Linear regression:

Linear regression is a **supervised machine learning** algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). The relationship between the variables is assumed to be linear, meaning that a change in the value of the predictor variable is associated with a constant change in the outcome variable.

Linear regression is widely used for tasks such as predicting house prices, stock prices, temperature, and many other real-world applications where a linear relationship between variables is reasonable.

The Best Fit Line / Regression Line

The "best fit line" in linear regression refers to the line that minimizes the sum of the squared differences between the predicted and actual values of the dependent variable. This line is also known as the "regression line" or "line of best fit." The process of finding this line involves determining the optimal values for the slope (m) and the y-intercept (b) in the linear equation $y=mx + b$

Here,

- y is the dependent variable (the variable we are trying to predict),
- x is the independent variable (the variable used for prediction),
- m is the slope of the line (representing the change in y for a unit change in x),
- b is the y-intercept (the value of y when x is zero).

Types of Linear Regressions:

1. Simple linear regression (with one predictor variable)
2. Multiple linear regression (with multiple predictor variables).

Simple Linear Regression

Simple linear regression is a supervised machine learning algorithm used for predicting a continuous target variable based on a single predictor variable. The relationship between the predictor variable (X) and the target variable (y) is assumed to be linear, represented by the equation ($y = mx + b$), where (m) is the slope and (b) is the y-intercept.

Example: Consider a scenario where we want to predict the salary (y) of employees based on their years of experience (X). The simple linear regression model will try to find the best-fit line that represents the linear relationship between experience and salary.

Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression that involves predicting a continuous target variable based on multiple predictor variables.

Example: Extending the employee salary prediction example, let's now consider multiple factors such as years of experience (x_1), education level (x_2), and age (x_3). The multiple linear regression model aims to find the best-fit hyperplane that represents the relationship between these factors and salary (y).

Cost Function:

The cost function measures the error between predicted and actual values and serves as the basis for optimizing the parameters of the linear regression model during the training process. The objective is to find the parameters that minimize this cost, resulting in a more accurate model.

Cost Function Formula (Mean Squared Error - MSE):

For a simple linear regression model with one independent variable, the cost function (MSE) is defined as:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

For multiple linear regression with n independent variables:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Here:

- $J(\theta)$ is the cost function.
- m is the number of training examples.
- $h_\theta(x^{(i)})$ is the predicted value for the i-th example.
- $y^{(i)}$ is the actual target value for the i-th example.

Source: ChatGPT

Gradient Descent

Gradient Descent is an optimization algorithm used to minimize the cost function in the context of linear regression (and other machine learning models). The primary goal is to find the optimal values of the model parameters that minimize the difference between predicted and actual values.

Formula

Formula:

For linear regression, the parameters to be optimized are $\theta_0, \theta_1, \ldots, \theta_n$, and the update rule for each parameter is given by:

$$\theta_j := \theta_j - \alpha \frac{\partial J}{\partial \theta_j}$$

Where:

- θ_j is the j-th parameter.
- α is the learning rate (step size), a hyperparameter chosen before training.
- J is the cost function.

The partial derivative $\frac{\partial J}{\partial \theta_j}$ represents the rate of change of the cost function with respect to the j-th parameter.

Example

Example:

Consider the simple linear regression model $y = mx + c$. The cost function is the mean squared error (MSE):

$$J(m, c) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$$

Where:

- m is the number of training examples.
- $h(x^{(i)})$ is the predicted value for the i-th example.

The parameters to be optimized are m and c. The update rules for each parameter become:

$$m := m - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)}) x^{(i)}$$
$$c := c - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})$$

Here, the terms $(h(x^{(i)}) - y^{(i)}) x^{(i)}$ and $(h(x^{(i)}) - y^{(i)})$ are the partial derivatives of the MSE with respect to m and c, respectively.

Advantage Simple Linear Regression:

1. Simple Linear Regression provides a clear and interpretable relationship between the independent and dependent variables.
2. It is computationally less demanding compared to more complex models, making it suitable for quick analysis and interpretation of relationships in relatively simple datasets.
3. Simple Linear Regression allows for effective visualization of the relationship between variables, aiding in the identification of patterns and trends in the data.

Disadvantages Simple Linear Regression:

1. Assumes a linear relationship between variables, which may not hold true in more complex real-world scenarios, leading to inaccurate predictions.
2. Simple Linear Regression is sensitive to outliers, and a single extreme data point can significantly impact the model's performance.
3. Considers only one predictor variable, neglecting potential interactions and dependencies that may exist among multiple variables in the dataset.