Machine Learning - Module 1

What is Machine Learning?

- Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

- Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning

Types of Machine Learning

1. Supervised : We train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).

   Categories - Regression , Classification.

   Advantages: These algorithms are helpful in predicting the output on the basis of prior experience.

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

   Applications of Supervised Learning : Image Segmentation, Medical Diagnosis, Fraud Detection, Spam detection, Speech Recognition

2) Unsupervised : It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision. In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.

Categories : Clustering, Association

Advantages: These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.

Disadvantages: The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior. Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

Applications of Unsupervised Learning: Network Analysis, Recommendation Systems, Anomaly Detection, Singular Value Decomposition

3) Semi-Supervised learning : It is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced.

Advantages: It is simple and easy to understand the algorithm.
It is highly efficient.
It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages: Iterations results may not be stable.
We cannot apply these algorithms to network-level data.
Accuracy is low.

4) Reinforcement learning : It works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.

Categories: Positive Reinforcement Learning, Negative Reinforcement Learning

Advantages: It helps in solving complex real-world problems which are difficult to be solved by general techniques.
Helps in achieving long term results.

Disadvantage: RL algorithms are not preferred for simple problems.
RL algorithms require huge data and computations.

Applications: Video Games, Resource Management, Robotics, Text Mining

Life Cycle of Machine Learning

1. Gathering Data : Identify various data sources → Collect data → Integrate the data obtained from different sources
2. Data preparation: a) Data exploration: It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data. A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.

   b) Data pre-processing: Now the next step is preprocessing of data for its analysis.
3. Data Wrangling: Data wrangling is the process of cleaning and converting raw data into a useable format. Removing Missing Values, Duplicate data, Invalid data, Noise

4) Analyse Data: The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.
This involves Selection of analytical techniques, Building models, Reviewing the result.
Then we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.

5) Train the model: Then, we train our model to improve its performance for better outcome of the problem. We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

6) Test Model: Then, we check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7) Deployment: If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not.

Weight, Bias and Curse of Dimensionality

- Weights: Weights are the real values that are attached with each input/feature. A weight decides how much influence the input will have on the output.
- Bias: bias represents how far off the predictions are from their intended value. Biases make up the difference between the function's output and its intended output.
- Curse of Dimensionality: An increase in the dimensions can in theory, add more information to the data thereby improving the quality of data but practically increases the noise and redundancy during its analysis. As the dimensionality increases, the number of data points required for good performance of any machine learning algorithm increases exponentially.

Testing Machine Learning Models

- Confusion Matrix: The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. The matrix is divided into two dimensions, that are predicted values and actual values along with the total number of predictions.
-

n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

True Negative: Model has given prediction No, and the real or actual value was also No.
True Positive: The model has predicted yes, and the actual value was also true.
False Negative: The model has predicted no, but the actual value was Yes, it is also called as Type-II error.
False Positive: The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

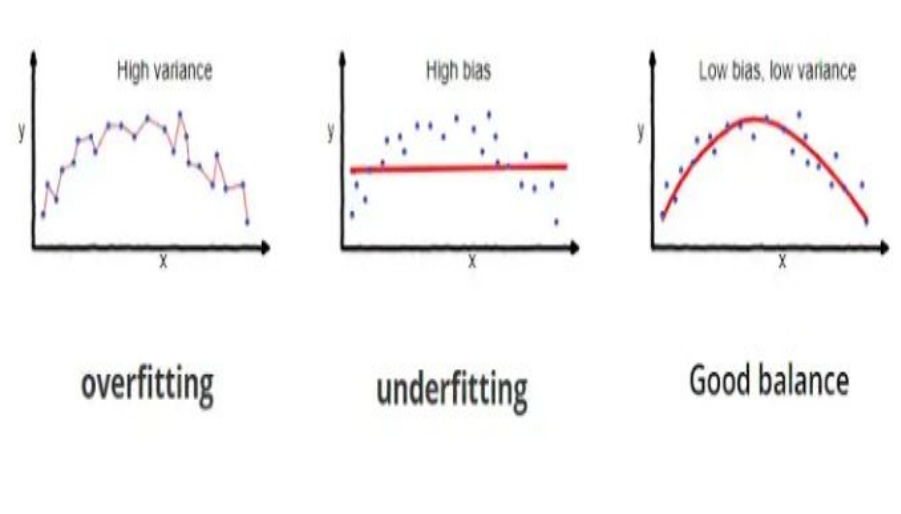
- Cross-Validation: Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

- Steps: Reserve a subset of the dataset as a validation set → Provide the training to the model using the training dataset → Evaluate model performance using the validation set. If the model performs well with the validation set, perform the further step, else check for the issues.

- Types: Validation Set Approach.
  Leave-P-out cross-validation
  Leave one out cross-validation
  K-fold cross-validation
  Stratified k-fold cross-validation

Bias-Variance Tradeoff

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
- In supervised learning, underfitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance.
- In supervised learning, overfitting happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance.



If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

Naive Bayes’ Classifier

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems (especially text classification).

It makes quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. It is called Bayes because it depends on the principle of Bayes' Theorem.

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B

Advantages:
Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
It can be used for Binary as well as Multi-class Classifications.
It performs well in Multi-class predictions as compared to the other Algorithms.

Disadvantages:
Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.