Linear Regression

Scenario-Based Question: Delivery Time Estimation

A logistics company wants to improve its delivery service by predicting delivery time (in hours) based on the distance (in kilometers) between the warehouse and the customer's location. This can help them optimize routes and better inform customers about delivery times.

The company's data team collects data from past deliveries, including the distance and actual delivery time, and decides to build a univariate linear regression model trained with gradient descent.

Based on this scenario, answer the following questions:

1. What type of supervised machine learning problem is this? Regression or classification? Justify your answer.

2. Identify the input feature and the target label in this context.

3. Write the hypothesis function h(x) for this model. Explain the meaning of θ_0 and θ_1 in this case.

4. Define the cost function used in linear regression. Why is it important?

5. Explain how the optimizer (gradient descent) updates the model parameters. Include the mathematical update rule.

6. What challenges might arise if the data contains outliers (e.g., a few deliveries with unusually long times)?

7. Assume the following data for 5 deliveries:

Distance (km)	Time (hrs)
5	1.2
10	2.3
15	3.1
20	4.2
25	5.0

8. Perform two iteration of gradient descent with an initial $\theta_0 = 0$, $\theta_1 = 0$ and learning rate $\alpha = 0.01$.

Predicting Salary using Linear Regression

Dataset Link:

Salary_Data.csv

Scenario:

You have recently joined a consulting firm as a junior data analyst. One of your clients is a recruitment company that wants to predict employee salaries based on their years of experience. They have shared a dataset containing historical records of their employees, including their years of experience and current salaries.

Your Task:

1. Download and load the dataset in Python using pandas.

2. Perform Exploratory Data Analysis (EDA):

    o   Show dataset info, summary statistics.

    o   Create scatter plots to visualize the relationship.

3. Build a Linear Regression model using Gradient Descent (not using scikit-learn's .fit() method).

4. Use Mean Squared Error (MSE) to evaluate your model.

5. Plot the regression line over the data points.

6. Interpret the coefficients (intercept and slope) in terms of salary prediction.

7. Predict the salary for a candidate with 6.5 years of experience.

8. Try to build the same model using scikit-learn's LinearRegression and compare results. Experiment with different learning rates in Gradient Descent.