Data Science Fundamentals with Python

Worksheet 2

June 18, 2025

We hope you enjoyed learning about regression and time series analysis. Let's try out the following questions:

1 Question A

Here is the link to a dataset containing information about 10000 students and the factors affecting their academic performance. These factors are:

- Hours studied
- Previous scores
- Extracurricular Activities
- Sleep Hours
- Sample Question Papers Practiced

Your task is to build and compare two Multiple Linear Regression models to predict the target variable: Performance Index. Ensure to split your dataset into train, test and split datasets.

Regression Model 1: From Scratch Implementation

- Implement the multiple linear regression models only using NumPy (for computations) and Pandas (for representation of data).
- As part of data pre-processing:
  - Apply Z-score standardization to all numerical features
  - Apply the appropriate encoding for categorical features

Regression Model 2: Scikit-learn Implementation

- Build a multiple linear regression model using scikit-learn's Linear Regression class.
- Perform the same data pre-processing steps (Z-score standardization and encoding)

Evaluation and Visualization

- Compare the predictions of both the models on the test set. Report and compare the Mean Squared Error (MSE) and R^2 score for both the models.
- Plot the actual vs predicted values for both models on the same graph.

Submission requirements: Create a single Jupyter notebook file containing both the models, as well as the evaluation/visualization code cells. Try to demarcate different sections of the code using text blocks.

2 Question B

1. What is the difference between ACF and PACF?

2. Describe how you would identify the optimal order of an AR model using ACF and PACF plots.

3. Load and visualize the time series data about air travel passengers from the following link.

- Plot the ACF and the PACF to select the optimal lag.
- Split the data such that the train dataset consists of the year 1949 till 1959 and the test dataset consists of the year 1960.
- Compare your prediction with real data, and thus evaluate and plot your results.

Submission requirements: Create a single Jupyter notebook file. Answer the first two questions in text blocks within the notebook, and then attempt the third part.

Create a zip file containing the Jupyter notebooks of both Question A and Question B. That will your final submission.

Hint: Try out creating an account on Kaggle and coding a Jupyter notebook on the website itself. It provides fewer rate limits on GPUs and TPUs, and it also makes accessing the datasets given in the above questions easier.