Basic Introduction

1. **Artificial Intelligence**

* Think AI as the big universe.
* AI is overall concept of creating machines and software that can think and act like humans
* Goal of an ai is to simulate human intelligence.
* Key features of ai is can work without human help after setup
* Example is a self-driving car that can drive itself, follow traffic rules and make decisions is an AI system.

1. **Machine Learning**

* It’s a part of AI
* ML is a subset of AI where machines are trained with data to learn patterns and make decisions.
* Key feature is that learns from data, improve with experience.
* Example is you give the self-driving car’s thousand video and it’ll learns how to drive by studying them.

1. **Deep learning**

* A subset of ML
* DL is more advanced type of ML that uses neural network just like human brains to process the data.
* Key feature is to works great with huge data and complex tasks. Like images and speech
* Example is that the car uses deep learning to detect pedestrians or read traffic signs from camera input.

1. **Data Science**

* Data science is about analyzing the data to get insights using statistics, programming and visualization.
* Helps in decision making by understanding the data.
* A company like Uber uses data science to analyze traffic data to plan best routes

Now Imagine a smart assistant like Siri understands your speech is deep learning learns your preferences is machine learning talk and respond to you is AI and analyze your usage data is data science.

**Now let’s get dive into machine learning**

There are 3 main types of machine learning

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

**1. Supervised Machine Learning**

* **Meaning:** The model learns using a **labeled dataset** (input + output is known).
* **Goal:** Predict the output based on input features.
* **Example:** Predict house prices.
* **Input (independent features):** Size of house, Number of rooms
* **Output (dependent feature):** Price of house
* if input changes, output also changes.

**Two Types of Problems:**

1. **Regression**
   * **Output:** Continuous values
   * **Example:** Predict salary, price, temperature
2. **Classification**
   * **Output:** Categorical values
   * **Example:** Pass or Fail, Yes or No, Disease type
   * **Types:**
     + **Binary Classification:** 2 categories (e.g., Pass/Fail)
     + **Multi-class Classification:** More than 2 categories (e.g., A/B/C grades)

**🔧 Common Algorithms:**

* Linear Regression, Logistic Regression
* Decision Tree, Random Forest
* AdaBoost, XGBoost, CatBoost
* Group similar customers (high salary + high spender, low salary + low spender, etc.)
* Hierarchical Clustering
* DBSCAN

**2 . Unsupervised Machine learning**

* **Meaning:** The model learns from data without output labels.
* **Goal:** Find Patterns or groups
* **Example:** Customer Segmentation
* We have Input features but we can not predict the output based on input.
* Used in marketing to find customer types for targeting.
* Used to find hidden patterns in data.

**Common Algorithms:**

* + - * K-Means Clustering
      * Hierarchical Clustering.
      * DBSCAN

1. **Reinforcement Learning**

* **Meaning:** The Model learns from trial and error using rewards and punishment.
* **Goal:** Taake the best actions in an environment to get maximum rewards.
* **Example:** Teaching a robot to walk, Game AI , Self driving car
* A baby tries to walk if it falls then we get negative feedback
* Means it’ll learns from it. Baby walks properly gets claps and rewards means positive feedback.

**Key Difference between Supervised and Unsupervised Learning:**

|  |  |  |
| --- | --- | --- |
| Aspect | Supervised Learning | Unsupervised Learning |
| Labeled Data | Uses labeled data (input-output pairs) | Uses unlabeled data (only inputs) |
| Goal | Predict outcomes or classify data | Discover hidden patterns or structure |
| Examples | Regression, Classification (e.g., Spam Detection) | Clustering, Dimensionality Reduction (e.g., Customer Segmentation) |
| Evaluation | Can be evaluated using metrics like accuracy, RMSE | Harder to evaluate objectively without labels |

**When Would You Use Unsupervised Learning Over Supervised Learning?**

* 1. You don’t have labelled data.
  2. You want to discover patterns.
  3. For improving visualization or performance.
  4. You’re exploring the data.

**Difference Between Supervised Learning and Reinforcement**

|  |  |  |
| --- | --- | --- |
| Feature | Supervised Learning | Reinforcement Learning |
| Training Data | **Requires labeled data (input-output pairs)** | **Learns via interaction with environment using rewards/punishments** |
| Learning Objective | **Learn a mapping from inputs to outputs** | **Learn a policy to maximize cumulative rewards** |
| Feedback Type | **Direct feedback (correct output is known)** | **Delayed feedback (reward comes after a sequence of actions)** |
| Example Use Cases | **Spam detection, image classification, sentiment analysis** | **Game playing (e.g. Chess, Go), robotics, self-driving cars** |
| Output | **Classification or regression** | **Sequence of actions (policy)** |

**Understanding the Equation of a line, Plane and Hyperplane in Machine Learning:**

Machine learning isn’t just about code, it’s deeply rooted in mathematics and geometry. Critical concepts behind many ML algorithms is the equation of straight line and its extension into higher dimensions – a hyperplane. These geometric elements are what allow algorithms like logistic Regression and support vector machines to make decisions.

**Equation of straight line (2D) : y = mx + c**

* Where m = slope and c = where the line crosses the y axis
* X and Y are coordinates on the graph.
* Example y = 2x + 3 here slope is 2. Every time x increases by 1, y increases by 2. Intercept is 3 the line touches the y axis at (0,3)
* The alternate form used in ML is ax + by + c = 0
* Or we can also use y = β₀ + β₁x
* You can also write as y = β0 + β1x (same meaning, just different symbols)
* Or as ax + by + c = 0, which is the same in a rearranged form.

**Equation of a Line in Vector Form**

In machine learning we often write the equation like: wᵀx + b = 0

General form of line/plane/hyperplane

Where w= weight vector , x= input features and b= bias or intercept

Then the equation be like : w1x1 + w2x2 + b = 0

W = means weights means howmuch each features matter.

**redicting House Prices**

* Suppose we have a dataset with features like:
  + **Size of the house** ( x1 )
  + **Number of bedrooms** ( x2 )
  + **Location rating** ( x3 )
* The model learns the weights ( w1, w2, w3)) to determine how much each feature impacts the price of the house.
* The final equation might look like: w1x1 + w2x2 + w3x3
* Example: If the size has a higher weight than location, that means **house size matters more than location** for predicting prices.

**Imagine: Predicting Exam Pass/Fail**

| **Hours Studied (x1)** | **Hours Slept (x2)** | **Result (y)** |
| --- | --- | --- |
| 3 | 2 | 1 (Pass) |
| 1 | 1 | 0 (Fail) |
| 4 | 3 | 1 (Pass) |

**Step-by-Step — Like a Learning Robot**

**Step 1 : Model makes a guess**

* The model randomly picks weights:
* W1=0.5 w2 =0.5 and bias b=0
* guess = w1 \* x1 + w2 \* x2 + b
* For a student who studied 3 hours and slept 2 hours:
* guess = 0.5 \* 3 + 0.5 \* 2 = 1.5 + 1 = 2.5
* but the real answer is 1 and we got 2.5 which is very high

**Step 2: Model Gets Feedback**

* If loss then the model calculates how far off .

**Step 3: Model Fixes Itself (Gradient Descent)**

* If study hours were more useful then increase w1
* If sleep wasn’t important decrease w2
* Keeps adjusting to reduce errors.

**Step 4: Repeat and Learn**

* It learns how much studying matters
* It learns how much sleeping matters.
* Now it can say: prediction = 2x1 + 3x2 -12

Cause : 2\*3 + 3\*2 -12 = 6+6-12 =0

* If output is >= 0 student pass
* If output < 0 student fails

**What is we have 3 features? 3D plane?**

In 3D instead of a line, we get a plane . w1x1 + w2x2 + w3x3 + b = 0

**Line/Plane passing through the origin**

If the intercept b = 0, then the line/Plane passes through the origin.

wᵀx = 0

**Geometry behind wᵀx = 0**

If wᵀx = 0, then cos a = 0 then a = 90°

This explanation walks through a fundamental linear algebra concept — **how to compute the distance of a point from a plane**, which is essential in machine learning algorithms like **logistic regression** and **Support Vector Machines (SVM)**.

* 1. **Plane Equation**

If a plane passes through the origin , it’s equation can be written as : **wᵀx = 0**

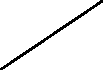
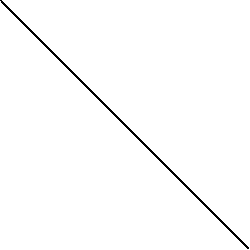
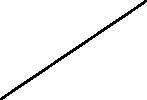
Where w is a vector normal (perpendicular) to the plane.

X is any point on the plane.

Let’s considered that I have a plane.



A(0,0)



This particular plane passes through the origin. For wTx = 0 where w is a vector, which is perpendicular to origin. Now how can we find the distance between s and b. In some of the problems like logistic regression we have to use this, like classification problems… let’s say there is one more point which is s^-1



Given a point s=[x1,x2,...,xn]s = [x\_1, x\_2, ..., x\_n]s=[x1​,x2​,...,xn​] in **n-dimensional space**, the **distance d** of the point from the plane is given by:

If the point lies above the plane at same direction as w then

If the point lies below the plane opposite direction of w then

So, above the plane it’s always Positive Distance

Below the Plane Negative distance

Example 1 : Let’s assume a 3D plane with normal vector w=[2,3,-1] passing through the origin. So the plane equation is 2x + 3y – z = 0

Point Above the plane let s = [1,1,1]

* 1. Compute wTs = 2(1) + 3(1) – 1(1) = 4
  2. ||W|| = =

Distance is positive meaning the point is above the plane.

If distance is negative then point is below the plane.

| **Feature** | **Instance-Based Learning** | **Model-Based Learning** |
| --- | --- | --- |
| Learning Style | Memorizes training data | Learns patterns from data |
| Dependency | Heavily dependent on training data | Learns and generalizes from training data |
| Pattern Recognition | No (or postponed until prediction) | Yes (during training itself) |
| Decision Making | Based on surrounding instances (local) | Based on learned model (global) |

**Instance-Based Learning vs Model-Based Learning**

**2. How Prediction Works**

* **Instance-Based**: Compares new instance to nearby examples in training data (e.g., K-Nearest Neighbors).
* **Model-Based**: Applies a learned function or decision boundary to make predictions.

**3.Technical Aspects**

| **Feature** | **Instance-Based** | **Model-Based** |
| --- | --- | --- |
| **Example Algorithms** | **K-Nearest Neighbors (KNN), Kernel methods** | **Linear/Logistic Regression, Decision Trees, SVM, Neural Networks** |
| **Training** | **Minimal (no actual "training")** | **Intensive (parameter estimation, optimization)** |
| **Storage** | **Requires full training data** | **Only model (parameters)** |
| **Speed (Prediction Time)** | **Slower (search through data)** | **Faster (apply function)** |
| **Storage Size** | **High (store all training data)** | **Low (model file, e.g., .pkl, .h5)** |
| **Generalization** | **Poor** | **Strong** |
| **Real-Time Use** | **Less suited** | **Well suited** |
| **Can discard training data after training?** | **No** | **Yes** |

**4. Analogy**

* **Instance-Based** = **Memorizing answers**
  + Like a student who memorizes questions and answers.
* **Model-Based** = **Understanding concepts**
  + Like a student who learns concepts and can solve new problems.

**5.Summary Table of Distance**

| Aspect | Instance-Based | Model-Based |
| --- | --- | --- |
| Learns Pattern? | No | Yes |
| Needs Training Data After Training? | Yes | No |
| Speed at Prediction | Slower | Faster |
| Generalization | Poor | Good |
| Model File Size | Large | Small |
| Example Algorithm | KNN | Linear Regression |
| Model Form | No fixed form | Mathematical function/model |

**Interview Questions and Answers: Data Science & Machine Learning from above topics:**

**1. What is Artificial Intelligence (AI)?**

**Answer:** AI is the science of creating intelligent machines that simulate human thinking and behaviour. It enables machines to learn, reason, and make decisions without human intervention.

**2. How is Machine Learning different from Artificial Intelligence?**

**Answer:** Machine Learning is a subset of AI where machines learn from data to make decisions. AI includes ML along with other areas like reasoning, robotics, and expert systems.

**3. What is Deep Learning and how is it related to Machine Learning?**

**Answer:** Deep Learning is a subset of ML that uses neural networks with many layers to analyze large and complex data like images, audio, and text.

**4. How would you explain Data Science to a non-technical person?**

**Answer**: Data Science involves analyzing data using statistics and programming to extract useful insights that support business decisions.

**5. Can you give an example where AI, ML, DL, and DS are used together?**

**Answer**: In a smart assistant like Siri: AI enables interaction, ML learns user preferences, DL processes speech, and DS analyzes usage data.

**Machine Learning Types**

**6. What are the main types of machine learning?**

**Answer**: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

**7. Explain supervised learning with an example.**

**Answer:** Supervised learning uses labeled data. For example, predicting house prices based on features like size and location.

**8. Explain unsupervised learning with an example.**

**Answer:** It uses unlabeled data to find hidden patterns, like customer segmentation in marketing.

**9. What is reinforcement learning? Give a real-world analogy.**

**Answer:** RL involves learning through trial and error using rewards and punishments. Example: training a robot to walk.

**10. How is reinforcement learning different from supervised learning?**

**Answer:** Supervised learning uses labeled data with direct feedback, while RL uses delayed feedback from interactions with the environment.

**Unsupervised Learning**

**15. When would you prefer unsupervised learning over supervised learning?**

**Answer:** When labeled data is unavailable or you want to discover hidden patterns in data.

**16. Name a few algorithms used for clustering.**

**Answer:** K-Means, DBSCAN, Hierarchical Clustering.

**17. What is customer segmentation and how is it useful?**

**Answer:** It groups customers by behavior or features to enable targeted marketing and better service.

**18. What are some challenges in evaluating unsupervised models?**

**Answer:** Lack of labeled data makes it hard to measure accuracy; evaluation is often subjective or based on domain knowledge.

**Reinforcement Learning**

**19. What is a reward in reinforcement learning?**

**Answer:** It is feedback given to the agent to indicate success or failure of its actions.

**20. Can you explain reinforcement learning using a game scenario?**

**Answer:** In a game like chess, the agent tries moves and learns strategies based on win/loss outcomes (rewards).

**21. How does feedback differ between supervised and reinforcement learning?**

**Answer:** Supervised learning has immediate and direct feedback; RL has delayed feedback based on sequences of actions.

**Mathematics & Geometry in ML**

**22. What is the equation of a straight line used in machine learning?**

**Answer:** y = mx + c or y = β₀ + β₁x

**23. How do you represent a hyperplane in vector form?**

**Answer:** wᵀx + b = 0, where w is the weight vector, x is input, and b is bias.

**24. What does the vector w represent in the equation wᵀx + b = 0?**

**Answer:** It represents the weight or importance of each feature in the input.

**25. What is the significance of bias b in the model?**

**Answer:** Bias allows the model to shift the decision boundary away from the origin.

**26. How do you interpret weights in a model like linear regression?**

**Answer:** Weights indicate how much each feature influences the output.

**Distance from a Point to a Plane**

**27. How do you calculate the distance of a point from a plane?**

**Answer:** Distance d = |wᵀs| / ||w||, where s is the point, and w is the plane's normal vector.

**28. What does it mean if the distance is positive or negative?**

**Answer:** Positive: point lies above the plane in the direction of w; Negative: point lies below the plane.

**29. Why is understanding this distance important in logistic regression and SVM?**

**Answer:** It helps determine how confidently a point belongs to a class by measuring margin from the decision boundary.

**Instance-Based vs Model-Based Learning**

**30. What is instance-based learning? Give an example.**

**Answer:** It memorizes training data and uses it to predict new instances, e.g., K-Nearest Neighbors.

**31. What is model-based learning? How is it different from instance-based?**

**Answer:** Model-based learns patterns and generalizes; unlike instance-based, it doesn't need the full dataset during prediction.

**32. What are some examples of instance-based algorithms?**

**Answer:** K-Nearest Neighbors (KNN), Kernel methods.

**33. What are the advantages and disadvantages of model-based learning?**

**Answer:** Advantages: fast prediction, small model size, good generalization. Disadvantages: requires training and tuning.

**34. Why is model-based learning faster at prediction time?**

**Answer:** Because it uses a trained function instead of comparing to all training data.

**35. Can you explain with an analogy the difference between instance-based and model-based learning?**

**Answer:** Instance-based is like memorizing answers for a test; model-based is like understanding concepts to solve new problems.

**Practical Understanding**

**36. How does a model learn feature importance using gradient descent?**

**Answer:** Gradient descent helps the model find the best weights for each feature. If a feature helps reduce the error more, its weight is increased. **Example:** If studying more helps predict exam pass better than sleeping, the model increases the weight for 'hours studied' more than 'hours slept'.

**37. What happens when the model gets a prediction wrong?**

**Answer:** It calculates how far the prediction is from the actual result (error). Then it updates its weights to reduce that error for future predictions. **Example:** If a student passed but the model predicted fail, it corrects the weights so it won't make the same mistake next time.

**38. What are training and prediction steps in a typical ML pipeline?**

**Answer:**

* **Training:** Feed labeled data to the model so it can learn patterns.
* **Prediction:** Use the trained model to make predictions on new, unseen data. **Example:** Train a model on 100 student records to predict pass/fail, then use it on new student data.

**39. Why can model-based systems discard training data after training?**

**Answer:** Because all the learning (patterns and parameters) is stored in the model. It no longer needs the original data to make predictions. **Example:** Once a student understands the concept, they don’t need to re-read the textbook to solve new questions.

**40. What is a serialized model, and why is it useful?**

**Answer:** A serialized model is a saved version of a trained model (e.g., .pkl, .h5). It lets you reuse the model without retraining. **Example:** You train a model once and save it as a file. Later, you load this file and use it to make predictions without training again.