UNIT I
Introduction to Machine Learning

Machine learning is programming computers to optimize a performance criterion using example data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data, or both. Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term “Machine Learning” in 1959 while at IBM. He defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed.”

The term "Machine Learning" was first coined by Arthur Samuel in 1959. He defined it as "the field of study that gives computers the ability to learn without being explicitly programmed.”

Definition of learning

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E.

Examples:

i) Handwriting recognition learning problem

- Task T: Recognising and classifying handwritten words within images
- Performance P: Percent of words correctly classified
- Training experience E: A dataset of handwritten words with given classifications

i) A chess learning problem

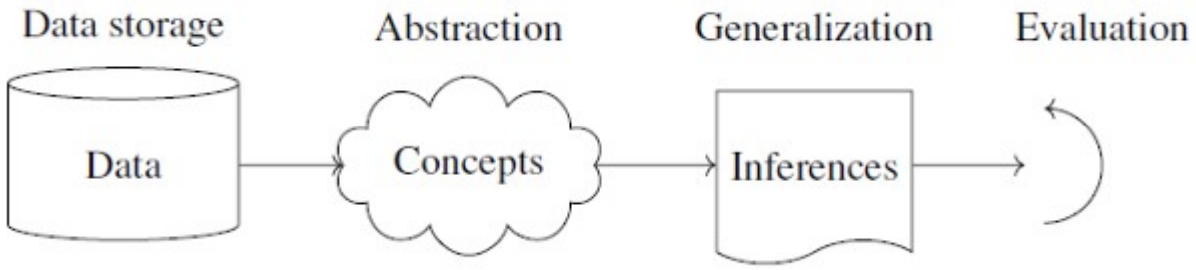
- Task T: Playing chess
- Performance measure P: Percent of games won against opponents
- Training experience E: Playing practice games against itself

A computer program which learns from experience is called a machine learning program or simply a learning program. Such a program is sometimes also referred to as a learner.

Components of Learning

Basic components of learning process

The learning process, whether by a human or a machine, can be divided into four components, namely, data storage, abstraction, generalization and evaluation. Figure illustrates the various components and the steps involved in the learning process.



1. **Data storage** Facilities for storing and retrieving huge amounts of data are an important component of the learning process. Humans and computers alike utilize data storage as a foundation for advanced reasoning.

- In a human being, the data is stored in the brain and data is retrieved using electrochemical signals.
- Computers use hard disk drives, flash memory, random access memory and similar devices to store data and use cables and other technology to retrieve data.

2. **Abstraction** - The second component of the learning process is known as abstraction. Abstraction is the process of extracting knowledge about stored data. This involves creating general concepts about the data as a whole. The creation of knowledge involves application of known models and creation of new models. The process of fitting a model to a dataset is known as training. When the model has been trained, the data is transformed into an abstract form that summarizes the original information.
3. **Generalization** - The third component of the learning process is known as generalisation. The term generalization describes the process of turning the knowledge about stored data into a form that can be utilized for future action. These actions are to be carried out on tasks that are similar, but not identical, to those what have been seen before. In generalization, the goal is to discover those properties of the data that will be most relevant to future tasks.
4. **Evaluation -** Evaluation is the last component of the learning process. It is the process of giving feedback to the user to measure the utility of the learned knowledge. This feedback is then utilised to effect improvements in the whole learning process.

Applications of machine learning

Application of machine learning methods to large databases is called data mining. In data mining, a large volume of data is processed to construct a simple model with valuable use, for example, having high predictive accuracy.

The following is a list of some of the typical applications of machine learning.

1. In retail business, machine learning is used to study consumer behaviour.
2. In finance, banks analyze their past data to build models to use in credit applications, fraud detection, and the stock market.
3. In manufacturing, learning models are used for optimization, control, and troubleshooting.
4. In medicine, learning programs are used for medical diagnosis.
5. In telecommunications, call patterns are analyzed for network optimization and maximizing the quality of service.
6. In science, large amounts of data in physics, astronomy, and biology can only be analyzed fast enough by computers.
7. In artificial intelligence, it is used to teach a system to learn and adapt to changes so that the system designer need not foresee and provide solutions for all possible situations.

Data set

Data set is the collection of data used for machine learning. Basically, the dataset is divdided into three categories. They are training data, testing Data and validation Data. Here, the training data is considered for initial training purpose. Testing data is used for checking the trained machine. Validation data is used for tuning the trained machine with the help of important parameters.

Four types of data are explained here, as it is often be handled in the process of dataset preparation or preprocessing. The data types are as given below.

1. **Numerical Data** : Numerical data is a datatype expressed in numbers. This further classified as continuous and discontinuous data.
2. **Categorical Data**: Categorical data is a collection of information that is divided into groups. They are further divided into two types such as ordinal and nominal.
3. **Time Series Data**: A Time Series is a sequence of data points that occur in successive order over some period of time
4. **Text Data**: Text data usually consists of documents, which can represent words, sentences or even paragraphs. Text data usually consists of documents, which can represent words, sentences or even paragraphs. Text data usually consists of documents, which can represent words, sentences or even paragraphs.

Networks Evaluation

Evaluating machine learning models is crucial for determining their effectiveness and reliability. This involves using evaluation metrics, which are quantitative measures that assess the model's performance. The choice of metrics depends on the type of problem, such as classification or regression. Some common evaluation metrics include:

Accuracy: The proportion of correct predictions out of the total predictions.
Precision: The proportion of positive predictions that were actually correct.
Recall (Sensitivity): The proportion of actual positive cases that were correctly identified.
Confusion Matrix: A table that summarizes the predictions of a classification model, showing true positives, true negatives, false positives, and false negatives.
ROC AUC Curve (Receiver Operating Characteristic - Area Under the Curve): A curve that illustrates the performance of a binary classifier system as its discrimination threshold is varied.
RMSE (Root Mean Squared Error): Commonly used for regression problems to measure the magnitude of the errors

The process of "learning" or training a machine learning model involves several stages:

Data Collection: Gathering a relevant and high-quality dataset is the foundational step.

Data Preprocessing and Cleaning: Raw data is often messy and needs to be cleaned, handled for missing values, transformed, and potentially standardized to improve the model's performance.

Feature Engineering: This involves transforming raw data into meaningful features or attributes that the model can learn from. Feature selection techniques are used to identify and select the most relevant subset of features to optimize performance and reduce complexity.

Model Selection: Choosing the appropriate machine learning algorithm (e.g., linear regression, decision trees, neural networks) based on the problem type and data characteristics.

Model Training: The selected model is trained on the preprocessed data, allowing it to adjust its internal parameters to identify patterns and relationships that map input features to target outputs. This fitting process involves using a loss function to measure errors and an optimization technique (like gradient descent) to minimize those errors.

Model Evaluation and Tuning: The model's performance is evaluated using metrics on a validation set, and adjustments are made to hyperparameters (external configuration variables) to improve accuracy and generalization. Techniques like cross-validation (discussed below) are crucial during this stage.

Cross-validation

Cross-validation is a robust technique used during model evaluation and tuning, particularly for assessing how well a model generalizes to unseen data and preventing overfitting. Instead of a single train-test split, cross-validation involves:

1.Partitioning the data into multiple subsets (folds).

2.Training and testing the model multiple times, using different folds for training and testing in each iteration.

3.Aggregating the results (e.g., averaging performance metrics) to provide a more reliable estimate of the model's performance and to make hyperparameter tuning decisions