School of Computer Science Engineering and Technology

Course- BTech	Type- Core
Course Code- 301	Course Name-AIML
Year- 2022	Semester- Even
Date- 04-02-2022	Batch- IV Sem Spl

4 - Lab Assignment # No. (4.1)

Objective: The task is to compare the linear regression and polynomial regression with different degrees based on regression performance metrics.

1. **Dataset:** Download the dataset from the link https://tinyurl.com/braziltraffic. The dataset contains 17 features and 1 target (Slowness in traffic (%) which is last column) (5)
2. **Encoding:** Load the dataset into the code for pre-processing. 1st feature 'Hour' can be discretized into labels such as [morning, noon, afternoon, evening, night], which can be further codes using one-hot encoding, where morning can be represented as [0,0,0,0,1], noon can be [0,0,0,1,0] and so on. This results single feature "Hour" to be represented using five features of binary values. This now makes the dataset to have four extra columns. Choice of discretization is up to you. You can have like [day, night] also.
3. **Normalization:** Since the features are in different ranges, each column can be normalized into 0 to 1 using different methods such as scaling, standardizing etc. Note: Normalization should not be done for the target feature.
4. **Data Splitting:** After the range normalization, its time to split the data into training and testing. Dataset contain 135 entries (5 days data, each day 27 entries), so keep the last 27 rows of the original dataset (data of last 1 day) for testing, and rest of them for training.
5. **Regression Models:** Train different models for regression such as Linear Regression and Polynomial Regression use degree 2. [use sklearn.linear_model.LinearRegression for linear regression model and sklearn.preprocessing.PolynomialFeatures for polynomial regression]
6. **Testing:** Test the model with the test data and compute the mean squared error (MSE) for test data. Use different train-test split ratio: 70:30, 80:20, 90:10 and see the effect on the performance of the model in testing set.
7. **Regression Evaluation Metrics:** Evaluate the trained model using regression measures such as mean squared error, mean absolute error, median absolute error, R2 score.
8. **Playing with the Model:** You can try different strategies to see whether testing error comes down or not. Strategies can be different 1. Encoding of features, 3. Shuffling of training samples, 5. Degree of polynomials (such as 2, 3 and 4, print parameters of the polynomial models), Check the model error for the testing data for each setup.