DS 5220: Supervised Machine Learning and Learning Theory
Summer, 2024

Dr. Fatema Nafa
In Class Work 3

Deadline: 05/ 24/ 2024. 11:59 PM

Learning Objectives:

- **Understand Linear Regression:** Students will learn how to implement a linear regression model in a statistical software or programming environment.
- **Data Preprocessing:** Handling and preprocessing data for regression analysis.
- **Interpretation of Results:** Students will learn to interpret the output of regression models, understanding concepts like the regression coefficient and intercept, and their significance in the context of the data.
- **Forecasting:** Apply the model to predict future values and understand the limitations and assumptions of the model used.

Scenario: Predicting Salary Based on Various Factors

Dataset

The dataset contains information about employees in a company, including their education, years of experience, job level, department, and salary. **The goal is to predict the salary based on the other factors.**

Steps for Predicting Salary:

1. **Reading the Data**
2. **Data Preprocessing:**

   Convert categorical variables (Education Level, Job Level, Department) into numerical values using one-hot encoding or label encoding.

   Normalize or scale the numerical features (Years of Experience) if necessary.

3. **Splitting the Data:**

   Split the dataset into training and testing sets, typically with a ratio of 80:20.

4. **Building the Model:**

   Use the training data to build a multiple linear regression model where the target variable is Salary, and the predictor variables are Education Level, Years of Experience, Job Level, and Department.

5. **Training the Model:**

   Train the model using the training dataset.

6. **Evaluating the Model:**

   Evaluate the model performance using the testing dataset by calculating metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

7. **Making Predictions:**

Use the trained model to predict the salary of employees in the testing dataset.

8. **implement and compare multiple multivariable regression models using Python to determine which model performs best on the given dataset. The models to be compared are Linear Regression, Ridge Regression, and Lasso Regression.**