School of Computer Science Engineering and Technology

Course- BTech
Course Code- CSEL301
Year- 2022
Date- 05-09-2022

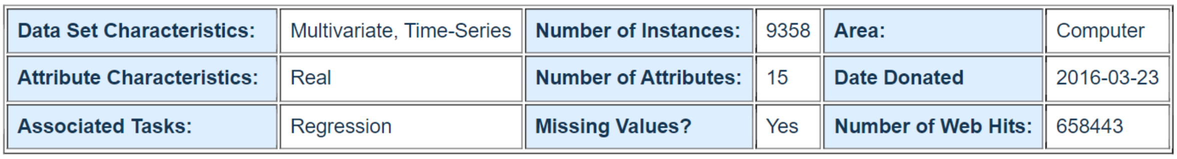
Type- Core
Course Name-AIML
Semester- Odd
Batch- 5th Sem

Lab Assignment No. 3.1_2

Exp. No.	Name	CO-1	CO-2	CO-3
3.1_2	Simple Linear regression	✓	✓	

Objective: To implement Simple Linear regression model using scikit-learn library.

About Dataset: The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensory Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidence of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.



Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	9358	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	15	**Date Donated**	2016-03-23
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	658443

.**Download** the dataset 'Student Performance' provided by UCI Machine Learning repository. Dataset link https://archive.ics.uci.edu/ml/datasets/Air+Quality

1. Load dataset into Pandas Data Frame (5)
2. Display the entries in data (5)
3. Check the shape (5)
4. Fetch the column name (5)
5. Check whether data contains missing value or not. if require, pre-process the data. (15)
6. Read and store the features “RH” in X and output variable “AH” in Y. (5)
7. Split the dataset into train and test in the following ratio (Hint: Use train_test_split class) (10)
   a) 70:30
   b) 80:20
8. Create Linear Regression Models on the splitting criterion as mentioned above (Hint: Use sklearn.linear_model.LinearRegression class) (10)

9. Find out the linear regression coefficients (i.e., m and c) (10)
10. Write the equation of SLR (10)
11. Perform the prediction on the test dataset. (10)
12. Check the performance of the model on test dataset by Calculating the ‘Mean Squared Error' (MSE) and R2-Score (Hint: sklearn.metrics.mean_squared_error function) (10)
13. Plot the regression line for test dataset (i.e., Y_pred vs Y_actual) (10)
    (Hint: Use scatter plot and line plot of Matplotlib Library)
14. Train the model against different any other 3 independent variable features one by one vs Dependent variable (i.e., Single independent variable vs dependent variable) and identify the most desirable feature for the dependent variable Y. (20)

Suggested Platform: Jupyter Notebook/Google Colab Notebook
Packages: numPy, Pandas, sklearn, matplotlib.pyplot