Notes

Linear Regression

* Imagine you are a shop owner trying to predict sales based on advertising budget.
* If you spend more on ads, you expect sales to go up. But how much will sales increase for every extra dollar spent?
* Linear Regression is a method used in machine learning to model the relationship between two variables by fitting a straight line through the data.
* Equation of a Straight Line:
  + Y = mx + b
  + Y = Dependent Variable (What you want to predict, e.g., Sales)
  + X = Independent Variable (What you control, e.g., Ad Budget)
  + M = Slope (How much Y changes for each unit of X)
  + B = Intercept (Y’s value when X = 0)
* Types of Linear Regression:
  + Simple Linear Regression (One Independent Variable)
  + Example: Predicting Sales based on Ad Spending
    - X: Ad Budget
    - Y: Sales Revenue
  + Multiple Linear Regression (More than One Independent Variable)
  + Example: Predicting house prices based on square footage, number of bedrooms, and location.
    - X1: House Size (sq ft)
    - X2: Number of Bedrooms
    - X3: Location Score
    - Y: House Price
* How does Linear Regression Work?
  + Linear Regression finds the best-fitting line by minimizing errors using Ordinary Least Squares (OLS).
    - Residuals (Errors) 🡪 The difference between actual and predicted values.
    - OLS minimize the sum of squared residuals to find the best slope m and intercept b.
* Assumptions of Linear Regression:
  + Linearity: The relationship between X and Y must be linear.
  + Independence: Data points should not be correlated
  + Homoscedasticity: Errors should have constant variance
  + Normality: Residuals should be normally distributed
* How do we Evaluate a Linear Regression Model?
  + Metrics to check Model Performance:
    - R-Squared: Measure how well the model explains the data (0 to 1, where 1 is perfect)
    - Mean Squared Error (MSE): Measures Average Error
    - Root Mean Squared Error (RMSE): Measures how far predictions are from actual values.
* Where is Linear Regression Used?
  + Stock Market Predictions 🡪 Predict future stock prices
  + Real Estate 🡪 Estimate house prices based on features
  + Business Analytics 🡪 Forecast sales and revenue
  + Healthcare 🡪 Predict disease risk based on patient data

Linear Regression finds relationships between variables.

It fits a straight line to minimize prediction errors.

Metrics

* R-Squared – Goodness of Fit
  + Measures how well the model explains the variability of the data.
  + Interpretation:
    - R-square = 1 🡪 Perfect model (all points fit exactly)
    - R-square close to 1 🡪 Good Model
    - R-square close to 0 🡪 Poor Model
    - R-square < 0 🡪 Worse than just predicting the mean
* Mean Absolute Error (MAE)
  + Measures the average absolute difference between actual and predicted values.
    - Example: If MAE = $5000, it means the model’s predictions are off by $5000 on average.
    - Easy to interpret.
    - Doesn’t penalize large errors heavily.
* Mean Squared Error (MSE)
  + Measures the average squared difference between actual and predicted values.
  + Why Square the Errors?
    - Punishes larger errors more (useful applications like stock market predictions).
    - However, it’s hard to interpret since it’s squared units (e.g., predicting house price in dollars but MSE is in dollars-square).
    - Penalizes large errors
    - Harder to impact
* Root Mean Squared Error (RMSE)
  + Same as MSE but takes the square root to get back to original units.
  + Example:
    - If RMSE = $4500, it means, on average, the model’s predictions are $4500 off from actual values.
    - More interpretable than MSE.
    - Still penalizes large errors.
* Adjusted R-Squared
  + Adding more features always R square, even if they don’t help.
  + Adjusted R Square fixes this by penalizing unnecessary features.

|  |  |  |
| --- | --- | --- |
| Metric | Meaning | Best For |
| R-Square | How well the model explains Y | General Model Evaluation |
| Adjusted R-Square | Fixes R-Square for extra variables | Feature selection & comparison |
| MAE | Average absolute error | Easy interpretation |
| MSE | Penalizes large errors more | Cases where big errors matter |
| RMSE | Similar to MSE but interprets | Most common choice |

Example:

A financial analyst wants to predict the closing price of a stock based on its opening price. The goal is to use historical stock market data and apply Linear Regression to find the relationship between Opening Price (X) and Closing Price (Y).

Objective

* Build a Linear Regression model that predicts the closing price of a stock given its opening price, using real or simulated stock market data.

Solution Approach

* We will use historical stock price data with columns:
  + Opening Price (X): The price at which the stock started trading in a day.
  + Closing Price (Y): The price at which the stock closed at the end of the day.
* Train a Linear Regression Model
  + Split the data into training (80%) and testing (20%)
  + Train the model using Ordinary Least Squares (OLS) to find the best-fit line.
  + Evaluate the model using metrics like R-Square, MSE, RMSE, and MAE.
* Interpret the Results
  + If R-Square is high (Close to 1) 🡪 The model can predict stock closing prices accurately.
  + If MAE and RMSE are low 🡪 Predictions are closer to actual values.
  + Regression coefficients will tell us how much the closing price is expected to change per unit increase in the opening price.