Introduction to Machine Learning

Fundamentals and Applications

Understanding Machine Learning

Machine learning is a branch of artificial intelligence and computer science that uses **data** and **algorithms** to imitate how humans learn. The data it uses in its learning phase is called **training data**, and it is the guiding principle for the machine learning system.

Training Data

X	Y
2	4
4	8
7	14
10	20

Predictor/Feature/Attribute/ Independent Variable

Target/Label/Response/Outcome/Dependent Variable

In machine learning, the machine learns the algorithm that defines the relationship between the predictor and the target

Understanding Machine Learning Model

A machine learning model is an **algorithm/mathematical expression** that defines the relationship between a target variable and one or more predictor variables.

X (Predictor)	Y (Target)
2	4
4	8
7	14
10	20
20	40

Model

Relationship
Y=2*X

The coefficient of X is called the **parameter** of the model. During the training phase, the model learns this parameter using the input data.

Types of Machine Learning

Based on supervision:

- Supervised Learning
  - Training data contains both the predictor and the target
- Unsupervised Learning
  - Training data contains only the predictor
- Semi Supervised Learning
  - Combination of supervised learning and unsupervised learning
- Reinforcement Learning
  - Doesn’t require any training data. Learns by itself through reward and penalty Technique.

Based on the target variable:

- Regression
  - The target is a numerical variable
  - Predicts a numeric value such as salary,temperature.
- Classification
  - The target is a categorical variable
  - Predicts a class such as positive,negative.

Introduction to Regression Models

Definition: A regression model is a method used to define the **relationship** between one or more **independent variables (predictors)** and a **dependent variable (target)**.

Purpose: The model helps in understanding how the value of the **dependent variable** changes in response to changes in the **independent variable(s)**.

Nature of the Target Variable: In a regression model, the target variable is always a **numerical variable**. This distinguishes regression from classification, where the target is categorical.

When the relationship between the variables is **linear**, we call it a **linear regression model.**

A linear regression model can be broadly categorized into two types:

- **Simple Linear Regression Model:**
  In Simple Linear Regression, we try to find the relationship between a single independent variable and a corresponding dependent variable.
- **Multiple Linear Regression Model:**
  In Multiple Linear Regression, we try to find the relationship between 2 or more independent variables and the corresponding dependent variable.

Exploring the Tips Dataset

The dataset has 244 observations(rows) with 7 features(columns). The features are:

total_bill: This column represents the total amount of the bill, including the cost of the meal and any additional items like drinks or desserts.

sex: This column denotes the gender of the person who paid the bill. It could be 'Male' or 'Female'.

smoker: This column indicates whether the party was composed of smokers or non-smokers. It's represented by 'Yes' for smokers and 'No' for non-smokers.

day: This column specifies the day of the week when the meal took place. It could be 'Thur' for Thursday, 'Fri' for Friday, 'Sat' for Saturday, or 'Sun' for Sunday.

time: This column represents the time of day when the meal occurred. It could be 'Lunch' or 'Dinner'.

size: This column denotes the size of the dining party. It indicates the number of people in the group.

tip: This column indicates the amount of tip left by the customer. It's usually a percentage of the total bill but can vary based on factors like service quality, personal preference, etc.

Understanding Pearson's Correlation Coefficient

The correlation between two numerical variables, *x* and y, is denoted by *r*.

The following formula measures this correlation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r = correlation coefficient

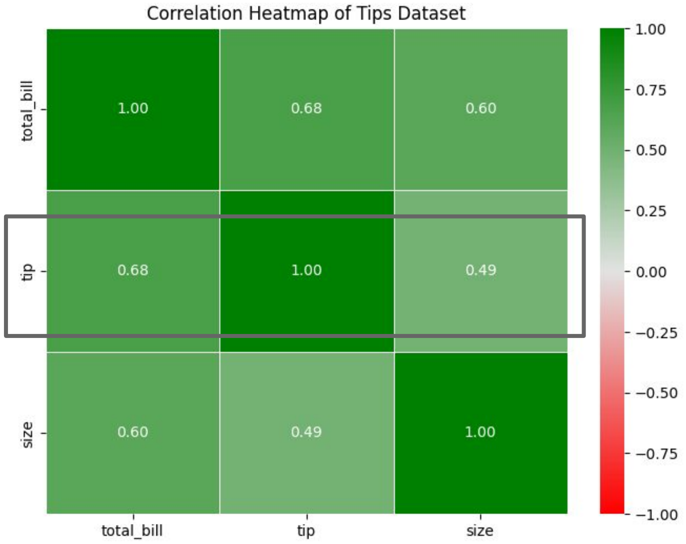
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

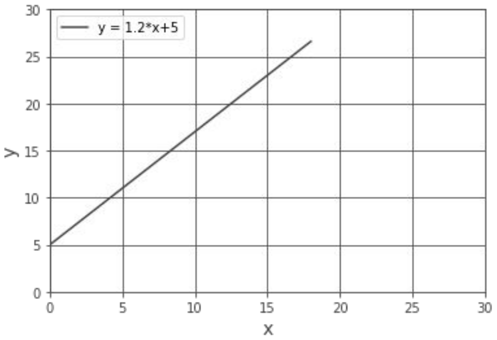
Pearson's Correlation Coefficient ranges from **-1 to 1**. A value of **-1** indicates a perfect negative linear relationship, **1** indicates a perfect positive linear relationship, and **0** indicates no linear relationship between the variables.



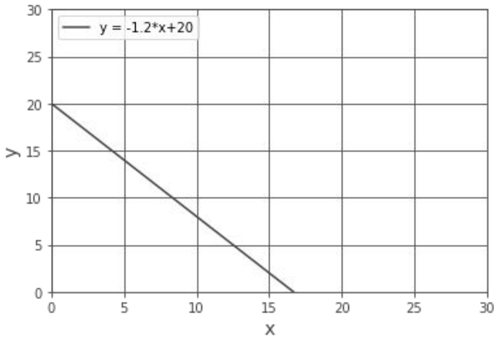
Each cell in the matrix represents the correlation between two variables.

The Equation of a Line

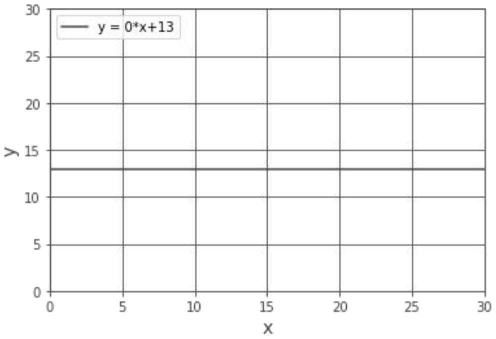
The equation of a line is typically written as **y = mx + b** where **m** is the slope and **b** is the y-intercept. The slope defines the direction and the steepness of the line where the y-intercept defines the expected value of y when x = 0.



The slope is **+1.2** and the y-intercept is **5**



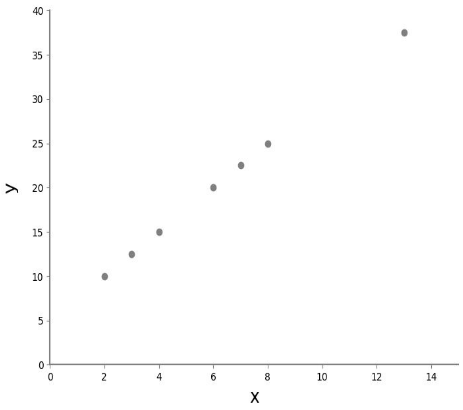
The slope is **-1.2** and the y-intercept is **20**

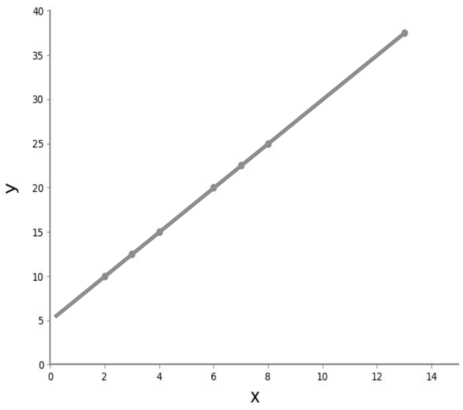


The slope is **0** and y-intercept is **13**

Linear Relationship in Data

X	Y
2	10
3	12.5
4	15
6	20
7	22.5
8	25
13	37.5

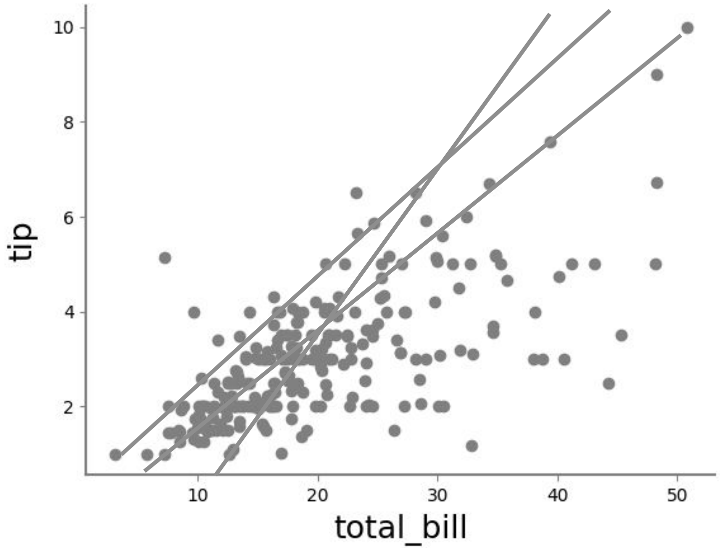




When data points follow a linear pattern, we can draw a line to express the relationship.

A simple linear regression model is a machine learning technique that fits a line to data points to describe their relationship. The intercept and slope of the line, known as the model parameters, are learned from the data.

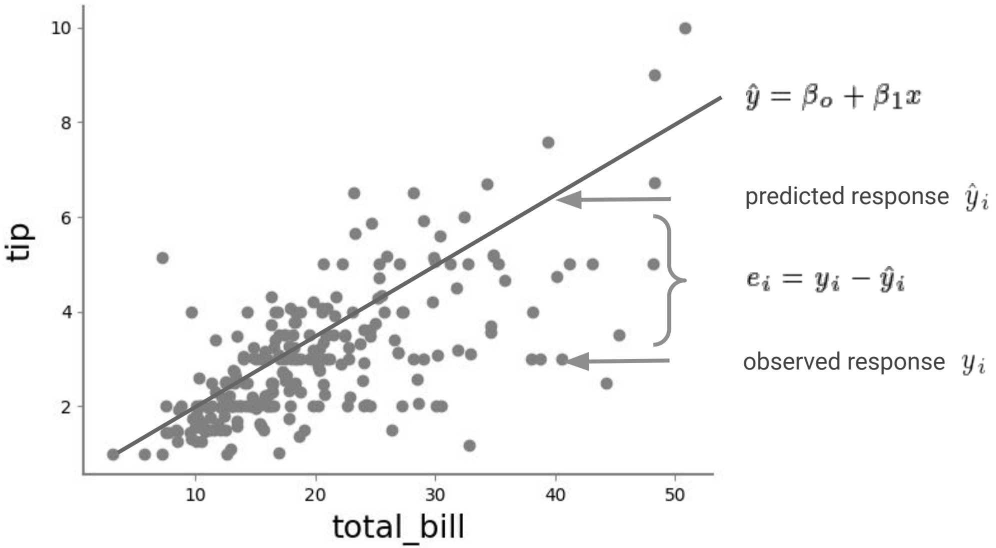
Fitting a Line to Real Data



No line can connect all the data points!

When it's impossible to fit a line perfectly to all data points, the goal is to estimate a line that provides the best possible fit.

Least Squares Method



Least Squares Method:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The least squares estimation is a technique that estimates parameters by minimizing the sum of squared errors (residuals), referred to as the cost function, loss function, or error function in machine learning.

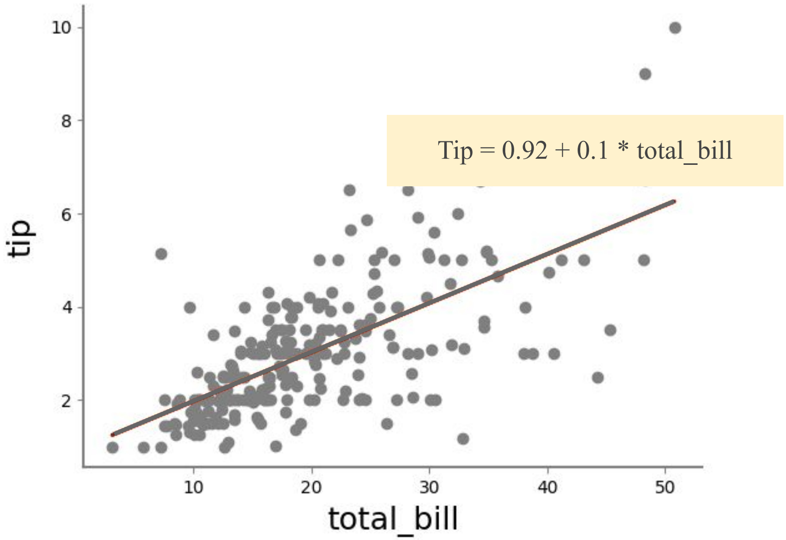
Algorithm for Finding the Best Fit Line

```python
import seaborn as sns
from sklearn.linear_model  import LinearRegression
# Load the tips dataset from seaborn
tips_df = sns.load_dataset( 'tips')

X = tips_df[[ 'total_bill' ]] #predictor
y = tips_df[ 'tip'] #target


# Creating and training the Linear Regression model
model = LinearRegression()
model.fit(X, y)

# Print the parameters
print(model.intercept_,model.coef_)
```




```python
# Making prediction
model.predict([[ 20]]) #[3.02]
model.predict([[ 22],[25]]) #[3.23,3.54]
```

Evaluating Model Performance

Evaluation metrics are measures or criteria used to assess the performance and effectiveness of a model or system. In the context of machine learning, these metrics quantify how well a model performs. The choice of evaluation metrics is crucial in determining the success of a model and its suitability for a given problem. Following are some common evaluation metrics for regression models:

- **Mean Absolute Error (MAE)**

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Measures the average magnitude of the errors.

- **Mean Square Error (MSE)**

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Measures the average of the squares of the errors.

- **Root Mean Square Error (RMSE)**

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Measures the standard deviation of the errors.

- **Coefficient of determination (R^2)**

$$R^2 = \frac{explained\ variation}{total\ variation}$$

The coefficient of determination, or R-squared, measures the amount of variation explained by the model and it ranges from 0 to 1, with 0 indicating no explanatory power and 1 signifying perfect prediction in a linear regression model.

Limitations of Linear Regression

- **Limited to Linear Relationships**

  Linear regression only looks at linear relationships between dependent and independent variables.
- **Sensitive to Outliers**

  Data outliers can damage the performance of a machine learning model drastically and can often lead to models with low accuracy.
- **Data Must Be Independent**

  Very often the inputs aren't independent of each other and hence any multicollinearity must be removed before applying linear regression.