Topic 3 - Regression

Linear Regression Simplilearn

Example: A Venture Capital firm is trying to understand which companies should they invest
We take the idea and then we:

- Decide companies to invest
- Predict the profit companies make
- Based on companies expenses

Independent Variable

A variable whose value **does not change** by the effect of other variables and is used to manipulate the dependent variable. It is often denoted as X

Example 2: Based on the amount of rainfall, how much would the crop yield?
In the example 2 is the **rainfall**

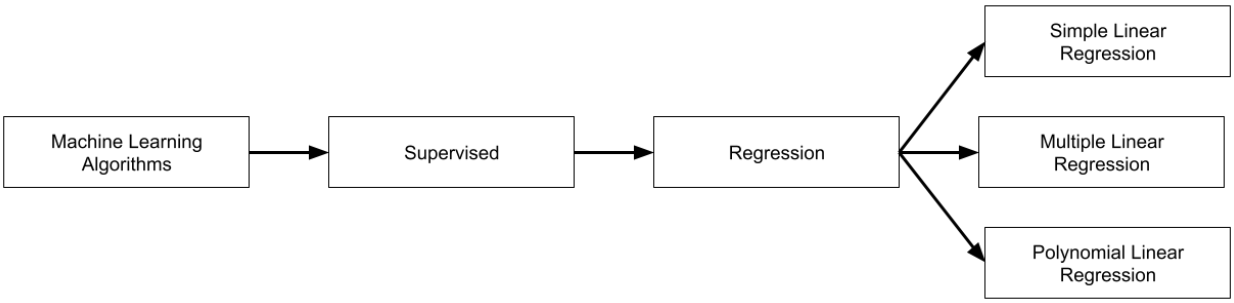
Dependent Variable

A variable whose **value changes** when there is any manipulation in the values of independent variables. It is often denoted as Y

In the example 2 is the **Crop yield**

Numerical and Categorical Values

Data	
Numerical	**Categorical**
Age	Color
Salary	Dog's Breed
Height	Gender



Understanding Linear Regression

- Linear Regression is a statistical model used to **predict the relationship** between independent and dependent variables by **examining two factors**
- First: Which **variables** in particular **are significant predictors** of the outcome variables
- How **significant is the Regression line** to make predictions with highest possible accuracy

Linear regression

- See if we can fit a line onto that data.
- $h_\theta(x)$ = hypothesis function
- θ_0 Is the intercept, which is also called the **bias term**.
- $\theta_1 x_1$The gradient determines the slope and that is θ_1

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

$$h_\theta = \sum_{j=0}^{1} \theta_1 x_1 (with\ x_0 = 1)$$

$$h_\theta = [1 \quad x] \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

Goal: Use data to learn hypothesis
Steps:

- Calculate the **L2 loss** or **Mean Squared Error**

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

Gradient Descent

- The gradient descent is used to optimize the line in linear regression or other machine learning algorithms
- The gradient descent algorithm applies derivatives in each of the the line's slopes to figure out if the gradient is decreasing or increasing

Data Scaling

- When you use a **multivariate** data, as you can see here, there are some **issues**, particularly **with gradient descent**
- Alters the perception of the graphics
- **Types:**
    - **Min-max normalization:** This is to force all of raw data to fall between zero and one

      $$x_j^s = \frac{x_j - min(x_j)}{max(x_j) - min(x_j)}$$
    - **Range normalization (centered on mean): Centering** the data **along the mean,** that **works very well** if your data tends to be **normally distributed**.

      $$x_j^s = \frac{x_j - mean(x_j)}{max(x_j) - min(x_j)}$$
    - **Standardization (z-score):** Subtract the mean. All the data is **centered on the mean**, but rather than between zero and one, **divide by standard deviation** of our data set

      $$x_j^s = \frac{x_j - mean(x_j)}{std(x_j)}$$

Polynomial regression

- Data can be quite complicated sometimes and a line doesn't always do it
- We can create a quadratic equation which may fit our data better.