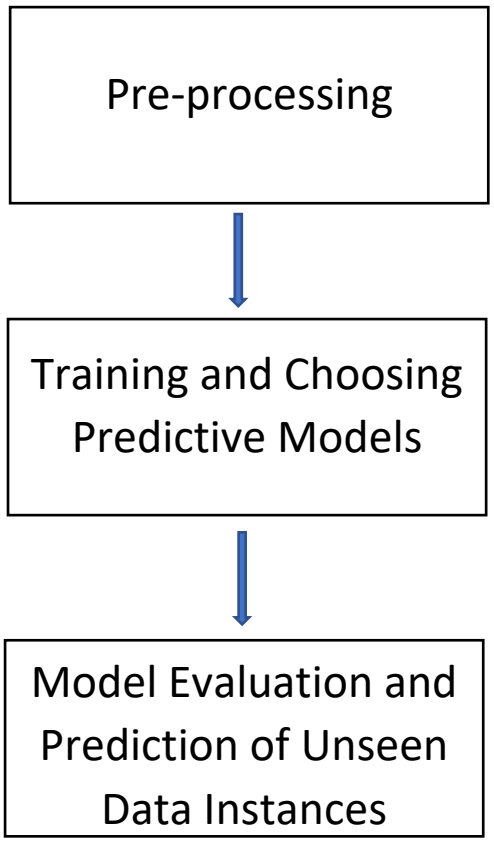
Introduction to Machine Learning

Machine Learning Systems- Typical Block Diagram



Three Types of Machine Learning

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

Supervised Learning – Predicting the Future

A supervised algorithm requires a detailed input of related data over a period of time. Once all the information is available to the computer, it is used to classify any new data relating to it.

The computer then does a series of calculation, comparisons, and analysis before it makes a decision.

This type of algorithm requires an extensive amount of information to be programmed into the system so that the computer can make the right decision. That way, when it needs to solve a problem, it will determine which mathematical function it needs to use in order to find the correct solution. With the right series of algorithms already programmed into the system, the machine can sit through all types of data in order to find the solution to a wide variety of problems in the related category.

Supervised algorithms are referred that way because they require human input to ensure that the computer has the right data to process the information it receives.

Supervised learning has one basic goal – to learn a specified model from "labelled" data for training purposes so that predictions can be made about future data or unseen data.

It is called "supervised" because the labels, or the output signals that we want, are known already.

Think about your email, about the spam filter in place. We could use supervised machine learning to train a model. The data would be labelled – emails already marked correctly as spam, those marked correctly as not spam – and the model would be trained to recognize to which group new emails belong. When you have supervised tasks such as this, using class labels, we call it a "classification task".

Predicting Class Labels using Classification

Classification is a subcategory with a goal of predicting under which class label, new instances would be placed. This is based on previous data. The labels are discrete with values in no order that can be understood to be memberships of the instances.

The example of spam filters is representative of a binary type of classification task – the algorithm will learn the rules so that it can work out the difference between potential classes – spam and not spam.

Although this is a binary classification task, the class labels are not required to be binary in nature. The learning algorithm will determine a predictive model that can then assign any of the labels from the dataset to a new instance that has no label.

One example of this type of multiclass classification is the recognition of hand written characters. We could have a dataset that has the letters of the alphabet written in several different hand writing styles. If the user were to use an input device to give a new character, the predictive model could predict which letter of the alphabet it was with accuracy.

However, the system could not recognize any digit from 0 to 9 because they would not be in the dataset used for training purposes.

Continuous Outcome Prediction using Regression

There is another type of supervised learning called regression analysis, and this is used for the prediction of continuous outcomes.

With regression analysis, we have several explanatory or predictor variables, along with a variable for a continuous response. We can predict outcomes by finding relationships between the variables.

Let assume we have a class of students and we want to try to predict their SAT scores. If we find a link between how long each student studied and their scores, we could then use that data to learn models that can use the length of time studied to predict the score achieved by future students.

Reinforcement Learning – Solving Interactive Problems

Reinforcement Learning is another area of machine learning and the goal here is to develop an agent or system that can improve its performance. This improvement is based on how the agent interacts with its environment. Information that details the current environment state, tends to include a rewards-signal, so this kind of machine learning can be considered a part of supervised learning.

However, the feedback in Reinforcement Learning is not the right value or truth label; rather, it is an indication of how the reward function related to the action. By interacting with its environment, the agent learns several actions to maximize the reward. By using Reinforcement Learning, the agent will go through an approach of either deliberate planning or trial and error to get the reward.

Reinforcement learning is commonly used in video games where the computer must navigate and adjust its movements in order to win the game.

A reward system is used so the computer knows and understands when it should make the right move, but there are also negative consequences whenever they make errors. This type of algorithm work best in situations where the computer has an obstacle that it must overcome like a rival in a game, or it could also be a self-driving car that needs to reach its destination. The entire focus of the computer is to accomplish certain tasks while navigating the unpredictable environment around it. With each mistake, the computer will readjust its moves in order to reduce the number of errors so it can achieve the desired result.

A chess engine is a great example of this type of machine learning. The agent will decide on the moves it will make, dependent on the environment (the board) and the reward is defined as winning or losing when the game is over.

Reinforcement Learning has a number of subtypes but, in general, the agent will carry out several environmental interactions in an attempt to get the maximum reward. Each of the states is associated with a negative or a positive reward – the reward is defined by achieving a specific goal, in this case, losing or winning the game of chess. For example, in a game of chess, the outcome of a move is a state of that environment.

Unsupervised Learning – Discovering Hidden Structures

With supervised learning, we already know the answer we want when the model is being trained.

With reinforcement learning, we provide a reward measured by specific actions performed by the agent.

With unsupervised learning, we are going one step further by using unlabelled data – data with unknown structure. By using these techniques, we can explore the data structure to get meaningful information without needing to be guided by a reward or by a variable with a known outcome.

An unsupervised algorithm implies that the computer does not have all the information to make a decision. Maybe it has some of the data needed but one or two factors may be missing. This is kind of like the algebra problems you encountered in school. You may have two factors in the problem but you must solve the third on your own.

A + b = c

If you know A but you have no idea what b is then you need to plug the information into an equation to solve the problem.

With unsupervised learning, this can be an extremely complex type of problem to solve. For this type of problem, you will need an algorithm that recognizes various elements of a problem and can incorporate that into the equation. Another type of algorithm will look for any inconsistencies in the data and try to solve the problem by analysing those.

Unsupervised algorithms clearly are much more complex than the supervised algorithms. While they may start with some data to solve a problem, they do not have all the information so they must be

equipped with the tools to find those missing elements without having a human to provide all the pieces of the puzzle for them.

Use Clustering to Find Subgroups

Clustering is a technique of exploratory data analysis which lets us organize a great amount of data into subgroups or clusters. We do not have any upfront knowledge about the group memberships. Each of the clusters identified from the analysis will define a specific group of objects that are similar to a certain degree but are not quite so similar to objects that are in the other clusters. This is why you often hear clustering being termed as “unsupervised classification”.

Clustering is absolutely one of the best techniques we have for providing a structure to as set of information and determining meaningful relationships from the information.

For example, think of marketers. They can use clustering to determine specific customer groups by their interests so that they can devise a targeted marketing program.

Dimensionality Reduction

This is another subfield of the unsupervised machine learning area. More often than not, high-dimensionality data is used. This means that each observation has several measurements. This presents a bit of a challenge where storage space is limited and affects the performance of the learning algorithms.

Dimensionality reduction, in term of unsupervised machine learning, is used quite often in feature pre-processing, to remove the noise from the data. The noise can cause degradation in the predictive performance of some algorithms. It is also commonly used to compress data so it fits into a smaller subspace while keeping the relevant information intact.

Occasionally, we can also use dimensionality reduction for visualizing data. For example, we could project a high-dimensionality feature set onto 1, 2, or 3 D feature spaces – this would allow us to see that feature set through 2D or 3D histograms or scatter plots.

Semi – supervised Learning

Semi-supervised learning is a blend of both supervised and reinforcement learning. The computer is given an incomplete set of data from which to work.

Some of the data include specific examples of previous decisions made with the available data while other data is missing completely. These algorithms work on solving a specific problem or performing very specific functions that will help them achieve their goals.