Introduction to Machine Learning: Types of Machine Learning

Prof. (Dr.) Honey Sharma

Reference Books:

- Alpaydin E., Introduction to Machine Learning, MIT Press (2014) 3rd Edition.
- https://www.geeksforgeeks.org/

ISSUES WITH MACHINE LEARNING

Lack of Training Data:

- The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. Less amount training data will produce inaccurate or too biased predictions.
- You decided to explain to a child how to distinguish between an apple and a watermelon. You will take an apple and a watermelon and show him the difference between both based on their color, shape, and taste. If we leave any one of the feature then there wil be great chance of mistake

Poor Quality of Data:

- Absence of good quality data.
- In the absence of good quality data our algorithm may produce inaccurate or faulty predictions. **The quality of data is essential to enhance the output.**
- we need to ensure that the process of **data preprocessing which includes removing outliers, filtering missing values, and removing unwanted features**, is done with the utmost level of perfection.

Underfitting of Training Data:

- This process occurs when data is unable to establish an accurate relationship between input and output variables.
- To overcome this issue:
  - Enhance the complexity of the model
  - Add more features to the data
  - Reduce regular parameters
  - Increasing the training time of model

Overfitting of Training Data:

- Overfitting happens when a model learns the **detail and noise in the training data** to the extent that it negatively impacts the performance of the model on new data.
- Let’s consider a model trained to differentiate between a cat, a rabbit, a dog, and a tiger. The training data contains 1000 cats, 1000 dogs, 1000 tigers, and 9000 Rabbits. Then there is a considerable probability that it will identify the cat as a rabbit.

Machine Learning is a Complex Process:

- The process is transforming, and hence there are high chances of error which makes the learning complex. It includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, and a lot more.

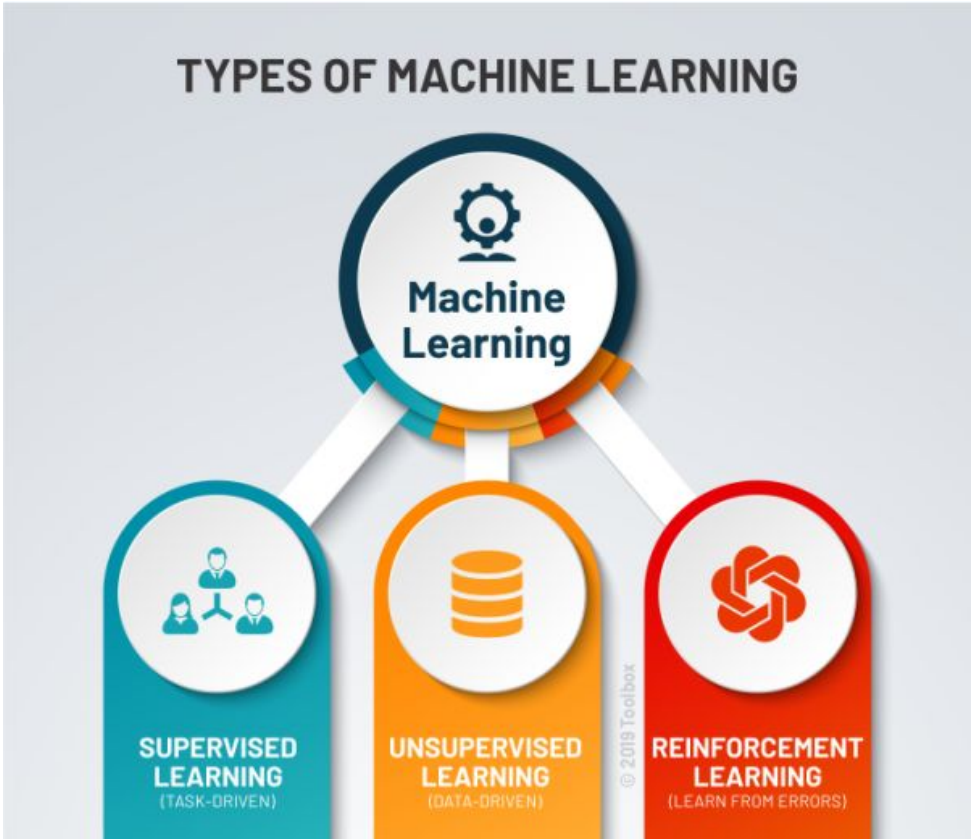
Slow Implementation:

This is one of the common issues faced by machine learning professionals. The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results.

Imperfections in the Algorithm When Data Grows:

Lets you have found quality data, trained it amazingly, and the predictions are really concise and accurate. You have learned how to create a machine learning algorithm!! But, the model may become useless in the future as data grows. The best model of the present may become inaccurate in the coming future and require further rearrangement.

Types of Machine Learnings



Supervised Learning

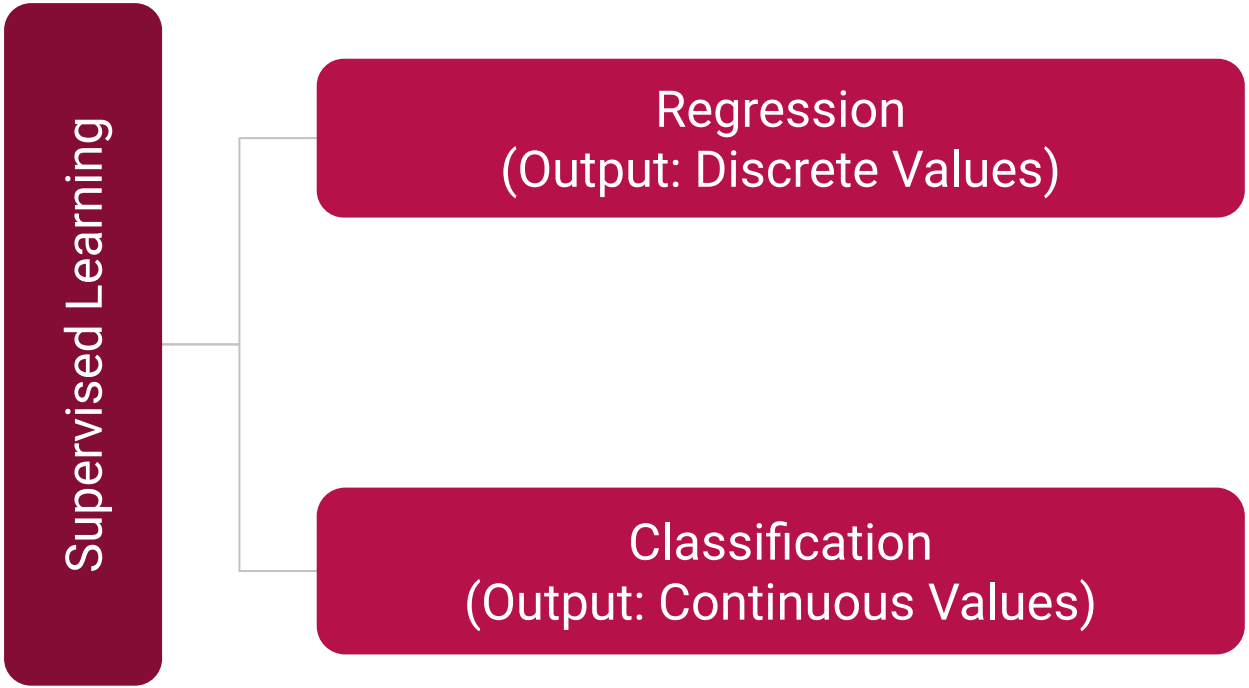
Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is **trained on labeled data**.

In this type of learning both training and validation, datasets are labelled.

The algorithm then **finds relationships between the parameters given (Fits data to a function) Regression Analysis**

This solution is then deployed for use with the final dataset, which it learns from in the same way as the training dataset.

This means that supervised machine learning algorithms will continue to improve even after being deployed, discovering new patterns and relationships as it trains itself on new data.


Supervised Learning
Regression
(Output: Discrete Values)
Classification
(Output: Continuous Values)

Classification: It is a Supervised Learning task where output is having defined labels(discrete value). It can be either binary or multi-class classification. In binary classification, the model predicts either 0 or 1; yes or no but in the case of multi-class classification, the model predicts more than one class.

For example :Output – Purchased has defined labels i.e. 0 or 1; 1 means the customer will purchase and 0 means that customer won't purchase.

Example: Gmail classifies mails in more than one class like social, promotions, updates, forums.

Regression: It is a Supervised Learning task where output is having continuous value.

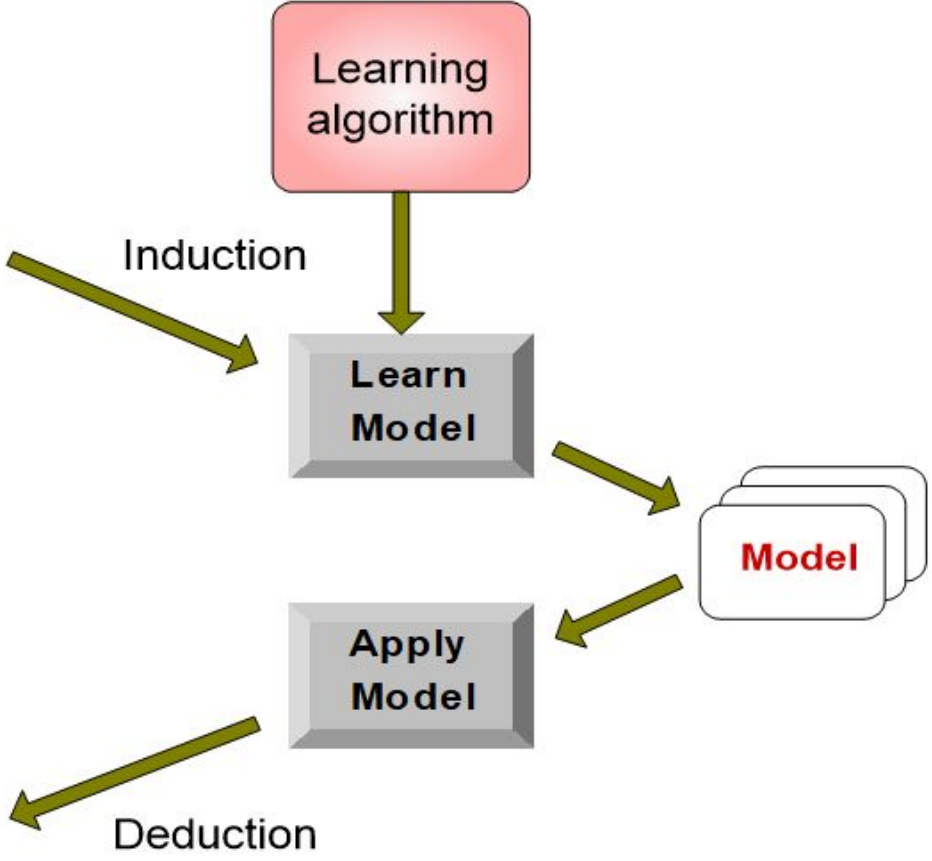
Output – Lets Wind speed depends on Temperature, Humidity, Pressure and Wind Direction. Wind Speed is not having any discrete value but is continuous in the particular range. The goal here is to predict a value as much closer to the actual output value as our model can and then evaluation is done by calculating the error value. The smaller the error the greater the accuracy of our regression model.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Un-Supervised Learning

In supervised learning, the aim is to learn a mapping from the input to an output whose correct values are provided by a supervisor.

In unsupervised learning, there is no such supervisor and we only have input data.

The aim is to find the regularities in the input. There is a structure to the input space such that certain patterns occur more often than others, and we want to see what generally happens and what does not.

For instance, suppose it is given an image having both dogs and cats which it has never seen.

Thus the machine has no idea about the features of dogs and cats so we can't categorize it as 'dogs and cats '.

But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having dogs in them and the second part may contain all pics having cats in them.

Here you didn't learn anything before, which means no training data or examples.

It allows the model to work on its own to discover patterns and information that was previously undetected. **It mainly deals with unlabelled data.**

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.

Learning Associations

A supermarket chain—one application of machine learning is basket analysis

Basket analysis is finding associations between products bought by customers:

If people who buy X typically also buy Y, and if there is a customer who buys X and does not buy Y, he or she is a potential Y customer.

Once we find such customers, we can target them for cross-selling

In finding an association rule, we are interested in learning a **conditional probability** of the form P (Y|X) where Y is the product we would like to condition on X, which is the product or the set of products which we know that the customer has already purchased.

Let us say, going over our data, we calculate that P (chips|colddrink) = 0.7. Then, we can define the rule: 70 percent of customers who buy colddrink also buy chips.

We may want to make a distinction among customers and toward this, estimate P (Y|X,D) where D is the set of customer attributes, for example, gender, age, marital status, and so on, assuming that we have access to this information

Reinforcement Learning

In some applications, **the output of the system is a sequence of actions**. In such a case, a single action is not important; what is important is the policy that is the sequence of correct actions to reach the goal.

There is no such thing as the best action in any intermediate state; an action is good if it is part of a good policy.

In such a case, the machine learning program should be able to assess the goodness of policies and learn from past good action sequences to be able to generate a policy.

Such learning reinforcement methods are called reinforcement learning algorithms.

Reinforcement learning can be used in robotics for industrial automation