training samples in x-axis and coefficient of determination in y-axis) to show percentage of training samples vs coefficient of determination over train and test samples separately. Store the graphs in the designated folder directly from code.
   vi. Calculate and analysis the value of Person correlation coefficient from slope parameters of different hypothesis over different training splits.
   vii. Write your own results analysis in the word document from the results and plots/-graphs (store in designated excel sheet and folders of plots/graphs).

2. Execute the above assignment (in question number 1) with inbuilt package for OLS