

Lab 02:

Supervised Learning- Linear Regression

Lab Objectives:

- Understand the basic concepts of supervised learning.
- Learn the principles of linear regression.
- Implement linear regression using a machine learning library.
- Evaluate and interpret the results of linear regression.

Learning Outcomes:

Upon completing this lab, students should be able to:

- Explain the principles of supervised learning.
- Understand the basic concepts of linear regression.
- Implement linear regression using a machine learning library (e.g., scikit-learn).
- Interpret the results of a linear regression model.
- Evaluate the performance of a linear regression model.

Supervised Learning:

Supervised learning is a type of machine learning where the algorithm learns from a labeled dataset, which means the input data is paired with the corresponding output.

The goal is to learn a mapping function that can predict the output for unseen data.

Linear Regression:

Linear regression is a simple and widely used statistical technique for predicting a continuous variable. The model assumes a linear relationship between the independent variable(s) and the dependent variable.

The equation for a simple linear regression is:

$$y=mx+b,$$

where y is the dependent variable, x is the independent variable, m is the slope, and b is the intercept.

Key Concepts:

Regression Line: The line that best fits the data points in a scatter plot.

Residuals: The differences between the observed and predicted values.

Mean Squared Error (MSE): A measure of the average squared difference between observed and predicted values.

Implementation with scikit-learn:

The scikit-learn library provides a simple and efficient tool for data analysis and modeling, including linear regression.

The Linear Regression class in scikit-learn is used for linear regression modeling.

Problem-Based Scenario:

A real estate agent wants help to predict the house price for regions in the USA. He gave you the dataset to work on and you decided to use Linear Regression Model. Create a model which will help him to estimate what the house would sell for.

Dataset contains 7 columns and 5000 rows with CSV extension. The data contains the following columns:

- 'Avg. Area Income': Avg. Income of householder of the city house is located in.
- 'Avg. Area House Age': Avg. Age of Houses in same city.
- 'Avg. Area Number of Rooms': Avg. Number of Rooms for Houses in same city.
- 'Avg. Area Number of Bedrooms': Avg. Number of Bedrooms for Houses in same city.

- 'Area Population': Population of city.
- 'Price': Price that the house sold at.
- 'Address': Address of the houses

Lab Tasks:

On the given dataset, do followings:

- Data Exploration and Data Preprocessing if necessary:
- Linear Regression Modeling:
- Feature Importance Analysis:
- Model Interpretation:
- Interpret the results of the linear regression model.
- Provide insights into which features are the most significant predictors of house prices.
- Evaluate the model's performance on the test set and print out the classification report:
- Do hyperparameter tuning to enhance the model's accuracy.