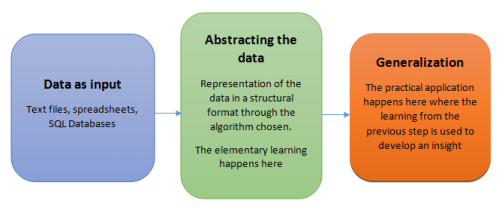
**UNIT-1**

**Machine learning** is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" from its past experience. (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. The past experience is developed through the data collected.

**But, how exactly do we teach machines?**

Teaching the machines involve a structural process where every stage builds a better version of the machine. For simplification purpose, the process of teaching machines can break down into 3 parts:

**What are the steps used in Machine Learning?**

There are 5 basic steps used to perform a machine learning task:

1. **Collecting data**: Be it the raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the future learning. The better the variety, density and volume of relevant data, better the learning prospects for the machine becomes.

2. **Preparing the data**: Any analytical process thrives on the quality of the data used. One needs to spend time determining the quality of data and then taking steps for fixing issues such as missing data and treatment of outliers. Exploratory analysis is perhaps one method to study the nuances of the data in details thereby burgeoning the nutritional content of the data.

3. **Training a model**: This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.

4. **Evaluating the model**: To test the accuracy, the second part of the data (holdout / test data) is used. This step determines the precision in the choice of the algorithm based on the outcome. A better test to check accuracy of model is to see its performance on data which was not used at all during model build.

5. **Improving the performance**: This step might involve choosing a different model altogether or introducing more variables to augment the efficiency. That’s why significant amount of time needs to be spent **in data collection and preparation.**

Be it any model, these 5 steps can be used to structure the technique and when we discuss the algorithms, you shall then find how these five steps appear in every model!

**Application of Machine learning:**

∙ **Healthcare**

Machine learning algorithms can process more information and spot more patterns than their human counterparts. One study used computer assisted diagnosis (CAD)when to review the early mammography scans of women who later developed breast cancer, and the computer spotted 52% of the cancers as much as a year before the women were officially diagnosed. Additionally, machine learning can be used to understand risk factors for disease in large populations. The company Medecision developed an algorithm that was able to identify eight variables to predict avoidable hospitalizations in diabetes patients.

∙ **Fraud Detection**

Machine learning is getting better and better at spotting potential cases of fraud across many different fields. PayPal, for example, is using machine learning to fight money laundering. The company has tools that compare

millions of transactions and can precisely distinguish between legitimate and fraudulent transactions between buyers and sellers.

∙ **Recommendations**

You’re probably familiar with this use if you use services like Amazon or Netflix. Intelligent machine learning algorithms analyze your activity and compare it to the millions of other users to determine what you might like to buy or binge watch next. These recommendations are getting smarter all the time, recognizing, for example, that you might purchase certain things as gifts (and not want the item yourself) or that there might be different family members who have different TV preferences.

∙ **Online Search**

Perhaps the most famous use of machine learning, Google and its competitors are constantly improving what the search engine understands. Every time you execute a search on Google, the program watches how you respond to the results. If you click the top result and stay on that web page, we can assume you got the information you were looking for and the search was a success. If, on the other hand, you click to the second page of results, or type in a new search string without clicking any of the results, we can surmise that the search engine didn’t serve up the results you wanted — and the program can learn from that mistake to deliver a better result in the future.

∙ **Natural Language Processing (NLP)**

NLP is being used in all sorts of exciting applications across disciplines. Machine learning algorithms with natural language can stand in for customer service agents and more quickly route customers to the information they need. It’s being used to translate obscure legalese in contracts into plain language and help attorneys sort through large volumes of information to prepare for a case.

∙ **Smart Cars**

IBM recently surveyed top auto executives, and 74% expected that we would see smart cars on the road by 2025. A smart car would not only integrate into the Internet of Things, but also learn about its owner and its environment. It might adjust the internal settings — temperature, audio, seat position, etc. — automatically based on the driver, report and even fix problems itself, drive itself, and offer real time advice about traffic and road conditions.

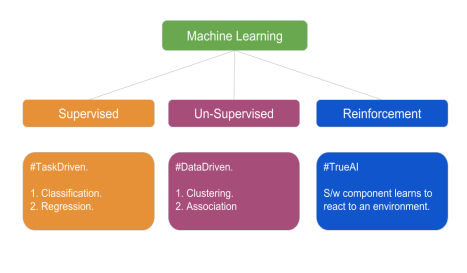
∙ **Banking & Financial services**: ML can be used to predict the customers who are likely to default from paying loans or credit card bills. This is of

paramount importance as machine learning would help the banks to identify the customers who can be granted loans and credit cards.

∙ **Healthcare**: It is used to diagnose deadly diseases (e.g. cancer) based on the symptoms of patients and tallying them with the past data of similar kind of patients.

∙ **Retail**: It is used to identify products which sell more frequently (fast moving) and the slow moving products which help the retailers to decide what kind of products to introduce or remove from the shelf. Also, machine learning algorithms can be used to find which two / three or more products sell together. This is done to design customer loyalty initiatives which in turn help the retailers to develop and maintain loyal customers.

**What are the types of Machine learning algorithms?**

****

**Supervised Learning / Predictive models:**

Predictive model as the name suggests is used to predict the future outcome based on the historical data. Predictive models are normally given clear instructions right from the beginning as in what needs to be learnt and how it needs to be learnt. This

class of learning algorithms is termed as **Supervised Learning.** All data is labeled and the algorithms learn to predict the output from the input data.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y = f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers; the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

For example: Supervised Learning is used when a marketing company is trying to find out which customers are likely to churn. We can also use it to predict the likelihood of occurrence of perils like earthquakes, tornadoes etc. with an aim to determine the Total Insurance Value. Some examples of algorithms used are: Nearest neighbor, Naïve Bayes, Decision Trees, and Regression etc.

Supervised learning problems can be further grouped into regression and classification problems.

∙ **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.

∙ **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”. Examples include real-valued labels denoting the amount of rainfall, the height of a person.

Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Linear Regression, Logistic Regression, CART, Naïve Bayes, SVM KNN are examples of supervised learning Algorithms

Ensemble is a type of supervised learning. It means combining the predictions of multiple different weak ML models to predict on a new sample. Algorithms Bagging with Random Forests, Boosting with XGBoost are examples of ensemble techniques.

Some popular examples of supervised machine learning algorithms are:

∙ Linear regression for regression problems.

∙ Random forest for classification and regression problems.

∙ Support vector machines for classification problems.

**Unsupervised learning / Descriptive models:**

It is used to train descriptive models where no target is set and no single feature is important than the other. All data is unlabeled and the algorithms learn to inherent structure from the input data.

Unsupervised learning is where you only have input data (X) and no corresponding output variables(Y). The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there are no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

For example: The case of unsupervised learning can be: When a retailer wishes to find out what is the combination of products, customers tends to buy more frequently. Furthermore, in pharmaceutical industry, unsupervised learning may be used to predict which diseases are likely to occur along with diabetes. Example of algorithm used here is: K- means Clustering Algorithm

Unsupervised learning problems can be further grouped into clustering and association problems.

∙ **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior. To group samples such that objects within the same cluster are more similar to each other than to the objects from another cluster

∙ **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. : To discover the probability of the co-occurrence of items in a collection. It is extensively used in market-basket analysis. Example: If a customer purchases bread, he is 80% likely to also purchase eggs.

∙ ***Dimensionality Reduction*:** True to its name, Dimensionality Reduction means reducing the number of variables of a dataset while ensuring that important information is still conveyed. Dimensionality Reduction can be done using Feature Extraction methods and Feature Selection methods. Feature Selection selects a subset of the original variables. Feature Extraction performs data transformation from a high-dimensional space to a low-dimensional space. Example

∙

Some popular examples of unsupervised learning algorithms are:

∙ K-means for clustering problems.

∙ Apriori algorithm for association rule learning problems.

∙ : PCA algorithm is a Feature Extraction approach.

**Semi-Supervised Machine Learning**

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems. Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.

These problems sit in between both supervised and unsupervised learning.

A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.

Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

You can use unsupervised learning techniques to discover and learn the structure in the input variables. You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

**Reinforcement learning (RL):**

Reinforcement learning is a type of machine learning algorithm that allows the agent to decide the best next action based on its current state, by learning behaviors that will maximize the reward.

It is an example of machine learning where the machine is trained to take specific decisions based on the business requirement with the sole motto to maximize efficiency (performance). The idea involved in reinforcement learning is: The machine/ software agent trains itself on a continual basis based on the environment it is exposed to, and applies its enriched knowledge to solve business problems. This continual learning process ensures less involvement of human expertise which in turn saves a lot of time!

Reinforcement algorithms usually learn optimal actions through trial and error. They are typically used in robotics – where a robot can learn to avoid collisions by receiving negative feedback after bumping into obstacles, and in video games – where trial and error reveals specific movements that can shoot up a player’s rewards. The agent can then use these rewards to understand the optimal state of game play and choose the next action.

An example of algorithm used in RL is Markov Decision Process.

**Important Note:** There is a subtle difference between Supervised Learning and Reinforcement Learning (RL). RL essentially involves learning by interacting with an environment. An RL agent learns from its past experience, rather from its continual trial and error learning process as against supervised learning where an external supervise or provides examples.

A good example to understand the difference is self driving cars. Self driving cars use Reinforcement learning to make decisions continuously – which route to take? What speed to drive on? Are some of the questions which are decided after interacting with the environment? A simple manifestation for supervised learning would be to predict fare from a cab going from one place to another.