Lab Assignment 6 – Evaluation Metrics

Build **Linear Regression, Polynomial Regression** using **Sklearn** for predicting Behaviour of the urban traffic of the city of Sao Paulo in Brazil. Apply different evaluation metrics for regression.

Steps

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).

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. This part you have to explore. 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:** This time we use **Sklearn** to train different models for regression such as Linear Regression and Polynomial Regression.

6. **Testing:** Test the model with the test data and compute the mean squared error (MSE) for test data. Explore different preprocessing strategies on how to improve 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, mean squared logarithmic error, max error, explained variance score, R2 score. Visualize the regression metrics as graphs.

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, 2. removal of some features, 3. normalization methods, 4. Shuffling of training samples, 5. Degree of polynomials (such as 2, 3, 4 etc print parameters of the polynomial models), Check the model error for the testing data for each setup.

9. **Classification Metrics:** Pick a Classification dataset of your choice and train the Perceptron and Logistic Regression model using Sklearn. Apply classification metrics such as confusion matrix, precision, recall, f-measure etc. Visualize the classification metrics as graphs.

Suggested Packages: Numpy, Pandas, Sklearn