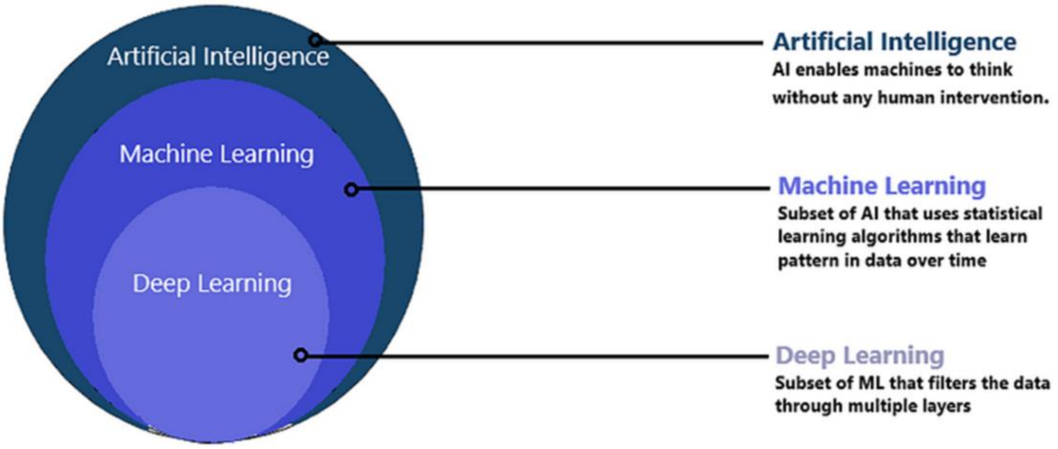
# Unit IV: Types of Machine Learning

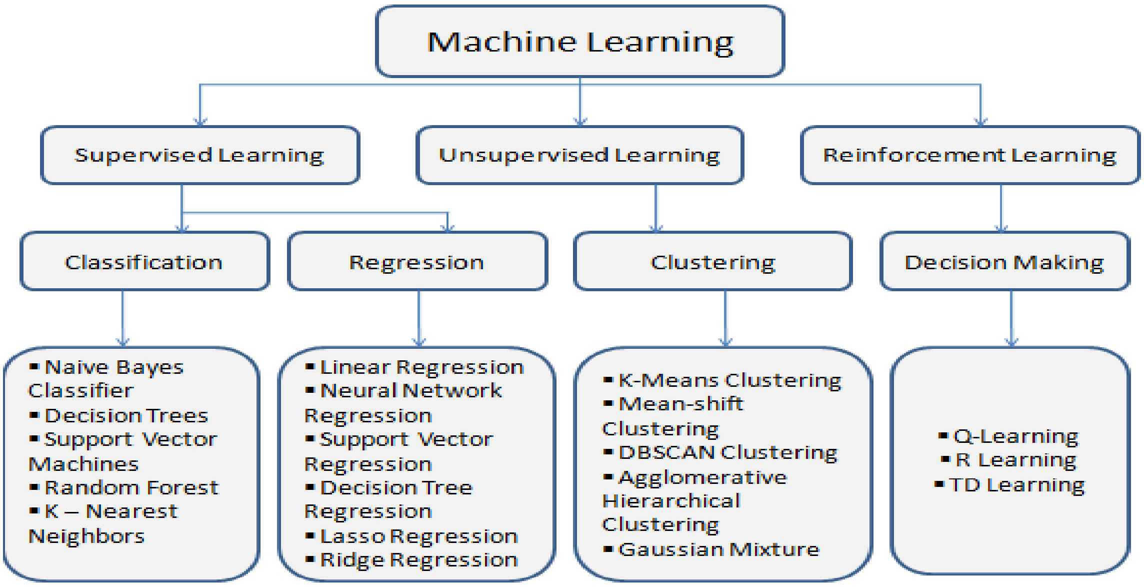
## Introduction to Machine Learning Techniques

Machine learning techniques can be broadly categorized based on how a model learns from data. Each category serves different types of tasks, data, and learning objectives. The fundamental goal of each type is to enable machines to identify patterns, make informed decisions, or improve performance over time based on experience or observed data.


Artificial Intelligence
Machine Learning
Deep Learning
Artificial Intelligence
AI enables machines to think without any human intervention.
Machine Learning
Subset of AI that uses statistical learning algorithms that learn pattern in data over time
Deep Learning
Subset of ML that filters the data through multiple layers

## Types of ML

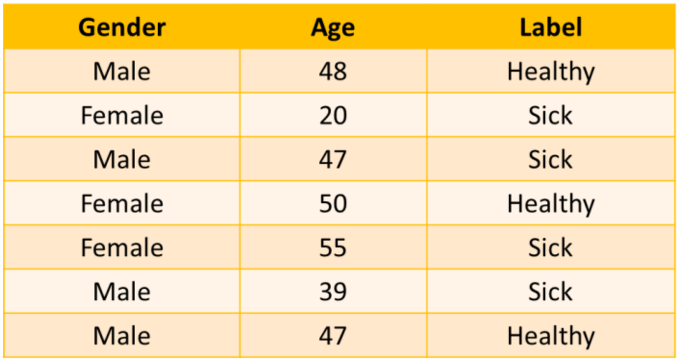
1. **Supervised learning:** training a model on labeled data
2. **Unsupervised learning:** training a model on unlabeled data
3. **Reinforcement learning:** training a model through trial and error
4. **Semi-supervised learning:** falls between unsupervised learning and supervised learning.



## Supervised Learning

In supervised learning, the model is trained using labeled data, meaning each input has an associated output. The model learns to map inputs to outputs by analyzing the training data, which includes both the features (input data) and labels (output data). Once trained, the model can predict outputs for new, unseen inputs.

Example — Consider the following data regarding patients entering a clinic . The data consists of the gender and age of the patients and each patient is labeled as “healthy” or “sick”.



| Gender | Age | Label |
|---|---|---|
| Male | 48 | Healthy |
| Female | 20 | Sick |
| Male | 47 | Sick |
| Female | 50 | Healthy |
| Female | 55 | Sick |
| Male | 39 | Sick |
| Male | 47 | Healthy |

## Types of Supervised Learning:

- **Regression**:
  - Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
  - It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.
  - In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data.

- **Types of Regression**:
  1. **Linear Regression**: Models a linear relationship between input and output variables. It’s simple and widely used for tasks where changes in the input result in proportional changes in the output.
  2. **Polynomial Regression**: Extends linear regression by modeling a non-linear relationship using polynomial terms.
  3. **Logistic Regression**: Used for binary classification tasks (despite the name, it’s a classification technique). It predicts probabilities for two possible outcomes, such as "yes" or "no".
  4. **Ridge and Lasso Regression**: These are regularized regression techniques that help prevent overfitting by adding penalties for larger coefficients in the model.

5. **Support Vector Regression (SVR)**: Uses support vector machines for regression tasks, particularly effective in high-dimensional spaces.

- **Regression Applications**:

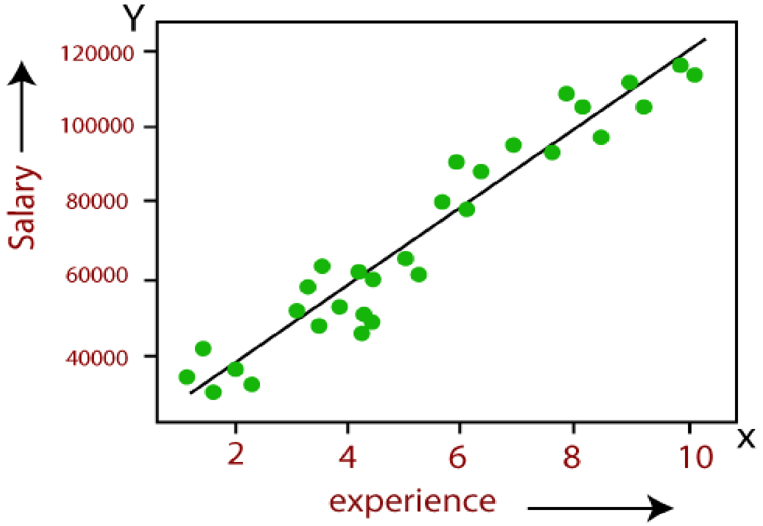
1. **Weather Forecasting**: Predicts temperature, humidity, and other weather variables based on historical weather data.
2. **Stock Market Prediction**: Forecasts stock prices or trends using past market data and economic indicators.
3. **Real Estate Valuation**: Determines house prices by analyzing features like location, size, and market conditions.

**Linear Regression:** Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.

The key point in Simple Linear Regression is that the *dependent variable must be a continuous/real value*. However, the independent variable can be measured on continuous or categorical values.

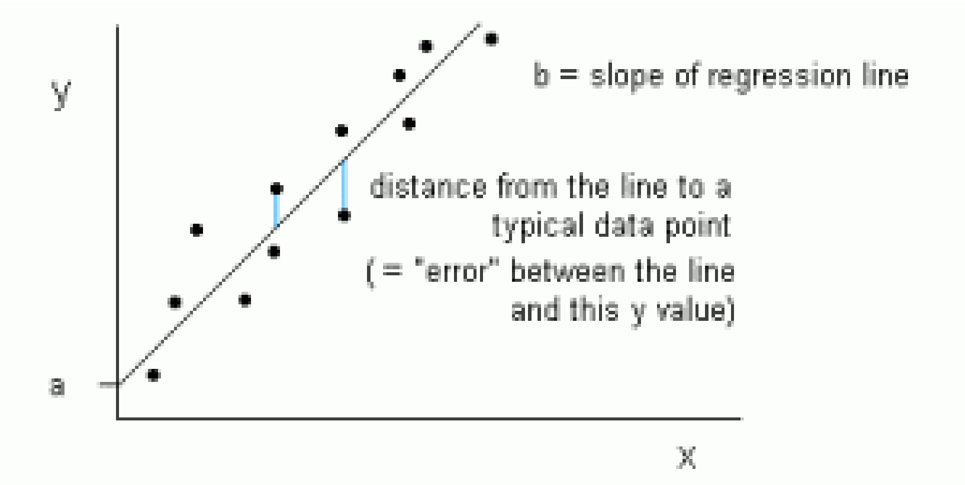
Simple Linear regression algorithm has mainly two objectives:

- **Model the relationship between the two variables.** Such as the relationship between Income and expenditure, experience and Salary, etc.
- **Forecasting new observations.** Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.



Mathematical equation for Linear regression: Y= a + bX

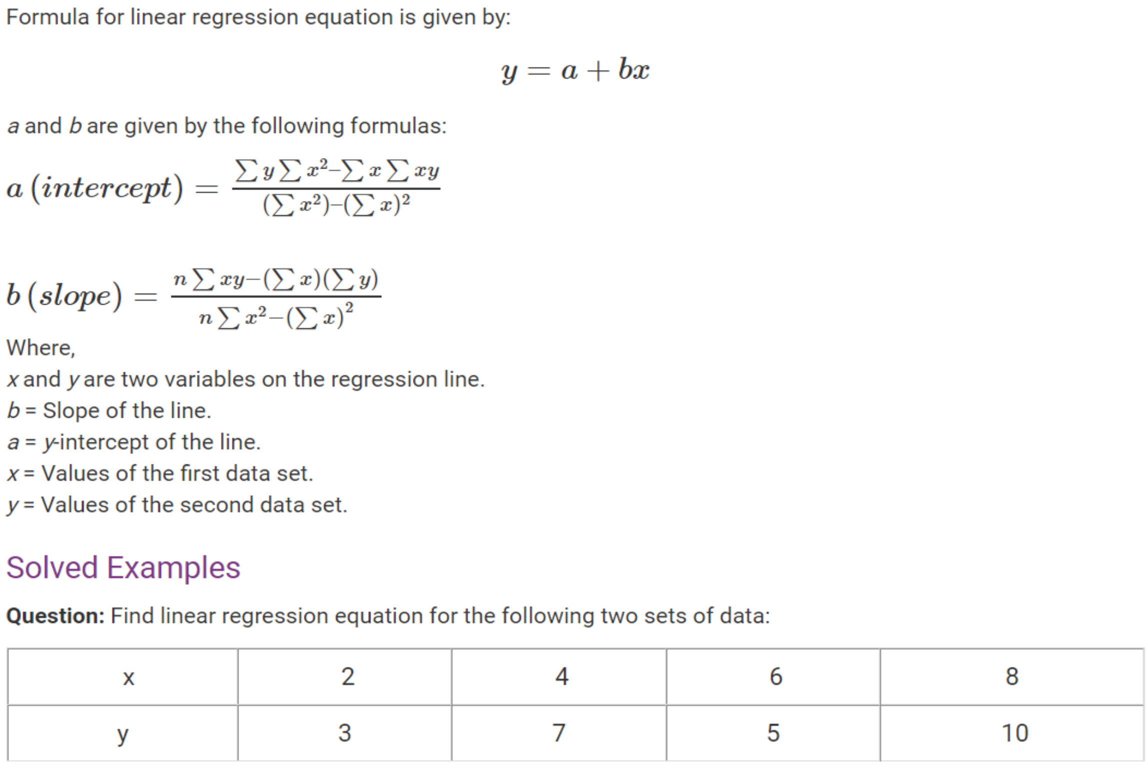
**Y = dependent variables (target variables),**
**X= Independent variables (predictor variables),**
**a and b are the linear coefficients**


y
b = slope of regression line
distance from the line to a typical data point (= "error" between the line and this y value)
a
x

Linear Regression model can be represented using the below equation: **Y= a + b X + ε**

Here,

- a = It is the intercept of the Regression line (can be obtained putting x=0)
- b = It is the slope of the regression line, which tells whether the line is increasing or decreasing.
- ε = The error term. (For a good model it will be negligible)



Formula for linear regression equation is given by:

$$y = a + bx$$

*a* and *b* are given by the following formulas:

$$a\,(intercept) = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Where,
*x* and *y* are two variables on the regression line.
*b* = Slope of the line.
*a* = *y*-intercept of the line.
*x* = Values of the first data set.
*y* = Values of the second data set.

## Solved Examples

**Question:** Find linear regression equation for the following two sets of data:

| x | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| y | 3 | 7 | 5 | 10 |

**Solution:**

Construct the following table:

| x | y | $x^2$ | xy |
|---|---|---|---|
| 2 | 3 | 4 | 6 |
| 4 | 7 | 16 | 28 |
| 6 | 5 | 36 | 30 |
| 8 | 10 | 64 | 80 |
| $\sum x$ = 20 | $\sum y$ = 25 | $\sum x^{2}$ = 120 | $\sum xy$ = 144 |

$b$ = $\frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^{2}-(\sum x)^{2}}$

$b$ = $\frac{4 \times 144 – 20 \times 25}{4 \times 120 – 400}$

b = 0.95

$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$

$a = \frac{25\times 120 - 20\times 144}{4(120) - 400}$

a = 1.5

Linear regression is given by:

y = a + bx

y = 1.5 + 0.95 x

## Multiple Linear Regression

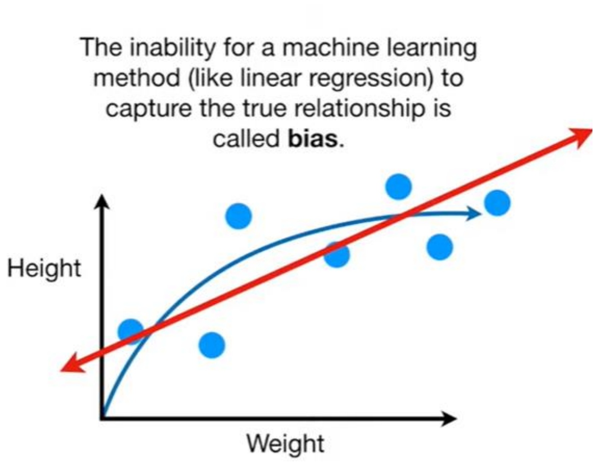
**Multiple linear regression** is used to estimate the relationship between **two or more independent variables** and **one dependent variable**. You can use multiple linear regression when you want to know:

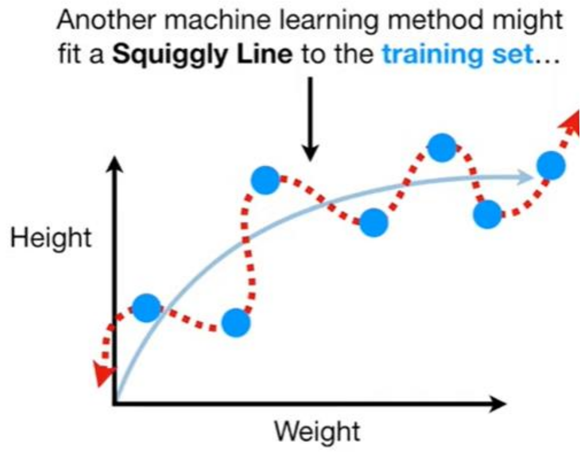
- How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
- The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).
- The Formula for Multiple regression is as follow:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

- **y** = the predicted value of the dependent variable
- **$B_0$** = the y-intercept (value of y when all other parameters are set to 0)
- **$B_1X_1$**= the regression coefficient ($B_1$) of the first independent variable (**$X_1$**) (a.k.a. the effect that increasing the value of the independent variable has on the predicted **y** value)
- **…** = do the same for however many independent variables you are testing
- **$B_nX_n$** = the regression coefficient of the last independent variable
- **e** = model error (a.k.a. how much variation there is in our estimate of **y**)

## Bias and Variance





**Bias**

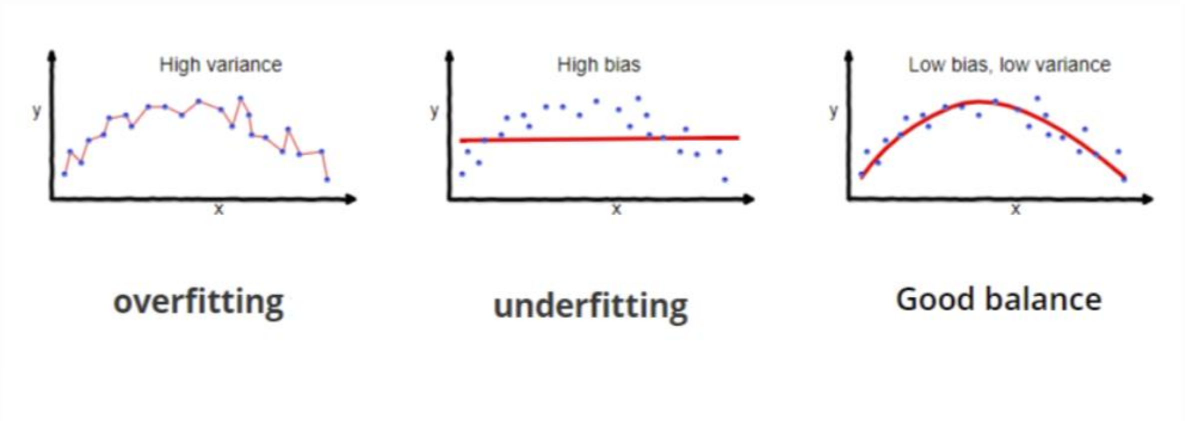
**Variance**

***Biases are the underlying assumptions that are made by data to simplify the target function.*** Bias does help us generalize the data better and make the model less sensitive to single data points. It also decreases the training time because of the decrease in complexity of target function High bias suggest that there is more assumption taken on target function. This leads to the **underfitting of the model** sometimes.

Examples of High bias Algorithms include Linear Regression, Logistic Regression etc.

***Variance refers to the degree to which a model is able to adapt to new data.*** Variance is a type of error that occurs due to a model's sensitivity to small fluctuations in the dataset. The high variance would cause an algorithm to model the outliers/noise in the training set. This is referred to as **overfitting**. In this situation, the model basically learns every data point and does not offer good prediction when it tested on a novel dataset.

Examples of High variance Algorithms include Decision Tree, KNN etc.



2. **Classification**:

Classification is a **supervised learning** technique where an algorithm learns to categorize data into predefined labels or classes. The goal of classification is to map input data (features) to a specific category or class based on patterns learned from labeled training data.

**How Classification Works**

1. **Input Data:** A dataset containing input features and corresponding class labels.
2. **Training Phase:** The algorithm learns patterns in the labeled data.
3. **Prediction Phase:** The model predicts the class label for new, unseen data.
4. **Evaluation:** The model's performance is assessed using metrics like accuracy, precision, recall, F1-score, and confusion matrix.

**Types of Classification**

1. **Binary Classification:**
   Classifies data into two categories.
   - Example: Spam email detection (Spam vs. Not Spam).
2. **Multi-class Classification:**
   Classifies data into more than two categories.
   - Example: Handwritten digit recognition (0–9 digits).
3. **Multi-label Classification:**
   Assigns multiple labels to a single data point.
   - Example: Tagging an image with labels like "beach," "sunset," and "vacation."
4. **Imbalanced Classification:**
   Handles datasets where one class significantly outnumbers the others.
   - Example: Fraud detection (fraud cases are rare).

**Common Classification Algorithms**

1. **Logistic Regression:** Predicts probabilities and maps them to class labels using a logistic function.
2. **Decision Trees:** Uses a tree-like model of decisions to classify data.
3. **Random Forest:** An ensemble method combining multiple decision trees for robust classification.
4. **Support Vector Machine (SVM):** Finds the hyperplane that best separates data points of different classes.
5. **K-Nearest Neighbors (KNN):** Classifies data points based on the majority class of their nearest neighbors.

6. **Naive Bayes:** A probabilistic classifier based on Bayes' theorem and assumes independence between features.
7. **Neural Networks:** Learns complex patterns in data using multiple layers of artificial neurons.

### Applications of Classification

1. **Medical Diagnosis:** Classifying diseases based on patient symptoms and test results.
2. **Email Spam Filtering:** Distinguishing spam emails from legitimate ones.
3. **Sentiment Analysis:** Classifying text as positive, negative, or neutral in reviews or social media.
4. **Credit Risk Analysis:** Predicting whether a loan applicant is likely to default.
5. **Image Recognition:** Identifying objects, animals, or scenes in images.
6. **Fraud Detection:** Detecting fraudulent transactions or activities.
7. **Speech Recognition:** Classifying spoken words into text categories.

### Advantages of Classification

1. Can handle a wide variety of tasks and datasets.
2. Provides interpretable results with some models (e.g., Decision Trees).
3. Highly adaptable to different problem domains.

### Limitations of Classification

1. Requires labeled data, which can be expensive and time-consuming to obtain.
2. Performance depends on data quality and feature selection.
3. May struggle with overlapping classes or imbalanced datasets.

## 3. Unsupervised Learning

In unsupervised learning, the model is provided with input data without any labels. The model’s task is to explore and identify patterns or structures within the data on its own. This type of learning is useful for discovering hidden structures and insights.

Example: Consider the following data regarding patients entering a clinic. The data consists of the gender and age of the patients.

| Gender | Age |
|---|---|
| Male | 48 |
| Female | 20 |
| Male | 47 |
| Female | 50 |
| Female | 55 |
| Male | 39 |
| Male | 47 |

**Clustering**: Clustering is an **unsupervised learning** technique in machine learning used to group data points into clusters based on their similarities. Unlike supervised learning, clustering works without labeled data, meaning the algorithm identifies patterns and relationships in the data without any predefined categories.

**How Clustering Works**

1. **Input Data:** A dataset with multiple data points.
2. **Similarity Measure:** The algorithm evaluates how similar or different data points are using a distance metric, such as Euclidean or Manhattan distance.
3. **Group Formation:** Data points that are similar are grouped into the same cluster, while dissimilar points are placed in separate clusters.

**Types of Clustering Algorithms**

1. **Partition-based Clustering**: Divides the data into distinct clusters.
    - Example: **K-Means**
        - Groups data into KKK clusters by minimizing the distance between data points and the cluster centroids.
2. **Hierarchical Clustering**: Builds a hierarchy of clusters, represented as a tree (dendrogram).
    - Example: **Agglomerative Clustering** (bottom-up), **Divisive Clustering** (top-down).
3. **Density-based Clustering**: Groups data points based on density regions in the data.
    - Example: **DBSCAN**
        - Identifies dense regions of data and marks outliers as noise.
4. **Model-based Clustering**: Assumes the data is generated from a mix of statistical distributions and finds the best fit.

- Example: **Gaussian Mixture Models (GMMs)**
  - Uses probabilities to assign data points to clusters.

**Applications of Clustering**

1. **Customer Segmentation**: Grouping customers based on purchasing behavior for targeted marketing.
2. **Image Segmentation:** Dividing an image into meaningful segments (e.g., medical imaging).
3. **Document Categorization:** Grouping similar documents together based on their content.
4. **Social Network Analysis:** Identifying communities within a network.
5. **Anomaly Detection:** Detecting outliers in data, such as fraud in transactions.
6. **Biology:** Classifying genes or organisms based on genetic information.

**Advantages of Clustering**

1. Handles unlabeled data effectively.
2. Helps discover hidden patterns in data.
3. Aids in feature engineering by identifying group-wise patterns.

**Limitations of Clustering**

1. Sensitive to the choice of distance metrics and the number of clusters (e.g., KKK in K-Means).
2. May struggle with overlapping clusters or noise in the data.
3. Computationally intensive for large datasets.

**4. Reinforcement Learning**

Reinforcement learning (RL) is a type of machine learning where an agent learns to make a sequence of decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties and adjusts its actions to maximize the cumulative reward over time. Unlike supervised learning, there are no labeled data points; instead, the agent explores to find the best actions.

Example: Consider teaching a dog a new trick: we cannot tell him what to do, what not to do, but we can reward/punish it if it does the right/wrong thing.

**Applications**:

1. **Robotics and Automation**: Trains robots to perform tasks, such as assembly line automation, where they learn the optimal sequence of movements.
2. **Gaming**: Used to develop AI players in games like Chess, Go, and video games, where the agent learns strategies to maximize its score or chances of winning.

3. **Self-Driving Cars**: Reinforcement learning is applied to make decisions for navigation, obstacle avoidance, and path optimization.

4. **Recommendation Systems**: Personalizes recommendations by learning user preferences and adapting over time, as seen in streaming services.

5. **Financial Portfolio Management**: Optimizes trading strategies to maximize returns based on past data and market conditions.

**Core Components**:

1. **Agent**: The decision-maker or learner.
2. **Environment**: Everything the agent interacts with.
3. **Actions**: The set of all possible moves the agent can make.
4. **Rewards**: Feedback received for each action, guiding the agent's learning.

**Techniques in RL**:

1. **Q-Learning**: A value-based approach where the agent learns the value of actions in specific states.

2. **Deep Q-Networks (DQN)**: Combines Q-learning with deep learning, often used in complex environments.

3. **Policy Gradient Methods**: Optimizes the agent's policy directly, useful for continuous action spaces.

### 5. Semi-Supervised Learning

Semi-supervised learning is a hybrid approach that leverages both labeled and unlabeled data. Typically, labeled data is limited and costly to obtain, while unlabeled data is abundant. Semi-supervised models aim to improve learning accuracy by utilizing the smaller labeled set to guide the interpretation of the larger unlabeled set.

**Applications**: Semi-supervised learning is ideal when only a small amount of labeled data is available, but there's an abundance of unlabeled data. This approach is commonly used in scenarios where data labeling is costly or time-consuming.

1. **Image and Video Classification**: Used in image recognition (e.g., tagging objects or people in photos), where only a few images are labeled, and the model learns to label the rest.
2. **Natural Language Processing (NLP)**: Improves performance in tasks like text classification or named entity recognition with limited labeled examples.
3. **Medical Imaging**: Helps classify medical images (e.g., X-rays or MRIs) for diagnosis by using a small labeled set of images along with a larger unlabeled set.
4. **Speech Recognition**: Enhances speech-to-text systems by training on a few labeled audio samples along with a large amount of unlabeled audio.
5. **Web Page Classification**: Assists in categorizing web pages (e.g., news, blogs, e-commerce) when only a limited number of labeled examples are available.

**Techniques in Semi-Supervised Learning**:

1. **Self-training**: The model initially trains on labeled data and then labels its own predictions on the unlabeled data to expand its training set.
2. **Co-training**: Uses multiple models trained on different features of the same data, each labeling the unlabeled data for the other model.
3. **Graph-based Methods**: Builds a graph of labeled and unlabeled data, where the relationships between data points are used to propagate labels to the unlabeled data.

### Types of Machine Learning Techniques and Applications

| Type of ML | Description | Key Techniques | Applications |
|---|---|---|---|
| **Supervised Learning** | - Learns a mapping function from labeled input data to outputs.<br>- Requires a dataset with input-output pairs.<br>- Focuses on predicting outcomes for new data based on historical data. | - Linear Regression<br>- Logistic Regression<br>- Decision Trees<br>- Support Vector Machines (SVM)<br>- Random Forest<br>- Neural Networks | - **Spam Detection**: Filtering emails into spam or not spam.<br>- **Credit Risk Analysis**: Predicting loan defaults.<br>- **Medical Diagnosis**: Disease classification.<br>- **Fraud Detection**: Identifying fraudulent transactions.<br>- **Stock Price Prediction**: Forecasting future stock prices. |
| **Unsupervised Learning** | - Identifies patterns or structures in data without labels.<br>- Groups similar data points or reduces dimensionality.<br>- No pre-defined output classes are required. | - K-Means Clustering<br>- DBSCAN<br>- Hierarchical Clustering<br>- Principal Component Analysis (PCA)<br>- Independent Component Analysis (ICA) | - **Customer Segmentation**: Dividing customers based on purchasing behavior.<br>- **Anomaly Detection**: Detecting unusual patterns in data.<br>- **Market Basket Analysis**: Finding relationships in transaction data.<br>- **Image Compression**: Reducing image size while retaining structure. |
| **Semi-supervised Learning** | - Combines a small amount of labeled data with a large amount of unlabeled data.<br>- Helps in scenarios where labeling data is expensive or time-consuming.<br>- Balances supervised and unsupervised methods. | - Self-training<br>- Graph-based Methods<br>- Co-training<br>- Generative Models | - **Speech Recognition**: Improving accuracy using a mix of labeled and unlabeled audio data.<br>- **Medical Imaging**: Identifying tumors with limited labeled samples.<br>- **Web Content Classification**: Categorizing webpages with minimal labeled data.<br>- **Fraud Detection**: Identifying anomalies in finance data. |
| **Reinforcement Learning** | - Learns through interaction with an environment.<br>- Rewards and penalties guide the learning process.<br>- Used for problems involving decision-making and sequential tasks. | - Q-Learning<br>- Deep Q-Networks (DQN)<br>- Policy Gradient Methods<br>- Actor-Critic Algorithms | - **Game AI**: Developing agents for games like Chess or Go (e.g., AlphaGo).<br>- **Robotics**: Teaching robots tasks like walking or picking objects.<br>- **Self-Driving Cars**: Navigating traffic and avoiding obstacles.<br>- **Dynamic Resource Allocation**: Optimizing network bandwidth or cloud resources. |

## Difference Between Classification and Regression in Machine Learning

| **Aspect** | **Classification** | **Regression** |
|---|---|---|
| **Definition** | Predicts **categories** or **labels** for given input data. | Predicts **continuous values** or **quantitative outputs** for given input data. |
| **Output Type** | Discrete (e.g., Yes/No, Spam/Not Spam, Class A/Class B). | Continuous (e.g., temperature, sales, house price). |
| **Goal** | Assign data points to one of several predefined classes or categories. | Predict a numerical value based on input features. |
| **Key Algorithms** | - Logistic Regression<br>- Decision Trees<br>- Random Forest<br>- Support Vector Machines (SVM)<br>- K-Nearest Neighbors (KNN)<br>- Neural Networks | - Linear Regression<br>- Polynomial Regression<br>- Support Vector Regression (SVR)<br>- Ridge Regression<br>- Lasso Regression |
| **Performance Metrics** | - Accuracy<br>- Precision<br>- Recall<br>- F1-score<br>- Confusion Matrix | - Mean Absolute Error (MAE)<br>- Mean Squared Error (MSE)<br>- Root Mean Squared Error (RMSE)<br>- R-squared |
| **Example Tasks** | - Spam email detection<br>- Disease diagnosis (e.g., diabetes: Yes/No)<br>- Customer churn prediction | - Predicting house prices<br>- Forecasting sales revenue<br>- Predicting temperature changes |
| **Input Data** | Data with features and corresponding labels (categories). | Data with features and corresponding continuous target values. |
| **Use Cases** | - Medical diagnosis (e.g., cancer detection)<br>- Fraud detection<br>- Sentiment analysis<br>- Object recognition | - Predicting stock prices<br>- Energy consumption forecasting<br>- Predicting employee salaries |
| **Nature of Prediction** | Determines the **class** or category of an input. | Determines the **value** or quantity associated with an input. |
| **Visualization of Output** | Bar charts or confusion matrix showing class distributions. | Scatter plots, line graphs, or regression lines showing trends. |

## Difference Between Supervised and Unsupervised Learning

| **Aspect** | **Supervised Learning** | **Unsupervised Learning** |
|---|---|---|
| **Definition** | A type of machine learning where the model learns from **labeled data** (input-output pairs). | A type of machine learning where the model learns from **unlabeled data** without predefined outputs. |
| **Goal** | To predict **outputs** or **labels** for new data based on patterns learned from the training data. | To find **patterns**, **groups**, or **relationships** in the data. |

| | | |
|---|---|---|
| **Input Data** | Data with both input features and corresponding output labels. | Data with only input features, no associated labels. |
| **Output** | Predicts specific outcomes (e.g., categories or numerical values). | Identifies hidden structures like clusters or dimensionality reduction. |
| **Key Algorithms** | - Linear Regression<br>- Logistic Regression<br>- Decision Trees<br>- SVM<br>- Neural Networks | - K-Means Clustering<br>- DBSCAN<br>- Hierarchical Clustering<br>- PCA<br>- ICA |
| **Performance Metrics** | - Accuracy<br>- Precision<br>- Recall<br>- Mean Squared Error (MSE)<br>- R-squared | - Silhouette Score<br>- Inertia (for clustering)<br>- Explained Variance Ratio (for dimensionality reduction) |
| **Example Tasks** | - Spam email detection<br>- Disease diagnosis (e.g., cancer: Yes/No)<br>- Predicting house prices | - Customer segmentation<br>- Market basket analysis<br>- Anomaly detection |
| **Learning Approach** | Relies on **guidance** from labeled examples to map inputs to outputs. | Explores **hidden patterns** without external supervision. |
| **Complexity of Output** | Produces specific, actionable predictions like a label or value. | Provides insights about data structure without specific actionable predictions. |
| **Real-World Applications** | - Fraud detection<br>- Image recognition<br>- Predictive analytics | - Recommender systems<br>- Customer segmentation<br>- Data visualization |
| **Ease of Implementation** | Easier to train because of clear guidance (labeled data). | More challenging due to lack of labels and reliance on inherent patterns. |