ML Unit-1 Notes by OMG

1. Compare Machine Learning vs. Artificial Intelligence.

Artificial Intelligence (AI) is a broad field of computer science that focuses on creating intelligent machines that can perform tasks that typically require human intelligence. This includes things like reasoning, problem-solving, learning, and understanding language. AI is the all-encompassing concept of creating smart machines.

Machine Learning (ML) is a **subfield of AI**. It's a specific approach to achieving AI. Instead of explicitly programming a machine to perform a task, ML algorithms allow the machine to **learn from data** and improve its performance over time.

Here's a table to summarize the key differences:

Feature	Artificial Intelligence (AI)	Machine Learning (ML)
Scope	Broad field, encompassing all aspects of creating intelligent machines.	A subfield of AI focused on learning from data.
Goal	To create intelligent machines that can simulate human intelligence.	To enable machines to learn from data without being explicitly programmed.
Approach	Can be rule-based (expert systems) or learning-based.	Primarily learning-based.
Example	A chess-playing program that uses a set of rules and heuristics to make moves.	A spam filter that learns to identify spam emails by analyzing a large dataset of emails.

In short, **all machine learning is AI, but not all AI is machine learning**.

---

2. Describe parametric and Non-parametric machine learning models.

Parametric models are those that make strong assumptions about the form of the data. They have a **fixed number of parameters**, regardless of the amount of training data. These models are generally simpler, faster to train, and require less data. However, if the assumptions about the data are wrong, the model's performance will be limited.

- **Examples:** Linear Regression, Logistic Regression, Naive Bayes.

Non-parametric models, on the other hand, do not make strong assumptions about the form of the data. The number of parameters in a non-parametric model **grows with the amount of training data**. These models are more flexible and can fit a wider range of data distributions, but they are also more prone to overfitting and can be computationally expensive.

- **Examples:** k-Nearest Neighbors (k-NN), Decision Trees, Support Vector Machines (SVMs).

---

3. Explain various Data formats that conform ML elements.

Machine learning models require data in a structured format. Some of the most common data formats include:

- **CSV (Comma-Separated Values):** A simple text file where values are separated by commas. It's easy to read and write and is widely supported by data analysis tools.
- **JSON (JavaScript Object Notation):** A lightweight data-interchange format that uses human-readable text to transmit data objects consisting of attribute-value pairs. It's often used for web applications.
- **XML (eXtensible Markup Language):** A markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. It's more verbose than JSON.
- **Databases:** Data can also be stored in relational databases (like MySQL, PostgreSQL) or NoSQL databases (like MongoDB). SQL is used to query relational databases, while

NoSQL databases have their own query languages.

No matter the format, data for ML usually consists of:

- **Features:** The input variables used to make predictions.
- **Target (or Label):** The output variable that the model is trying to predict.

---

4. Explain supervised, unsupervised and semi-supervised learning.

These are the three main types of machine learning:

- **Supervised Learning:** In supervised learning, the model learns from **labeled data**. This means that each data point in the training set has both the input features and the corresponding output label. The goal is to learn a mapping function that can predict the output for new, unseen data.
  - **Examples:**
    - **Classification:** Predicting a category, like "spam" or "not spam" for an email.
    - **Regression:** Predicting a continuous value, like the price of a house.
- **Unsupervised Learning:** In unsupervised learning, the model learns from **unlabeled data**. The goal is to find hidden patterns and structures in the data without any predefined labels.
  - **Examples:**
    - **Clustering:** Grouping similar data points together, like customer segmentation.
    - **Dimensionality Reduction:** Reducing the number of variables in the data while preserving the important information.
- **Semi-supervised Learning:** This is a combination of supervised and unsupervised learning. It uses a small amount of **labeled data** and a large amount of **unlabeled data**. This is useful when labeling data is expensive or time-consuming.

---

5. Describe various statistical learning approaches.

Statistical learning is the foundation of machine learning. It involves building statistical models to understand data and make predictions. The main approaches include:

- **Regression:** Used to predict a continuous output variable.
  - **Linear Regression:** Models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

    - **Polynomial Regression:** A type of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x.
- **Classification:** Used to predict a categorical output variable.
    - **Logistic Regression:** A statistical model that in its basic form uses a logistic function to model a binary dependent variable.
    - **Support Vector Machines (SVMs):** A supervised learning model that uses a hyperplane to separate data into different classes.
    - **Decision Trees:** A tree-like model of decisions and their possible consequences.
- **Clustering:** Used to group similar data points together.
    - **K-Means Clustering:** An algorithm that partitions a dataset into K distinct, non-overlapping clusters.
    - **Hierarchical Clustering:** An algorithm that builds a hierarchy of clusters.
- **Dimensionality Reduction:** Used to reduce the number of variables in the data.
    - **Principal Component Analysis (PCA):** A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

---

6. Compare Machine Learning with traditional programming.

Feature	Traditional Programming	Machine Learning
Approach	**Rule-based:** The programmer explicitly writes rules for the program to follow.	**Data-driven:** The model learns from data to make predictions or decisions.
Input	Data and a program (rules).	Data and the desired output.
Output	The program's output.	A program (the trained model).
Process	The programmer analyzes the problem and writes	The model learns patterns from the data to solve the

	code to solve it.	problem.
Example	A program that calculates the factorial of a number.	A program that recognizes handwritten digits.

Types of Machine Learning with Examples:

- **Supervised Learning:**
  - **Example:** A credit card company uses a supervised learning model to predict whether a transaction is fraudulent or not based on past transaction data.
- **Unsupervised Learning:**
  - **Example:** An e-commerce website uses an unsupervised learning model to group customers into different segments based on their purchasing behavior. This can be used for targeted marketing.
- **Reinforcement Learning:**
  - **Example:** A self-driving car uses reinforcement learning to learn how to drive by getting rewards for making good decisions (like staying in the lane) and penalties for making bad decisions (like hitting a curb).

---

7. What are various Statistical Learning Approaches?

As mentioned in question 5, statistical learning approaches are the methods used to build models from data. Here's a more in-depth look:

- **Regression:** When you want to predict a continuous value.
  - *When to use:* Predicting house prices, stock prices, or temperature.
- **Classification:** When you want to predict a category.
  - *When to use:* Email spam detection, image classification (e.g., cat vs. dog), or medical diagnosis.
- **Clustering:** When you want to find natural groupings in your data.
  - *When to use:* Customer segmentation, grouping similar documents, or anomaly detection.
- **Dimensionality Reduction:** When you have a large number of features and want to reduce the complexity of your data.
  - *When to use:* To speed up model training, to visualize high-dimensional data, or to avoid the curse of dimensionality.

---

8. Explain different data formats used in Machine Learning.

This is a repetition of question 3. To add more detail, here are some pros and cons of each format in the context of ML:

- **CSV:**
  - **Pros:** Simple, human-readable, widely supported.
  - **Cons:** Not suitable for complex, nested data structures.
- **JSON:**
  - **Pros:** Flexible, can represent complex data structures, commonly used in web APIs.
  - **Cons:** Can be more verbose than other formats.
- **XML:**
  - **Pros:** Very flexible, well-defined standard.
  - **Cons:** Verbose, can be complex to parse.
- **Databases:**
  - **Pros:** Efficient for storing and querying large amounts of data, can handle complex relationships between data.
  - **Cons:** Can be more complex to set up and maintain than flat files.

---

9. What is Machine Learning? Explain applications of Machine Learning in data science.

Machine Learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output valu[1]es.

Applications of Machine Learning in Data Science:

- **Predictive Modeling:** Building models that can predict future outcomes based on historical data. For example, predicting customer churn or sales forecasts.
- **Customer Segmentation:** Grouping customers into different segments based on their characteristics and behavior. This is used for targeted marketing and product recommendations.
- **Recommendation Engines:** Recommending products, movies, or articles to users based on their past behavior. This is used by companies like Netflix and Amazon.

- **Fraud Detection:** Identifying fraudulent transactions or activities. This is used by banks and financial institutions.
- **Natural Language Processing (NLP):** Analyzing and understanding human language. This is used for applications like sentiment analysis, chatbots, and machine translation.
- **Computer Vision:** Enabling computers to "see" and interpret images and videos. This is used for applications like self-driving cars, facial recognition, and medical imaging analysis.

---

10. Explain Geometric Model and Probabilistic Model with suitable examples.

Geometric Models use geometric concepts like lines, planes, and distances to make predictions. They represent data as points in a high-dimensional space and try to find a geometric separation between different classes.

- **Examples:**
  - **Linear Regression:** Fits a line (or hyperplane) to the data.
  - **Support Vector Machines (SVMs):** Finds the hyperplane that best separates the data into different classes.
  - **k-Nearest Neighbors (k-NN):** Classifies a new data point based on the majority class of its k-nearest neighbors.

Probabilistic Models use probability theory to model the relationships between variables. They aim to learn the underlying probability distribution of the data.

- **Examples:**
  - **Naive Bayes:** A classification algorithm based on Bayes' theorem with the "naive" assumption of independence between features.
  - **Logistic Regression:** A classification algorithm that models the probability of a certain class or event existing.
  - **Gaussian Mixture Models (GMMs):** A probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

---

11. How a machine-learning model works? Explain various steps involved.

Here's a breakdown of the typical workflow for a machine learning model:

1. **Data Collection:** Gathering data from various sources. The quality and quantity of data are crucial for the model's performance.
2. **Data Preprocessing:** Cleaning and preparing the data for the model. This includes handling missing values, scaling features, and splitting the data into training and testing sets.
3. **Model Selection:** Choosing the right machine learning algorithm for the problem. This depends on the type of problem (e.g., classification, regression), the size of the dataset, and the nature of the data.
4. **Model Training:** Feeding the training data to the model to let it learn the underlying patterns.
5. **Model Evaluation:** Assessing the model's performance on the testing data. This is done using metrics like accuracy, precision, recall, and F1-score for classification, and mean squared error (MSE) for regression.
6. **Parameter Tuning:** Adjusting the model's parameters to improve its performance. This is also known as hyperparameter optimization.
7. **Prediction/Deployment:** Using the trained model to make predictions on new, unseen data. The model is then deployed into a production environment where it can be used by other applications.

---

12. What are the main differences between supervised learning and reinforcement learning?

Feature	Supervised Learning	Reinforcement Learning
Data	Labeled data (input-output pairs).	No predefined data, the agent learns by interacting with the environment.
Learning Process	Learns a mapping function from input to output.	Learns a policy (a mapping from state to action) through trial and error.
Goal	To predict the correct output for new data.	To maximize the cumulative reward over time.

Feedback	Direct and immediate feedback (the correct label).	Delayed and sparse feedback (rewards or punishments).
Examples	Image classification, spam detection, price prediction.	Self-driving cars, game playing (e.g., AlphaGo), robotics.

---

13. Explain geometric models and its types.

As mentioned in question 10, **geometric models** use geometric concepts to make predictions. Here's a more detailed explanation of the types:

- **Linear Models:** These models assume that the decision boundary between classes is a line (or a hyperplane in higher dimensions).
  - **Examples:** Linear Regression, Logistic Regression, Linear Discriminant Analysis (LDA).
- **Distance-based Models:** These models use the distance between data points to make predictions.
  - **Examples:** k-Nearest Neighbors (k-NN), K-Means Clustering.
- **Kernel-based Models:** These models use kernel functions to map the data into a higher-dimensional space where it can be separated by a hyperplane.
  - **Example:** Support Vector Machines (SVMs).

---

14. Compare Artificial intelligence and Machine learning.

This is a repetition of question 1. To provide a different perspective, think of it this way:

- **AI is the goal:** To create intelligent machines.
- **ML is a tool:** A powerful tool that we can use to achieve that goal.

Another analogy:

- **AI is like building a car.** There are many different ways to build a car.
- **ML is like a specific type of engine.** You can use a gasoline engine, a diesel engine, or

an electric motor. ML is one type of "engine" that can power an AI system.

---

15. Describe grouping and grading models.

- **Grouping Models (Clustering):** These are unsupervised learning models that group similar data points together. The goal is to find natural groupings in the data, without any predefined labels.
    - **Types of Clustering:**
        - **K-Means:** Partitions the data into K clusters.
        - **Hierarchical Clustering:** Creates a tree of clusters.
        - **DBSCAN:** A density-based clustering algorithm.
- **Grading Models (Classification/Regression):** These are supervised learning models that assign a "grade" or a continuous value to an input.
    - **Classification:** Assigns a discrete label (a "grade") to an input. For example, "spam" or "not spam".
    - **Regression:** Predicts a continuous value (a numerical "grade"). For example, the price of a house.