MACHINE LEARNING LAB-CSP-317

1. What is machine learning?

   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 algorithm use historical data as input to predict new output values.

2. Features of Machine Learning

   i. It is a data-driven technology.
   ii. It can learn from past data and improve automatically.
   iii. It uses data to detect various patterns in a given dataset.
   iv. It is much similar to data mining as it also deals with the huge amount of the data.

3. Types of Machine Learning

   1. Supervised learning
   2. Unsupervised learning
   3. Semi supervised learning
   4. Reinforcement learning

   1. **Supervised learning :** Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on the basis, it predicts the output.

      The goal of supervised learning is to map input data with the output data.

      Supervised learning can be grouped further into two categories pf algorithms :

      i. Classification
      ii. Regression

      Classification : Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as Yes or No, Male or Female etc.

      The classification algorithms predict the categories present in the dataset.

      *Real world examples :

      - Spam Detection
      - Email filtering

      Classification algorithms :-

      1. Random Forest Algorithms
      2. Decision Tree Algorithms
      3. Logistic Regression Algorithm
      4. Support Vector Machine Algorithms

Regression : Regression are used to solve regression problems in which there is a linear relationship between input and output variables.

These are used to predict continuous output variables.

*Real world examples :

- Market trends
- Weather prediction

Regression algorithms :-

1. Simple Linear Regression Algorithm
2. Multivariate Regression Algorithm
3. Decision Tree Algorithm
4. Lasso Regression

Advantages of supervised learning

1. Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
2. These algorithms are helpful in predicting the output on the basis of prior experience.

Disadvantages of supervised learning

1. These algorithms are not able to solve complex tasks.
2. It may predict the wrong output if the test data is different from the training data.
3. It requires lots of computational time to train the algorithms.

Applications of supervised learning

1. Image Segmentation
2. Medical Diagnosis
3. Fraud Detection
4. Spam Detection
5. Speech Recognition

2. **Unsupervised learning :** Unsupervised learning is a learning method in which a machine learns without any supervision.

   The training is provided to the machine with the set of data that has not been labeled, classified or categorized, and the algorithm need to act on that data without any supervision.

   The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

   It can be further classifieds into two categories of algorithms :

   i. Clustering
   ii. Association

Clustering : The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups.

*An example of the clustering algorithm is grouping the customers by their purchasing behavior.

Clustering algorithms :-

1. K-Means Clustering Algorithm
2. Means-Shift Algorithm
3. DBSCAN Algorithm
4. Principle Component Analysis
5. Independent Component Analysis

Association : Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit.

*This algorithm is mainly applied in Market Basket analysis, Web usage mining, continuous production etc.

Association algorithms :-

1. Apriori Algorithm
2. Eclat
3. FP-growth Algorithm

Advantages of Unsupervised learning

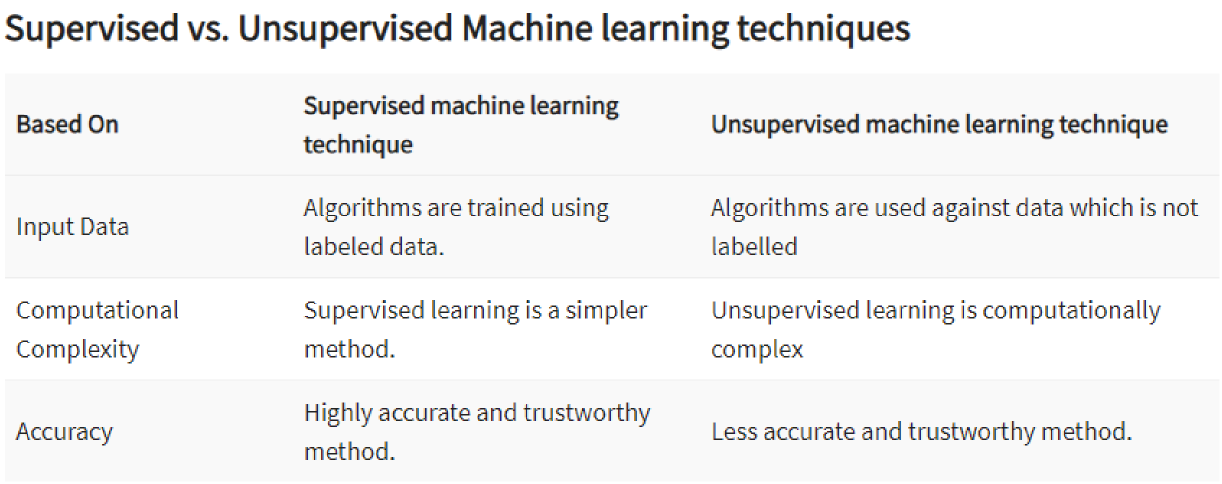
1. Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labeled dataset.

Disadvantages of Unsupervised learning

1. Working with Unsupervised learning is more difficult as it works with the unlabeled dataset that does not map with the output.

Applications of Unsupervised learning

1. Network Anlysis
2. Recommendation Systems
3. Anomaly Detection
4. Singular Value Decomposition



Supervised vs. Unsupervised Machine learning techniques

Based On	Supervised machine learning technique	Unsupervised machine learning technique
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labelled
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Accuracy	Highly accurate and trustworthy method.	Less accurate and trustworthy method.

3. **Semi-Supervised Learning**

Semi-Supervised learning is a type of Machine Learning algorithm that represents the intermediate ground between Supervised and Unsupervised learning algorithms. It uses the combination of labeled and unlabeled datasets during the training period.

*Real-world applications

1. Speech Analysis
2. Web content classification
3. Text document classifier
4. Protein sequence classification

*Difference between Semi-supervised and Reinforcement Learning.

Reinforcement learning is different from semi-supervised learning, as it works with rewards and feedback. ***Reinforcement learning aims to maximize the rewards by their hit and trial actions, whereas in semi-supervised learning, we train the model with a less labeled dataset.***

4. **Reinforcement Learning**

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

Types of Reinforcement learning

i. Positive Reinforcement

ii. Negative Reinforcement

Applications of Reinforcement learning

1. Robotics
2. Control
3. Game Playing
4. Chemistry

1. **Data Set - A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table.

2. **Model -** A data structure that store representation of A data set models are creative when you train a algorithm on a data set.

3. **Noise -** Any irrelevant information is called Noise.

4. **Outliner -** An observation that significantly from other observation in the data set.

5. **Desk Set -** A set of observation used at the end of model training and validation to access. The predictive power of your model.

6. **Training -** A set of observation used to generate machine learning model.

7. **Data collection -** Collect the data that the algorithm will learn from.

8. **Data preparation -** Format the data into the optical form.

9. **Evaluation -** Test the model to see how it will perform.

10. **Tuning -** Fine tuning the model to maximize its performance.

11. **Features -** With respect to a data set a feature and attributes a value combination. For example – Color is a attribute and blue color is a feature.

12. **Data Analysis -** Data Analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information by informing conclusions and supporting decision making.

13. **Data Visualization -** Data visualization is the graphical representation of information and data in a pictorial or graphical format(Example: charts, graphs, and maps). Data visualization tools provide an accessible way to see and understand trends, patterns in data, and outliers. Data visualization tools and technologies are essential to analyzing massive amounts of information and making data-driven decisions.

14. **Seaborn -** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

15. **Pandas -** Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

16. **NumPy -** Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

17. **Datapoints -** Values are organised in structures called datapoints.

18. **Linear regression**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

19. **Support Vector Machine**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Types of SVM

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

SVM algorithm can be used for Face detection, image classification, text categorization, **etc.**

20. **Naïve Bayes Classifier Algorithm**

1. Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
2. It is mainly used in *text classification* that includes a high-dimensional training dataset.
3. It helps in building the fast machine learning models that can make quick predictions.
4. **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
5. Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

21. K-Nearest Neighbor(KNN) Algorithm for Machine Learning

1. K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
2. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
3. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
4. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
5. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
6. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
7. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
8. **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure.

22. **Decision Tree Classification Algorithm**

    1. Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
    2. In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
    3. The decisions or the test are performed on the basis of features of the given dataset.
    4. ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***

23. **Random Forest Algorithm**

    Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

    As the name suggests, **"*Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.*"**

24. **Bias :** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.

25. **Variance :** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

26. **Overfitting**

    Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.

    The overfitted model has **low bias** and **high variance.**

    Overfitting is the main problem that occurs in supervised learning.

27. **Underfitting**

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

An underfitted model has high bias and low variance.

28. **K-Means Clustering Algorithm**

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

29. **Principal component analysis**

Principal component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

30. **Association rule**

Association rule mining Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.