Machine Learning Overview

Definition

Field of study that gives computers the ability to learn without being explicitly programmed.

Knowledge Pyramid

- **Data** (facts) → **Information** (processed data) → **Knowledge** (condensed data) → **Intelligence** (applied knowledge) → **Wisdom**
  Note: Machines do not possess wisdom.

Basic Outlook

- **Data → Learning Program (Model) → Decisions**

Info Byte

- **Deep Learning (DL)** is a sub-branch of Machine Learning (ML).
- **Artificial Intelligence (AI)** is the superset of ML.
- **ML** is a branch of **Data Science**.

Relationship of ML with Other Major Fields

1. **Big Data**
   - **Volume:** Huge data sets.
   - **Variety:** Different formats, e.g., images, videos.
   - **Velocity:** Speed at which data is generated and processed.
   - Big Data is used by many ML algorithms.
2. **Data Mining (DM)**
   - Finds hidden patterns and valuable information in data.

- **DM:** Extracts hidden patterns.
  ML: Uses these patterns to make predictions.

3. **Data Analysis (DA)**
   - A branch of Data Science aiming to extract useful knowledge from crude data.
   - ML and DA are closely related.
4. **Pattern Recognition**
   - An engineering field that uses ML for pattern analysis and classification.

Methods of Representing Patterns in Data

1. Mathematical equations
2. Relational diagrams (e.g., trees/graphs)
3. Logical if/else rules
4. Grouping (clusters)

Types of Machine Learning

1. Supervised Learning

- **Classification**
- **Regression (Prediction)**

2. Unsupervised Learning

- **Cluster Analysis** (Grouping)
- **Association Mining** (Finding hidden patterns and valuable information)
- **Dimension Reduction** (Reducing dataset features to simplify data)

3. Semi-Supervised Learning

4. Reinforcement Learning

Info Byte

- **Attributes:** General properties of data (often refer to features or characteristics).
- **Features:** Specific input variables used to make predictions.
- **Labels:** Output variables that the model aims to predict.

- **Note:** Features and attributes are often the same.

Labeled vs Unlabeled Data

1. Labeled Data

Definition:

Labeled data is data that has both input information (features) and the correct output (label) for each example. It is used in supervised learning to teach a model the relationship between inputs and outputs.

- Data with both input information (**features**) and correct output (**labels**) for each example.

- Usage: Supervised learning to teach models the relationship between inputs and outputs.

- Example:

```
- Features: Patient age, blood pressure, cholesterol levels.
- Label: Diagnosis (e.g., "Healthy" or "Diabetic").
```

2. Unlabeled Data

Definition:

Unlabeled data consists only of input features, without any associated target labels. It is used in **unsupervised learning**, where the model tries to find patterns, groupings, or structures in the data without predefined categories or outcomes.

- Data that consists only of input features, without associated labels.
- **Usage:** Unsupervised learning to find patterns or structures.
- **Example:**
  - Customer demographic data (e.g., age, location, purchase history) without predefined categories.
    The model identifies patterns (e.g., high spender, low spender).

Supervised Learning vs Unsupervised Learning

Supervised Learning

Definition:

Supervised learning is a type of machine learning where the model is trained on labeled data. It learns the relationship between input features and corresponding labels (outputs) to make predictions on new, unseen data.

- Model is trained on **labeled data**.
- Learns the relationship between **features** and corresponding **labels** to make predictions on new data.
- **Example:** Recognizing whether an email is spam or not, using labeled examples (spam or not spam).

Unsupervised Learning

Definition:

Unsupervised learning is a type of machine learning where the model is trained on unlabeled data. It tries to find patterns, groupings, or structures in the data without predefined labels or outcomes.

- Model is trained on **unlabeled data**.
- Finds patterns, groupings, or structures without predefined labels or outcomes.
- **Example:** Clustering customers into groups based on purchasing behavior without predefined categories.

Machine Learning Process

1. Understand the business ↔ Understand the data
2. Data Preprocessing
3. Modeling
4. Model Evaluation
5. Model Deployment

Challenges of Machine Learning

1. Problems (ML solves **only well-posed problems**; cannot solve ill-posed problems.)

2. Huge Data
3. High Computation Power
4. Complexity of Algorithms
5. Bias/Variance Tradeoff

Bias

Definition:
Bias is the error introduced by overly simplifying the model's assumptions about the data. It represents how much the model's predictions deviate from the true values. High bias means the model makes strong assumptions and is too simplistic, leading to systematic errors.

Explanation:
When a model has high bias, it cannot capture the underlying patterns in the data because it is too simple. This often leads to underfitting, where the model misses important relationships. Low bias means the model is more flexible and can better approximate the true data, but may still face other issues like overfitting.

High Bias Example:
Imagine you're trying to predict house prices based only on the number of rooms using a linear regression model. If the true relationship between house prices and features like location, size, and condition is more complex, the linear model will oversimplify the situation, resulting in high bias. This model might consistently underpredict or overpredict prices because it cannot capture the real-world complexity.

Low Bias Example:
If you use a decision tree model to predict house prices, which splits data based on multiple factors like number of rooms, location, and age of the house, it can capture the more complex relationships between these features. This results in lower bias, as the model adapts better to the true underlying patterns in the data.

Variance

Definition:
Variance refers to the model's sensitivity to the specific data used for training. High variance means the model fits the training data very well, but this can cause overfitting, where the model captures noise or random fluctuations in the data instead of general trends.

Explanation:
When a model has high variance, it may perform very well on the training data, but poorly on unseen test data, as it becomes too tailored to the training data. Low variance means the model is more stable and generalizes better to new data, but may not fit the training data as perfectly.

High Variance Example:
Imagine you're predicting house prices with a deep decision tree model. If the tree has too many splits, it might perfectly fit the training data, but it will also be highly sensitive to small changes in the data. For example, if a new house in the test set has slightly different features, the model might make wildly inaccurate predictions, leading to high variance and poor generalization.

Low Variance Example:
If you use a linear regression model to predict house prices with a small set of features (like number of rooms), the model will not capture as much detail as a decision tree. While it might not fit the training data as well as the decision tree, it is less likely to be influenced by small changes in the data. This leads to lower variance, making the model more stable and likely to perform better on new data.