Linear Regression

Key Points

- **Definition**: A statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.
- **Purpose**: To predict the value of the dependent variable (target) based on the values of independent variables (features).
- **Equation**: For a simple linear regression with one feature, the model is represented as: $$y = \beta_0 + \beta_1 x + \epsilon$$ Where:
  - y = dependent variable
  - β_0 = intercept
  - β_1 = slope coefficient
  - x = independent variable
  - ϵ = error term
- **Assumptions**:
  - Linearity: The relationship between dependent and independent variables is linear.
  - Independence: Observations are independent of each other.
  - Homoscedasticity: Constant variance of the error terms.
  - Normality: The residuals (errors) are normally distributed.
- **Evaluation Metrics**:
  - **Mean Absolute Error (MAE)**: Average of absolute differences between predicted and actual values.
  - **Mean Squared Error (MSE)**: Average of squared differences between predicted and actual values.
  - **Root Mean Squared Error (RMSE)**: Square root of the MSE.
  - **R-squared**: Proportion of variance in the dependent variable that is predictable from the independent variable(s).

Algorithm

1. **Initialize Parameters**:
   - Start with initial estimates for the intercept (β_0) and slope coefficients (β_1).
2. **Calculate Predictions**:
   - For each data point, compute the predicted value using the linear equation: $$\hat{y} = \beta_0 + \beta_1 x$$
3. **Compute Error**:
   - Calculate the error (residuals) for each prediction: $$\text{Error} = y - \hat{y}$$
4. **Optimize Parameters**:
   - Use optimization techniques (e.g., Gradient Descent) to minimize the cost function, which is typically the Mean Squared Error (MSE): $$\text{Cost Function} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
   - Update the parameters (β_0 and β_1) iteratively to reduce the cost function.
5. **Repeat**:
   - Continue updating the parameters until convergence is reached (i.e., when changes in the cost function become minimal).

6. **Predict**:
    - Use the optimized parameters to make predictions on new data: $$\hat{y}_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}}$$
7. **Evaluate**:
    - Assess the model's performance using evaluation metrics (MAE, MSE, RMSE, R-squared) to determine how well the model fits the data.