Machine Learning basics

- **Work done by Tharunithi TJ**

Module 1: Types of Machine Learning Systems

What is Machine Learning?

Quiz:

1. **What is the primary goal of machine learning?**

The primary goal of machine learning is to enable computers to learn from data and improve their performance on specific tasks over time.

The primary goal of machine learning is to create machines that can perform only a single task.

The primary goal of machine learning is to replace human decision-making entirely.

2. What is the difference between supervised and unsupervised learning?

Supervised learning and unsupervised learning are interchangeable terms with no distinct differences.

Supervised learning only uses unlabeled data and does not require any correct answers during training.

Supervised learning involves training models on labeled data, where the correct answers are provided, and the model learns to map input data to output labels.

Unsupervised learning analyzes unlabeled data to discover patterns or relationships without explicit output labels.

3. **Select two examples of applications for machine learning in real-world scenarios.**

Machine learning is exclusively used in robotics

In marketing, machine learning is utilized for customer segmentation, allowing businesses to target specific groups with personalized advertisements.

Machine learning is only used in theoretical research and does not have practical applications in real-world scenarios.

Machine learning is applied in healthcare for disease prediction and diagnosis, using patient data to identify potential health issues.

AI vs. Machine Learning: How Do They Differ?

Quiz:

1.Artificial Intelligence is _____

is a branch of computer science constituting underlying technology that governs expert machines

a branch of computer science that constitutes underlying technology that enables computers to simulate human intelligence

a branch of computer algorithms that facilitates an expert machine to make accurate predictions

2. **Which of the following statements about machine learning is TRUE?**

Machine learning is the intelligence technology that provides computers with advanced abilities to execute processes without being specifically programmed to do so

Machine learning and Artificial intelligence are synonymous

Machine learning is the intelligence technology developed for expert machines to facilitate the learning processes of various tasks

Machine learning is the intelligent technology an expert system uses to make accurate predictions

3. **True or False: The late 1980s marked a resurgence in the development of artificial intelligence (AI), driven by advancements in areas like chess and computer vision.**

False

True

Pipeline of Machine Learning

Quiz:

1. **Why is data gathering a critical first step in the machine learning workflow?**

Models can learn effectively without a representative dataset.

Data gathering provides the foundation for testing models only, ensuring they are representative of real-world scenarios.

Data gathering provides the foundation for training and testing models, ensuring they are representative of real-world scenarios.

Data gathering is an optional step in the machine learning workflow and doesn't significantly impact the model's performance.

2. **What is the purpose of the training and testing phases in the workflow?**

The training and testing phases in the workflow are irrelevant and can be skipped since models inherently understand all types of data without any need for learning or evaluation.

The purpose of the training and testing phases in the workflow is to ensure that the model can learn patterns from the data and generalize its knowledge to make accurate predictions on new, unseen data.

The purpose of the training and testing phases in the workflow is to ensure that the model can learn patterns from the data and memorize the learned patterns.

3. **Why is model evaluation crucial in the machine learning workflow?**

Model evaluation provides a way to quickly complete the machine learning project.

Model evaluation measures the performance of the model and ensures its effectiveness on new, unseen data.

Model evaluation is unnecessary as machine learning models always perform perfectly.

Module 2: Types of Machine Learning Systems

Types of Machine Learning Systems

Quiz:

1. **What is the key distinction between supervised and unsupervised learning?**

Unsupervised learning uses labeled data, while supervised learning works with unlabeled data to find patterns

Supervised learning uses labeled data, while unsupervised learning works with unlabeled data to find patterns

2. **Select the correct example of a real-world application for each type of machine learning.**

Supervised learning in email spam filtering, reinforcement learning in customer segmentation, and unsupervised learning in training self-driving cars.

Supervised learning in email spam filtering, unsupervised learning in customer segmentation, and reinforcement learning in training self-driving cars.

3. **Select the correct machine learning type for Supervised learning**

Trains models on unlabeled data to discover hidden patterns and groupings.

Uses labeled data to make predictions based on previously known outcomes.

Involves agents interacting with an environment to maximize rewards through learned actions.

Types of Supervised Learning

Quiz:

1. **What is the difference between supervised and unsupervised machine learning?. Which of the following statements are true?**

Supervised learning problems can be grouped into clustering and association.

Supervised learning is used when we want to predict a certain outcome from a given input.

There are two major types of supervised learning problems, called clustering and regression.

The goal for unsupervised learning is to model the underlying structure or distribution in the data.

2. **Select which of the following scenarios are regression problems.**

Predict how much a company will spend on electricity the next semester.

Predict whether a user will churn from the service.

Given a tweet, determine whether or not it contains text against or on favor for a presidential candidate.

Predict the score that a student will achieve in an exam whose grade can be 0.1, 2,. . . , 10

3. **Select which of the following scenarios are classification problems.**

Predict the prices of a house in Boston based on zipcode, neighbourhood, the per capita crime rate by town, etc

An algorithm is trained to recognize spam email by learning the characteristics of what constitutes spam vs non-spam email.

Determine whether a customer is likely to purchase more items or not

Impact of blood alcohol content on coordination

4. **Suppose you want to develop a supervised machine learning model to predict whether a superhero will fly or not. Which of the following statements are true?**

A classification model provide the best approach.

A regression model is the best way to predict the probability to fly.

We'll use unlabeled examples to train the model.

This is not a machine learning problem

Types of Unsupervised Learning

Quiz

1. **True or False: Clustering is a common task in unsupervised learning, where data points are grouped together based on similarity**

False

True

2. **Imagine you work for an e-commerce company. The company has a large database of customer transactions, and your task is to segment customers into different groups for**

targeted marketing. Which unsupervised learning technique would you use?

For customer segmentation in e-commerce, you can use a dimensional reduction algorithms like PCA.

For customer segmentation in e-commerce, you can use clustering algorithms like K-means or hierarchical clustering.

3. **You are a financial analyst at a bank and need to detect potentially fraudulent transactions in a credit card dataset. How could unsupervised learning be applied to identify unusual transaction patterns indicative of fraud?**

Unsupervised learning can be applied to detect fraudulent transactions by clustering normal and abnormal transaction patterns or using outlier detection methods to identify unusual transaction behavior.

Unsupervised learning can be applied to detect fraudulent transactions by training a supervised machine learning model on a labeled dataset of fraudulent and non-fraudulent transactions.

Unsupervised learning can NOT be applied to detect fraudulent transactions.

Knowledge Test

1. **Customer Segmentation for Marketing Strategy**

Given a dataset containing customer demographics and purchase history, how can we group customers based on their similarities to tailor marketing strategies?

Supervised Learning

Unsupervised Learning

2. **Anomaly Detection in Network Traffic**

How can we identify unusual patterns or anomalies in network traffic that may indicate a security breach, without having prior labeled examples of such incidents?

Unsupervised Learning

Supervised Learning

Both options are possible

3. **Predicting Housing Prices based on Features**

Given historical data on housing prices and features such as location, size, and amenities, can we build a model to predict the prices of new houses?

Unsupervised Learning

Both options are possible

Supervised Learning

4. **redicting Customer Satisfaction for an E-commerce Website**

Given customer feedback data on an e-commerce website where customers can rate their satisfaction on a scale from 1 to 5, should we model this problem as a classification or a regression task?

Classification problem

Regression problem

Both options are possible

5.You are a healthcare researcher aiming to identify potential subgroups of patients based on their medical records to personalize treatment plans. How could unsupervised learning be utilized to uncover distinct patient clusters with similar medical profiles, allowing for more targeted and effective healthcare interventions?

Unsupervised learning can only be used if the dataset contains labeled patient groups.

Unsupervised learning is not applicable in this scenario, and supervised learning should be used instead.

Unsupervised learning can segment patients into clusters based on their medical records, revealing distinct subgroups without using any predefined labels.

Unsupervised learning can classify patients into predefined categories based on their medical records.

6. **You are an e-commerce manager and want to predict whether customers are likely to make a purchase during their website visit. How can supervised learning be applied to develop a predictive model that helps classify customers into 'potential buyers' and 'non-buyers' based on historical data and labeled purchase information?**

Supervised learning is not suitable for this scenario, and unsupervised learning should be used instead.

Supervised learning requires an unsupervised learning pre-processing step to classify customers effectively.

Supervised learning can only be used if the purchase data is anonymized and doesn't contain labels.

Supervised learning can be employed to classify customers into groups based on their purchase history and other features, allowing for targeted marketing strategies.

Module 3: Data quality and quantity

What data features are important for machine learning projects or applications?

Quiz

1. **True or False: Incorporating irrelevant features in a machine learning model can improve its predictive performance**

True

False

2. **True or False: Feature selection in machine learning helps in reducing the dimensionality of the dataset.**

True

False

3. **Scenario Question**

You are tasked with developing a spam email filter using machine learning. The dataset you have contains email

samples labeled as "spam" or "not spam." Which features would you consider important for this classification task?

A. Sender's Email Address B. Word Frequency in the Email C. Font Style and Size D. Number of Links in the Email E. Email Attachment Size

Select the correct options:

B, D, A

A, C, E

A, D, E

B, D, E

4. **Scenario Question**

You are working on a machine learning project to predict customer churn for a subscription-based service. The dataset you have includes customer information such as age, usage patterns, subscription type, and a unique customer ID for each record.

Why would using the customer ID as a feature for predicting churn be ineffective?

The customer ID is not related to the underlying reasons for churn and including it may introduce noise into the predictive model.

The customer ID is a unique identifier and does not contain predictive information regarding the likelihood of churn.

Using the customer ID as a feature may raise privacy concerns and ethical considerations, as it directly identifies individuals.

How much data is needed for machine learning?

Quiz

1. **True or False: Having a large dataset guarantees high data quality**

False

True

2. **True or False: Data quality is a subjective term and does not have a universally agreed-upon definition**

False

True

3. **Scenario Question**

You are working on a project to predict student performance based on various factors such as study hours, attendance, and extracurricular activities. The dataset you have includes information about these features for each student.

In this scenario, what would be an example of a high-quality dataset ?

A. A dataset with accurate and up-to-date information about study hours, attendance, and extracurricular activities for a diverse set of students.

B. A dataset with study hours as the only feature for a small subset of students.

C. A dataset with random and inconsistent entries for study hours, attendance, and extracurricular activities.

D. A dataset with missing values for most of the entries related to study hours, attendance, and extracurricular activities.

Please select the most appropriate option and provide reasoning for your choice.

B

D

C

A

Module 4: Hypothesis in machine learning

What is a Hypothesis in Machine Learning?

Quiz

1. **What does the hypothesis (h (x)) represent in machine learning?**

Activity description...

The features of the dataset.

The target variable to be predicted.

The ground truth or actual output of the training data.

The predicted output based on the input features and model parameters.

Hypothesis space

Quiz

1. **True or False: A larger hypothesis space always leads to better model performance.**

True

False

2. **True or False: Overfitting occurs when the hypothesis space is too complex for the given data.**

False

True

3. **Scenario Question**

You are a data scientist working on a binary classification problem. You have tried two different models for the task. Model A uses a simple hypothesis space with a linear model, while Model B employs a more complex hypothesis space with a high-degree polynomial. After evaluating both models, you notice that Model B fits the training data almost perfectly, but its performance on new, unseen data is not as good. On the other hand, Model A generalizes better to unseen data.

Based on this scenario, which model is likely suffering from overfitting?

Model A, because it uses a linear hypothesis space.

Neither model is suffering from overfitting.

Model B, because it fits the training data almost perfectly.

Both models are suffering from overfitting.

Module 5: Industry applications

Real case of study

Quiz

1. **True/False: The primary goal of the case study is to understand the practical aspects of applying machine learning techniques to predict and mitigate customer churn.**

False

True

2. **True/False: The dataset for this case study includes customer demographics, usage patterns, contract details, and customer service interactions.**

True

False