Task 5

Supervised/ Unsupervised Models

Machine learning

Machine learning is a subset of AI, in which the algorithms are 'trained' and learn from their past experiences and examples.

Supervised learning

Supervised learning makes use of regression analysis and classification analysis. It is used to predict future outcomes based on past data. It requires both input and output values to be used in the training process.

- The system requires both an input and an output to be given to the model so it can be trained.
- The model uses labelled data, so the desired output for a given input is known.
- Algorithms receive a set of inputs and the correct outputs to permit the learning process.
- Once trained, the model is run using labelled data.
- The results are compared with the expected output; if there are any errors, the model needs further refinement.
- The model is run with unlabelled data to predict the outcome. An example of supervised learning is categorising emails as relevant or spam/ junk without human intervention.

Types of Supervised Learning:

Supervised learning in machine learning is generally divided into two categories: classification and regression.

Regression

Regression is used to make predictions from given data by learning some relationship between the input and the output. It helps in the understanding of how the value of a dependent variable changes when the values of independent variables are also changed. This makes it a valuable tool in prediction applications, such as weather forecasting. In machine learning, this is used to predict the outcome of an event based on any relationship between variables obtained from input data and the hidden parameters.

A common example of a regression task might be predicting a salary based on work experience. For instance, a supervised learning algorithm would be fed inputs related to work experience (e.g., length of time, the industry or field, location, etc.) and the corresponding assigned salary amount, as illustrated below.

Classification

Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data. For instance, an algorithm can learn to predict whether a given email is spam or ham (no spam)

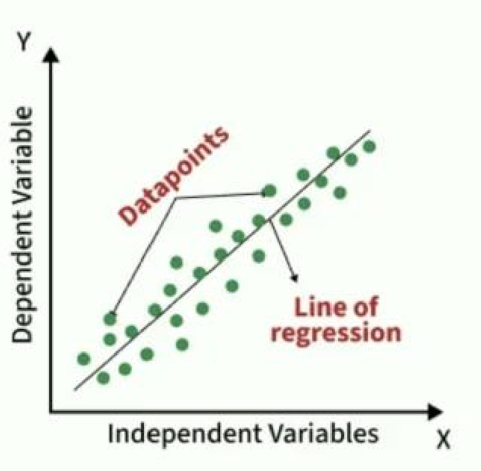
1. Linear Regression

Linear regression is a type of supervised machine-learning algorithm that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets.

$$Y=\beta 0 \quad +\beta 1 \quad X1 \quad +\beta 2 \quad X2 \quad +\cdots+\beta n \quad Xn \quad +\epsilon$$

It assumes that there is a linear relationship between the input and output. rThis relationship is represented by a straight line.

- Independent variable (input): Hours studied because it's the factor we control or observe.
- Dependent variable (output): Exam score because it depends on how many hours were studied.



2. Logistic Regression

Logistic Regression is used for classification problems. It is used for binary classification where the output can be one of two possible categories such as Yes/No, True/False or 0/1. It uses sigmoid function to convert inputs into a probability value between 0 and 1.

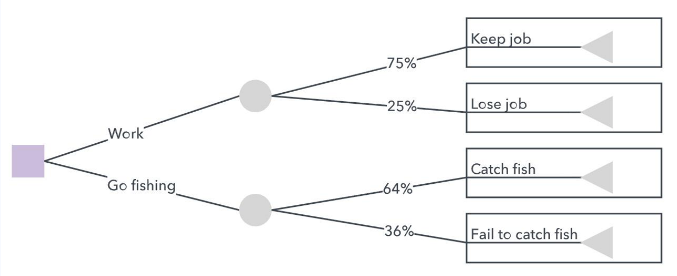
$$P(Y=1 \mid X)=\frac{1}{1+e-(\beta 0 \quad +\beta 1 \quad X1 \quad +\ldots+\beta n \quad Xn \quad)}$$

Logistic regression can be classified into three main types based on the nature of the dependent variable:

1. Binomial Logistic Regression: This type is used when the dependent variable has only two possible categories. Examples include Yes/No, Pass/Fail or 0/1.
2. Multinomial Logistic Regression: This is used when the dependent variable has three or more possible categories that are not ordered. For example, classifying animals into categories like "cat," "dog" or "sheep.".
3. Ordinal Logistic Regression: This type applies when the dependent variable has three or more categories with a natural order or ranking. Examples include ratings like "low," "medium" and "high." It takes the order of the categories into account when modeling.

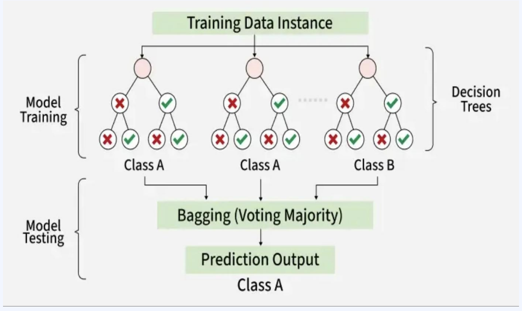
3. Decision Tree

It works by splitting data into branches based on decision rules, creating a tree-like structure where each internal node represents a test on a feature, each branch represents the outcome of that test, and each leaf node represents a final prediction or decision.



4. Random Forest

Random Forest is a machine learning algorithm that uses many decision trees to make better predictions. Each tree looks at different random parts of the data and their results are combined by voting for classification or averaging for regression. This helps in improving accuracy and reducing errors.

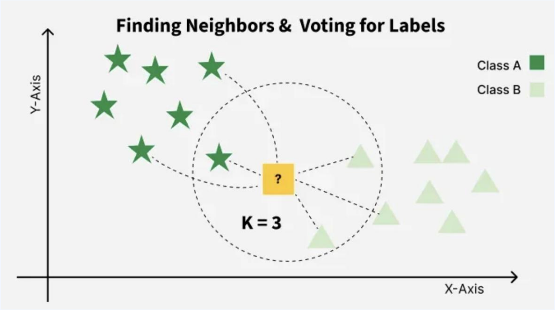


5. Support Vector Machine (SVM)

Support Vector Machine (SVM) tries to find the best boundary known as hyperplane that separates different classes in the data. It is useful when you want to do binary classification like spam vs. not spam or cat vs. dog.

6. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) works by finding the "k" closest data points (neighbors) to a given input and makesa predictions based on the majority class (for classification) or the average value (for regression). Since KNN makes no assumptions about the underlying data distribution it makes it a non-parametric and instance-based learning method.



7. Naive Bayes Classifier

The Naive Bayes classifier is a supervised machine learning algorithm designed to assign a label or category to an object based on its features. The algorithm analyzes a dataset with labeled examples (e.g., emails classified as “spam” or “not spam”) and calculates the probabilities of features for each target class. For new data (e.g., a new email), the algorithm uses the calculated probabilities to determine the most likely target class.

8. Gradient Boosting

Gradient Boosting is a ensemble learning method used for classification and regression tasks. It is a boosting algorithm which combine multiple weak learner to create a strong predictive model. It works by sequentially training models where each new model tries to correct the errors made by its predecessor.

Unsupervised learning

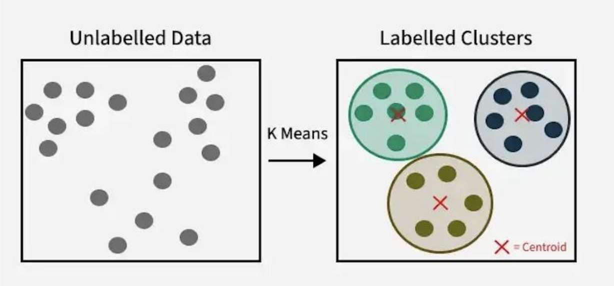
Systems are able to identify hidden patterns from the input data provided; they are not trained using the 'right' answer.
By making data more readable and more organised, patterns, similarities and anomalies will become evident (unsupervised learning makes use of density estimation and k-mean clustering; in other words, it classifies unlabelled real data).

Algorithms evaluate the data to find any hidden patterns or structures within the data set.

An example is used in product marketing: a group of individuals with similar purchasing behaviour are regarded as a single unit for promotions.

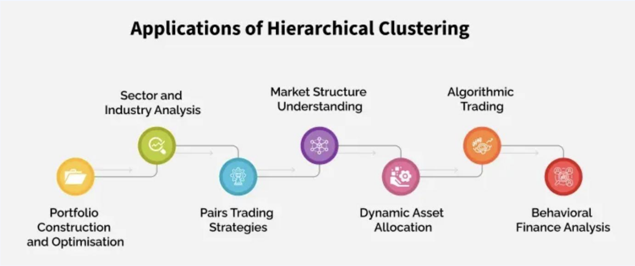
1. K-Means Clustering

K-Means Clustering groups unlabeled dataset into different clusters. It is used to organize data into groups based on their similarity. For example online store uses K-Means to group customers based on purchase frequency and spending creating segments like Budget Shoppers, Frequent Buyers and Big Spenders for personalised marketing.



2. Hierarchical clustering

Hierarchical clustering is used to group similar data points together based on their similarity creating a hierarchy or tree-like structure. The key idea is to begin with each data point as its own separate cluster and then progressively merge or split them based on their similarity.


Applications of Hierarchical Clustering
Sector and Industry Analysis
Market Structure Understanding
Algorithmic Trading
Portfolio Construction and Optimisation
Pairs Trading Strategies
Dynamic Asset Allocation
Behavioral Finance Analysis

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups data points that are closely packed together and marks outliers as noise based on their density in the feature space. It identifies clusters as dense regions in the data space separated by areas of lower density.

4. Principal Component Analysis (PCA)

PCA (Principal Component Analysis) is a dimensionality reduction technique used in data analysis and machine learning. It helps you to reduce the number of features in a dataset while keeping the most important information. It changes your original features into new features these new features don’t overlap with each other and the first few keep most of the important differences found in the original data.

5. Independent Component Analysis (ICA)

Independent Component Analysis (ICA) is a technique used to separate mixed signals into their independent, non-Gaussian components. Its aim to find a linear transformation of data that maximizes statistical independence among the components. ICA is widely used in fields like audio, image processing and biomedical signal analysis to isolate distinct sources from mixed signals.

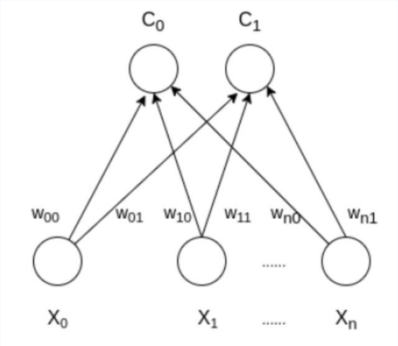
6. Autoencoders (Neural Networks)

Autoencoders are a special type of neural networks that learn to compress data into a compact form and then reconstruct it to closely match the original input. They consist of an:

- Encoder that captures important features by reducing dimensionality.
- Decoder that rebuilds the data from this compressed representation.

7. Self-Organizing Maps (SOM)

A Self Organizing Map (SOM) is an unsupervised neural network algorithm based on biological neural models from the 1970s. It uses a competitive learning approach and is primarily designed for clustering and dimensionality reduction. SOM effectively maps high-dimensional data to a lower-dimensional grid enabling easier interpretation and visualization of complex datasets.



8. Gaussian Mixture Model

Gaussian Mixture Model is a probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions with unknown parameters. Unlike hard clustering methods like K-Means that assigns each data point to one cluster based on the closest centroid which often doesn't align with the complexity of real-world data. GMMs perform soft clustering meaning each data point belongs to multiple clusters with certain probabilities. Each Gaussian in the mixture is defined by:

- Mean (μ): The center of the distribution.
- Covariance (Σ): Describes the spread and orientation.
- Mixing coefficient (π): Represents the proportion of each Gaussian in the mixture.