Statistical Machine Learning

Assignment: 4 **Max marks: 100**

Instructions:

1. You are free to use math libraries like Numpy, Pandas; and use Matplotlib, Seaborn library for plotting.
2. Add all the analysis related to the question in the written format, anything not in the report will not be marked.
3. Implement code that is Modular in nature and generalized to be executed for any input.
4. You can't use the inbuilt libraries to implement any algorithm except for calculation purposes like variance, Eigen decomposition, SVD, etc.
5. Code should be submitted in Python/Matlab file format only(.py/.m)

• Regression

1. Implement the normal equation (closed form) regression for the Boston housing dataset. The dataset description can be found here. The target feature is variable no. 14, 'MEDV', and the input variables are the remaining 13 variables.

a. Divide the dataset into training and testing using an 80:20 split ratio.
b. Perform Linear regression for all features and compute the RMSE for training as well as the testing set. (**Note:** There is no need to perform k-fold cross-validation for this part.)
c. Select the feature named 'LSTAT' for polynomial regression.
d. Perform k-fold cross-validation for k=5 on the training dataset. (**Note:** You can not use any inbuilt library to implement k-fold cross-validation.)
e. Perform step (d) for different degrees of polynomials using Polynomial Regression (Ex. For degree=1 perform 5-fold cross-validation, For degree=2, perform 5-fold cross-validation and so on.)

f. Use RMSE as an evaluation metric (**Note:** You can't use any inbuilt library for it). Compute mean RMSE of training and validation set separately from 5-fold cross-validation for each degree of the polynomial and plot it.
g. Choose the degree of a polynomial with the least mean validation RMSE and use that degree of polynomial to perform final regression on the whole training dataset (i.e., 80% dataset). State the RMSE of the test dataset (i.e., 20% dataset).

2. Use the following data that contains only 1 input feature and 1 target variable i.e X and Y. Consider the dataset as a whole i.e. Don't split it into train or test data.

a. Perform the steps (d), (e), (f) of Part-1.
b. Show the plots of line/curve fitted for the dataset using the different degrees of polynomials (degree). I.e, degree = [1,2,4,5,10,15,30]. Compute and state their RMSE also.

Note: Mention all the observations, results, analysis, visualizations and experiments in the report for each part (if any).