I. Introduction to Machine Learning

- Fundamental concepts of Machine Learning
  - Define the concept and characteristics of supervised and unsupervised learning

Supervised Learning

Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or "supervise" algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.
Supervised learning can be separated into two types of problems when data mining: classification and regression:

- **Classification** problems use an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Or, in the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox. Linear classifiers, support vector machines, decision trees and random forest are all common types of classification algorithms.
- **Regression** is another type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables. Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.

Basically supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

Supervised learning is classified into two categories of algorithms:

- **Classification**: A classification problem is when the output variable is a category, such as "Red" or "blue" , "disease" or "no disease".
- **Regression**: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

Supervised learning deals with or learns with "labeled" data. This implies that some data is already tagged with the correct answer.
Types:

- Regression
- Logistic Regression
- Classification
- Naive Bayes Classifiers
- K-NN (k nearest neighbors)
- Decision Trees

- Support Vector Machine

Advantages:

- Supervised learning allows collecting data and produces data output from previous experiences.
- Helps to optimize performance criteria with the help of experience.
- Supervised machine learning helps to solve various types of real-world computation problems.
- It performs classification and regression tasks.
- It allows estimating or mapping the result to a new sample.
- We have complete control over choosing the number of classes we want in the training data.

Disadvantages:

- Classifying big data can be challenging.
- Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
- Supervised learning cannot handle all complex tasks in Machine Learning.
- Computation time is vast for supervised learning.
- It requires a labelled data set.
- It requires a training process.

Unsupervised Learning

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are “unsupervised”).
Unsupervised learning models are used for three main tasks: clustering, association, and dimensionality reduction:

- **Clustering** is a data mining technique for grouping unlabeled data based on their similarities or differences. For example, K-means clustering algorithms assign similar data points into groups, where the K value represents the size of the grouping and granularity. This technique is helpful for market segmentation, image compression, etc.
- **Association** is another type of unsupervised learning method that uses different rules to find relationships between variables in a given dataset. These methods are frequently used for market basket analysis and recommendation engines, along the lines of “Customers Who Bought This Item Also Bought” recommendations.
- **Dimensionality reduction** is a learning technique used when the number of features (or dimensions) in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity. Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

Unsupervised learning is classified into two categories of algorithms:

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Types of Unsupervised Learning:

Clustering

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

Clustering Types:

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis

Advantages of unsupervised learning:

- It does not require training data to be labeled.
- Dimensionality reduction can be easily accomplished using unsupervised learning.
- Capable of finding previously unknown patterns in data.
- **Flexibility**: Unsupervised learning is flexible in that it can be applied to a wide variety of problems, including clustering, anomaly detection, and association rule mining.
- **Exploration**: Unsupervised learning allows for the exploration of data and the discovery of novel and potentially useful patterns that may not be apparent from the outset.
- **Low cost**: Unsupervised learning is often less expensive than supervised learning because it doesn’t require labeled data, which can be time-consuming and costly to obtain.

Disadvantages of unsupervised learning :

- Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
- The results often have lesser accuracy.
- The user needs to spend time interpreting and label the classes which follow that classification.
- **Lack of guidance**: Unsupervised learning lacks the guidance and feedback provided by labeled data, which can make it difficult to know whether the discovered patterns are relevant or useful.
- **Sensitivity to data quality**: Unsupervised learning can be sensitive to data quality, including missing values, outliers, and noisy data.
- **Scalability**: Unsupervised learning can be computationally expensive, particularly for large datasets or complex algorithms, which can limit its scalability.

- Define the concept of the probabilistic model

Probabilistic modeling is a statistical technique used to take into account the impact of random events or actions in predicting the potential occurrence of future outcomes.
Based on the fact that randomness or uncertainty plays a role in predicting outcomes, predictive modeling is used in a wide variety of fields and disciplines, from predicting the weather to potential nuclear fallout.

Probabilistic modeling is a statistical approach that uses the effect of random occurrences or actions to forecast the possibility of future results. It is a quantitative modeling method that projects several possible outcomes that might even go beyond what has happened recently.
Probabilistic modeling considers new situations and a wide range of uncertainty while not underestimating dangers. The three primary building blocks of probabilistic modeling are adequate probability distributions, correct use of input information for these distribution functions, and proper accounting for the linkages and interactions between variables. The downside of the probabilistic modeling technique is that it needs meticulous development, a process that depends on several assumptions and large input data.

Probabilistic models are an essential component of machine learning, which aims to learn patterns from data and make predictions on new, unseen data. They are statistical models that capture the inherent uncertainty in data and incorporate it into their predictions. Probabilistic models are used in various applications such as image and speech recognition, natural language processing, and recommendation systems. In recent years, significant progress has been made in developing probabilistic models that can handle large datasets efficiently.

- Explain the differences between supervised and unsupervised learning

Supervised machine learning relies on labelled input and output training data, whereas unsupervised learning processes unlabelled or raw data.

To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.

In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately

- **Goals:** In supervised learning, the goal is to predict outcomes for new data. You know up front the type of results to expect. With an unsupervised learning algorithm, the goal is to get insights from large volumes of new data. The machine learning itself determines what is different or interesting from the dataset.
- **Applications**: Supervised learning models are ideal for spam detection, sentiment analysis, weather forecasting and pricing predictions, among other things. In contrast, unsupervised learning is a great fit for anomaly detection, recommendation engines, customer personas and medical imaging.
- **Complexity:** Supervised learning is a simple method for machine learning, typically calculated through the use of programs like R or Python. In unsupervised learning, you need powerful tools for working with large amounts of unclassified data. Unsupervised learning models are computationally complex because they need a large training set to produce intended outcomes.
- **Drawbacks**: Supervised learning models can be time-consuming to train, and the labels for input and output variables require expertise. Meanwhile, unsupervised learning methods can have wildly inaccurate results unless you have human intervention to validate the output variables.

Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.

In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
Supervised learning can be categorized in **Classification** and **Regression** prob lems.	Unsupervised Learning can be classified in **Clustering** and **Associations** probl ems.
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.
Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

The distinction between supervised and unsupervised learning depends on whether the learning algorithm uses pattern-class information. Supervised learning assumes the availability of a teacher or supervisor who classifies the training examples into classes, whereas unsupervised learning must identify the pattern-class information as a part of the learning process.

Supervised learning algorithms utilize the information on the class membership of each training instances. This information allows supervised learning algorithms to detect pattern misclassifications as a feedback to themselves. In unsupervised learning algorithms, unlabeled instances are used. They blindly or heuristically

process them. Unsupervised learning algorithms often have less computational complexity and less accuracy than supervised learning algorithms.

- Identify the difference between the concepts of regression and Classification

Regression helps predict a continuous quantity, classification predicts discrete class labels. There are also some overlaps between the two types of machine learning algorithms.

- A regression algorithm can predict a discrete value which is in the form of an integer quantity
- A classification algorithm can predict a continuous value if it is in the form of a class label probability

Regression Algorithms	Classification Algorithms
The output variable must be either continuous nature or real value.	The output variable has to be a discrete value.
The regression algorithm's task is mapping input value (x) with continuous output variable (y).	The classification algorithm's task mapping the input value of x with the discrete output variable of y.
They are used with continuous data.	They are used with discrete data.
It attempt to find the best fit line, which predicts the output more accurately.	Classification tries to find the decision boundary, which divides the dataset into different classes.
Regression algorithms solve regression problems such as house price predictions and weather predictions.	Classification algorithms solve classification problems like identifying spam e-mails, spotting cancer cells, and speech recognition.
We can further divide Regression algorithms into Linear and Non-linear Regression.	We can further divide Classification algorithms into Binary Classifiers and Multi-class Classifiers.

	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can

	divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.

Classification	**Regression**
In this problem statement, the target variables are discrete.	In this problem statement, the target variables are continuous.
Problems like Spam Email Classification, Disease prediction like problems are solved using Classification Algorithms.	Problems like House Price Prediction, Rainfall Prediction like problems are solved using regression Algorithms.
In this algorithm, we try to find the best possible decision boundary which can separate the two classes with the maximum possible separation.	In this algorithm, we try to find the best-fit line which can represent the overall trend in the data.
Evaluation metrics like Precision, Recall, and F1-Score are used here to evaluate the performance of the classification algorithms.	Evaluation metrics like Mean Squared Error, R2-Score, and MAPE are used here to evaluate the performance of the regression algorithms.
Here we face the problems like binary Classification or Multi-Class Classification problems.	Here we face the problems like Linear Regression models as well as non-linear models.
Input Data are Independent variables and categorical dependent variable.	Input Data are Independent variables and continuous dependent variable.
Output is Categorical labels.	Output is Continuous numerical values.
Objective is to Predict categorical/class labels.	Objective is to Predicting continuous numerical values.
Example use cases are Spam detection, image recognition, sentiment analysis	Example use cases are Stock price prediction, house price prediction, demand forecasting.

References:

https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

https://www.geeksforgeeks.org/supervised-unsupervised-learning/

https://www.appsflyer.com/glossary/probabilistic-modeling/#:~:text=Probabilistic%20modeling%20is%20a%20statistical,potential%20occurrence%20of%20future%20outcomes.

https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-are-probabilistic-models#what_is_probabilistic_modeling

https://www.geeksforgeeks.org/probabilistic-models-in-machine-learning/

https://www.seldon.io/supervised-vs-unsupervised-learning-explained#:~:text=The%20main%20difference%20between%20supervised,processes%20unlabelled%20or%20raw%20data.

https://www.javatpoint.com/difference-between-supervised-and-unsupervised-learning

https://www.geeksforgeeks.org/difference-between-supervised-and-unsupervised-learning/

https://www.springboard.com/blog/data-science/regression-vs-classification/#:~:text=The%20most%20significant%20difference%20between,types%20of%20machine%20learning%20algorithms.

https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article#the_difference_between_regression_vs_classification

https://www.javatpoint.com/regression-vs-classification-in-machine-learning

https://www.geeksforgeeks.org/ml-classification-vs-regression/