PH3022- Machine Learning and Neural Computation

Assignments

1. Explain the following terms regarding the datasets.
    a. Training set
    b. Validation set
    c. Test set
2. Describe the following common problems found in machine learning models.
    a. Underfitting
    b. Overfitting
3. Data splitting
    a. Present the importance of randomly splitting data.
    b. Explain how to use the train_test_split() function in sklearn.model_selection to split the given data.
    c. Explain the purpose of the test_size parameter in the above function.
    d. Please provide an example to illustrate your explanation.
4. Boston Housing Dataset
    a. Write a description about the Boston Housing dataset. Find the labels of each column of the dataset.
    b. Develop a machine learning model based on linear regression in Scikit-learn. Follow the following steps.
        i. Load the Boston Housing Dataset from sklearn.datasets
        ii. Extract the features (X) and target (y) values from the dataset.
        iii. Split the data into training (X_train, y_train) and testing (X_test, y_test) sets.
        iv. Develop a linear regression model.
        v. Train the linear regression model.
        vi. Predict the y (y_pred) values using the trained model for X_test.
        vii. Evaluate the model using mean squared error (MSE) and coefficient of determination (R2 score) after training.
        viii. Plot y_pred Vs y_test graph. Comment on the plot.
        ix. Calculate MSE for both training and testing sets. Based on the results, comment on underfitting or overfitting of the developed model.
    c. Change the test_size parameter in train_test_split() function from 0.1 to 0.9 with 0.1 step size.

i. Calculate MSE and R2 score for both training and testing sets for each test_size value.

ii. Plot MSE for both training and testing sets against test_size parameter.

test_size	MSE (Train)	MSE (Test)
0.1		
0.2		
0.3		
0.4		
0.5		
0.6		
0.8		
0.9		

iii. Plot R2 score for both training and testing sets against test_size parameter.

test_size	R2 Score (Train)	R2 Score (Test)
0.1		
0.2		
0.3		
0.4		
0.5		
0.6		
0.8		
0.9		

iv. Comment on underfitting or overfitting of the developed model when test_size alters.

(Note: The results and conclusions may be only valid for the above dataset and the linear regression model. You cannot generalize the results for all the machine learning models.)