MACHINE LEARNING Notes - 201CS6T01

Unit – I

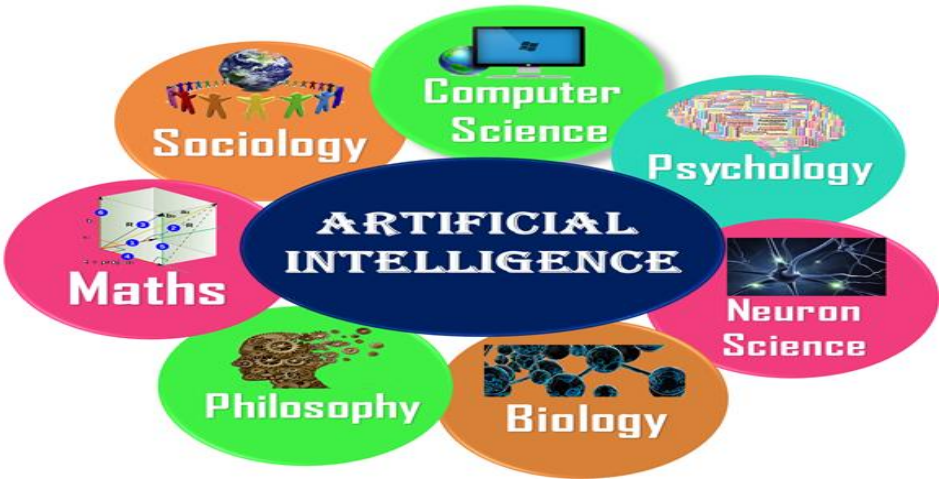
Introduction- Artificial Intelligence, Machine Learning, Deep learning, Types of Machine Learning Systems, Main Challenges of Machine Learning. Statistical Learning: Introduction, Supervised and Unsupervised Learning, Training and Test Loss, Trade-offs in Statistical Learning, Estimating Risk Statistics, Sampling distribution of an estimator, Empirical Risk Minimization.

TOPIC-1: Introduction- Artificial Intelligence, Machine Learning, Deep learning:

- **Artificial Intelligence (AI):** In today's world, technology is growing very fast, and we are getting in touch with different new technologies day by day.
- Here, one of the booming technologies of computer science is Artificial Intelligence which is ready to create a new revolution in the world by making intelligent machines.
- Artificial Intelligence is composed of two words Artificial and Intelligence, where Artificial defines "man-made," and intelligence defines "thinking power", hence AI means "a man-made thinking power."
- So, we can define AI as: "It is a branch of computer science by which we can create intelligent machines which can behave like a human, think like humans, and able to make decisions."
- Artificial Intelligence exists when a machine can have human based skills such as learning, reasoning, and solving problems.

Why Artificial Intelligence?

- With the help of AI, you can create such software or devices which can solve real-world problems very easily and with accuracy such as health issues, marketing, traffic issues, etc.
- With the help of AI, you can create your personal virtual Assistant, such as Cortana, Google Assistant, Siri, etc.
- With the help of AI, you can build such Robots which can work in an environment where survival of humans can be at risk.
- AI opens a path for other new technologies, new devices, and new Opportunities.



Machine Learning:

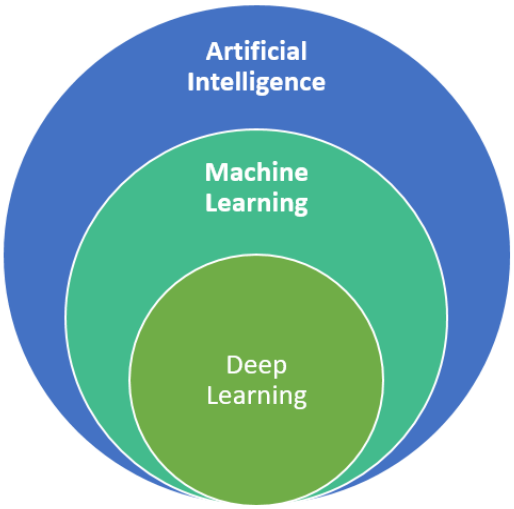
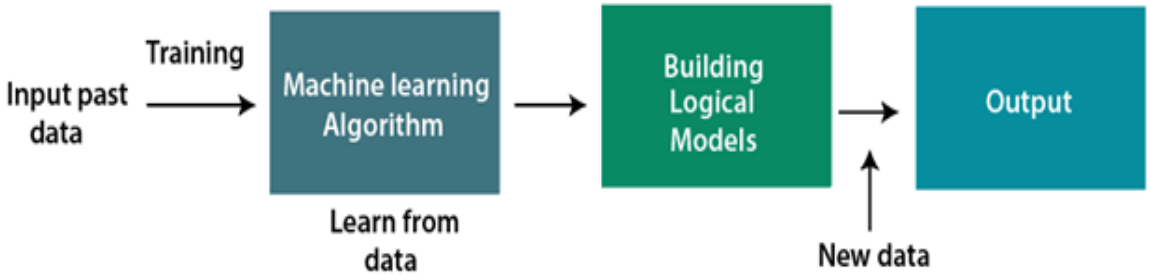


Figure 1: artificial intelligence, machine leaning and deep learning Source: Nadia BERCHANE (M2 IESCI, 2018)

- Machine learning is a growing technology which enables computers to learn automatically from past data.
- Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information.
- Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

Arthur Samuel

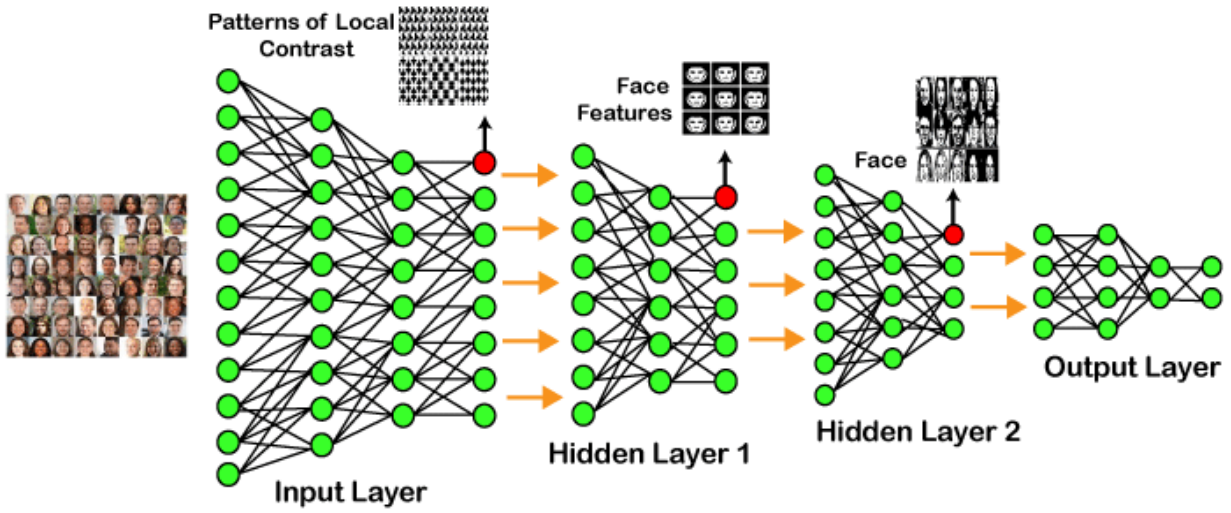
- The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as:
- **Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.**



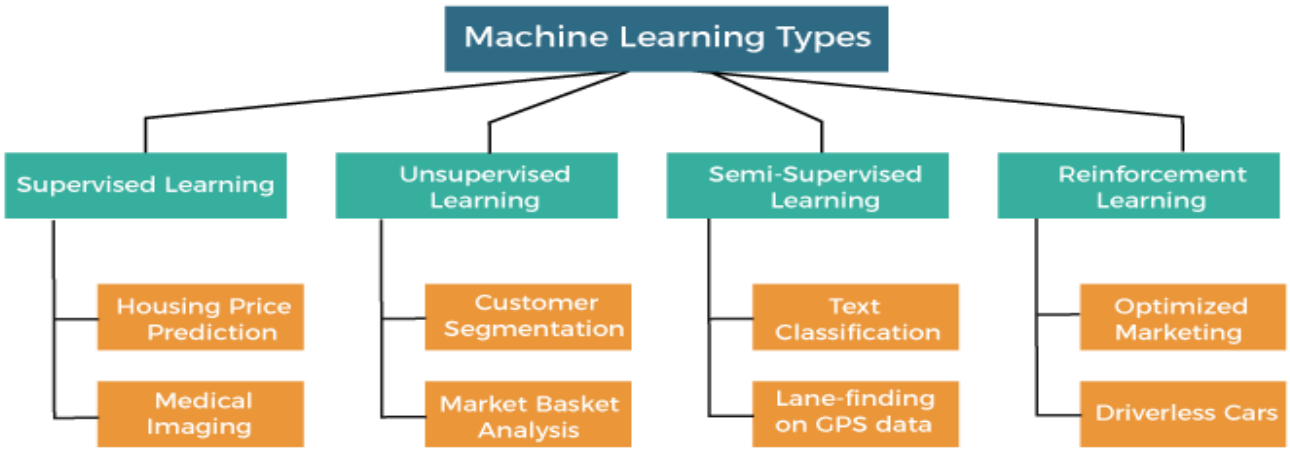
Deep Learning:

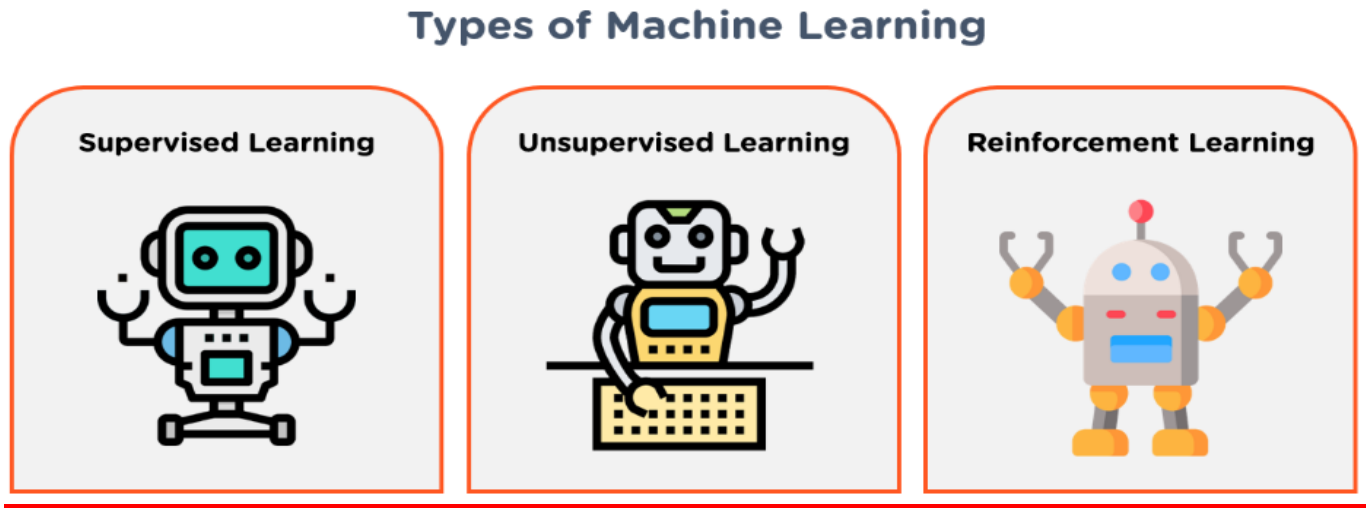
- Deep learning is based on the branch of machine learning, which is a subset of artificial intelligence.

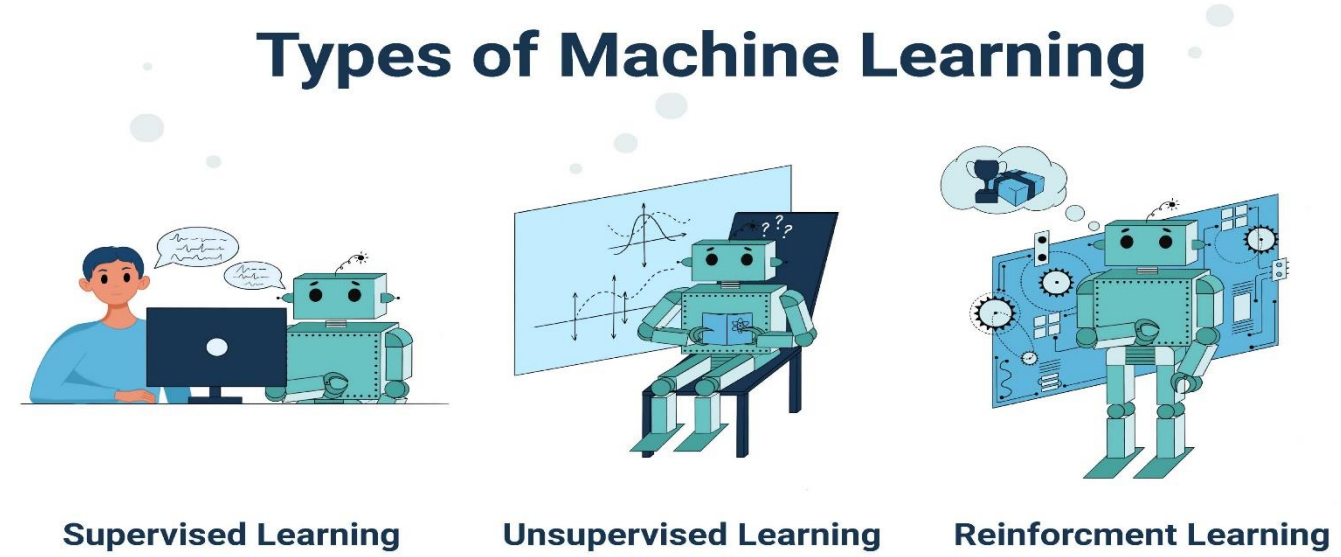
- Since neural networks imitate the human brain and so deep learning will do. In deep learning, nothing is programmed explicitly.
- Basically, it is a machine learning class that makes use of numerous nonlinear processing units so as to perform feature extraction as well as transformation.
- IDEA: Deep learning is implemented with the help of Neural Networks, and the idea behind the motivation of Neural Network is the biological neurons, which is nothing but a brain cell.
- Deep learning is a collection of statistical techniques of machine learning for learning feature hierarchies that are actually based on artificial neural networks.
- **Example of Deep Learning:**



TOPIC-2: Types of Machine Learning Systems







There are so many different types of Machine Learning systems that it is useful to classify them in broad categories, based on the following criteria:

1. Whether or not they are trained with human supervision (supervised, unsupervised, semi supervised, and Reinforcement Learning)

2. Whether or not they can learn incrementally on the fly (online versus batch learning)

3.Whether they work by simply comparing new data points to known data points, or instead by detecting patterns in the training data and building a predictive model, much like scientists do (instance-based versus model-based learning).

1. Supervised Machine Learning: As its name suggests, supervised machine learning is based on supervision.

- It means in the supervised learning technique, 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). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

Categories of Supervised Machine Learning:

- Supervised machine learning can be classified into two types of problems, which are given below:
- **Classification**
- **Regression**

Classification: Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc.

- The classification algorithms predict the categories present in the dataset.

- Some real-world examples of classification algorithms are Spam Detection, Email filtering, etc.

Some popular classification algorithms are given below:

- Random Forest Algorithm
- Decision Tree Algorithm
- Logistic Regression Algorithm
- Support Vector Machine Algorithm

Regression:

- Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables.
- These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

- Simple Linear Regression Algorithm
- Multivariate Regression Algorithm
- Decision Tree Algorithm
- Lasso Regression

Advantages and Disadvantages of Supervised Learning:

Advantages:

- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- 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.

2. Unsupervised Machine Learning:

- Unsupervised learning is different from the supervised learning technique; as its name suggests, there is no need for supervision.
- It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output w
- **The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.**
- Machines are instructed to find the hidden patterns from the input dataset.

Categories of Unsupervised Machine Learning:

Unsupervised Learning can be further classified into two types, which are given below:

- **Clustering**
- **Association**

1) Clustering:

- The clustering technique is used when we want to find the inherent groups from the data.
- It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups.
- An example of the clustering algorithm is grouping the customers by their purchasing behavior.

Some of the popular clustering algorithms are given below:

- K-Means Clustering algorithm
- Mean-shift algorithm
- DBSCAN Algorithm
- Principal Component Analysis
- Independent Component Analysis

2) Association:

- Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset.
- The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit.
- Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

Advantages and Disadvantages of Unsupervised Learning Algorithm:

Advantages:

- These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
- Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled 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 unlabeled dataset that does not map with the output.

3. Semi-Supervised Learning:

- **Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning**.
- It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced.

- We can imagine these algorithms with an example. Supervised learning is where a student is under the supervision of an instructor at home and college.
- Further, if that student is self- analyzing the same concept without any help from the instructor, it comes under unsupervised learning.
- Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of an instructor at college.

Advantages:

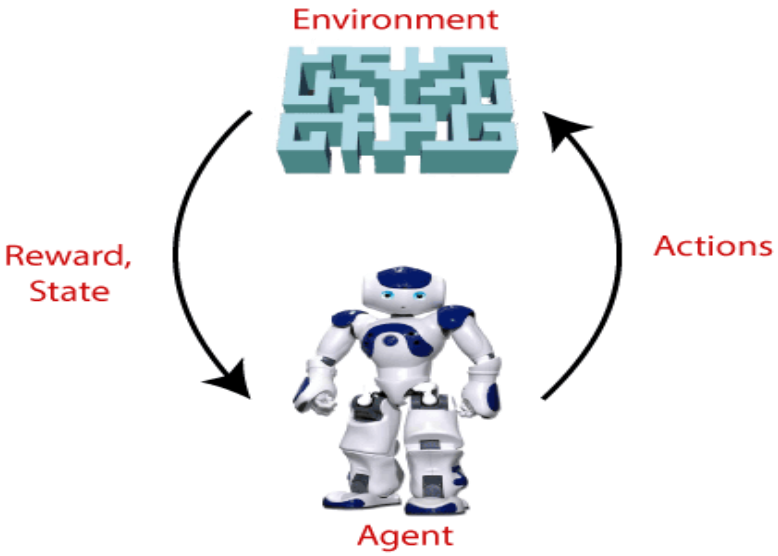
- It is simple and easy to understand the algorithm.
- It is highly efficient.
- It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages:

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

4. Reinforcement Learning:

- **Reinforcement learning 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.**
- Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.
- In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.



- The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life.
- An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score.
- Agent receives feedback in terms of punishment and rewards.
- Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.**

Categories of Reinforcement Learning:

- Reinforcement learning is categorized mainly into two types of methods/algorithms:
- **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behavior would occur again by adding something. It enhances the strength of the behavior of the agent and positively impacts it.
- **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behavior would occur again by avoiding the negative condition.

Real-world Use cases of Reinforcement Learning

- **Video Games**
- **Robotics**
- **Text Mining**

TOPIC-3: Main Challenges of Machine Learning:

1) Lack Of Quality Data

One of the main issues in Machine Learning is the absence of good data. While upgrading, algorithms tend to make developers exhaust most of their time on artificial intelligence.

- Data can be noisy which will result in inaccurate predictions.
- Incorrect or incomplete information can also lead to faulty programming through Machine Learning.

2) Fault In Credit Card Fraud Detection

Although this AI-driven software helps to successfully detect credit card fraud, there are issues in Machine Learning that make the process redundant.

3) Getting Bad Recommendations

Proposal engines are quite regular today. While some might be dependable, others may not appear to provide the necessary results. Machine Learning algorithms tend to only impose what these proposal engines have suggested.

4) Talent Deficit

Albeit numerous individuals are pulled into the ML business, however, there are still not many experts who can take complete control of this innovation.

5) Implementation

Organizations regularly have examination engines working with them when they decide to move up to ML. The usage of fresher ML strategies with existing procedures is a complicated errand.

6) Making The Wrong Assumptions

ML models can’t manage datasets containing missing data points. Thus, highlights that contain a huge part of missing data should be erased.

7) Deficient Infrastructure

ML requires a tremendous amount of data stirring abilities. Inheritance frameworks can’t deal with the responsibility and clasp under tension.

8) Having Algorithms Become Obsolete When Data Grows

ML algorithms will consistently require a lot of data when being trained. Frequently, these ML algorithms will be trained over a specific data index and afterwards used to foresee future data, a cycle which you can only expect with a significant amount of effort.

9) Absence Of Skilled Resources

The other issues in Machine Learning are that deep analytics and ML in their present structures are still new technologies.

10) Customer Segmentation

Let us consider the data of human behaviour by a user during a time for testing and the relevant previous practices. All things considered, an algorithm is necessary to recognize those customers that will change over to the paid form of a product and those that won't.

The lists of supervised learning algorithms in ML are:

- Neural Networks
- Naive Bayesian Model
- Classification
- Support Vector Machines
- Regression
- Random Forest Model

11) Complexity

Although Machine Learning and Artificial Intelligence are booming, a majority of these sectors are still in their experimental phases, actively undergoing a trial and error method.

12) Slow Results

Another one of the most common issues in Machine Learning is the slow-moving program. The Machine Learning Models are highly efficient bearing accurate results but the said results take time to be produced.

13) Maintenance

Requisite results for different actions are bound to change and hence the data needed for the same is different.

14) Concept Drift

This occurs when the target variable changes, resulting in the delivered results being inaccurate. This forces the decay of the models as changes cannot be easily accustomed to or upgraded.

15) Data Bias

This occurs when certain aspects of a data set need more importance than others.

16) High Chances Of Error

Many algorithms will contain biased programming which will lead to biased datasets. It will not deliver the right output and produces irrelevant information.

17) Lack Of Explainability

Machine Learning is often termed a “Black box” as deciphering the outcomes from an algorithm is often complex and sometimes useless.

TOPIC-4 Statistical Learning: Introduction

- Structuring and visualizing data are important aspects of data science, the main challenge lies in the mathematical analysis of the data.
- When the goal is to interpret the model and quantify the uncertainty in the data, this analysis is usually referred to as statistical learning.

There are two major goals for modeling data:

- 1) to accurately predict some future quantity of interest, given some observed data, and
- 2) To discover unusual or interesting patterns in the data.

TOPIC-5 Supervised and Unsupervised Learning:

1. Feature, Response:

- Given an input or feature vector x, one of the main goals of machine learning is to predict response an output or response variable y.
- For example, x could be a digitized signature and y a binary variable that indicates whether the signature is genuine or false.

2. Prediction function:

- Another example is where x represents the weight and smoking habits of an expecting mother and y the birth weight of the baby.
- The data science attempt at this prediction is encoded in a mathematical prediction function g, called the prediction function function, which takes as an input x and outputs a guess g(x) for y.

3. Regression, classification:

- In regression problems, the response variable y can take any real value.
- In contrast, regression when y can only lie in a finite set, say $y \in \{0. . . c - 1\}$, then predicting y is conceptually the same as classifying the input x into one of c categories, and so prediction becomes a classification problem.
- loss function:
- We can measure the accuracy of a prediction by with respect to a given response y by loss function using some Loss(y,y').
- In a regression setting the usual choice is the squared error loss $(y-y')\ 2$.

TOPIC-6 Training and Test Loss:

Given an arbitrary prediction function g, it is typically not possible to compute its risk $\ell(g)$ in (2.1). However, using the training sample $\mathcal{T}$, we can approximate $\ell(g)$ via the empirical (sample average) risk

$$\ell_{\mathcal{T}}(g) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}(Y_i, g(\boldsymbol{X}_i)), \tag{2.3}$$

which we call the *training loss*. The training loss is thus an unbiased estimator of the risk (the expected loss) for a prediction function g, based on the training data.

To approximate the optimal prediction function g^* (the minimizer of the risk $\ell(g)$) we first select a suitable collection of approximating functions $\mathcal{G}$ and then take our *learner* to be the function in $\mathcal{G}$ that minimizes the training loss; that is,

$$g_{\mathcal{T}}^{\mathcal{G}} = \underset{g\in\mathcal{G}}{\text{argmin}}\, \ell_{\mathcal{T}}(g). \tag{2.4}$$

The prediction accuracy of new pairs of data is measured by the *generalization risk* of the learner. For a *fixed* training set τ it is defined as

$$\ell(g_{\tau}^{\mathcal{G}}) = \mathbb{E}\,\text{Loss}(Y, g_{\tau}^{\mathcal{G}}(\boldsymbol{X})), \tag{2.5}$$

For any outcome τ of the training data, we can estimate the generalization risk without bias by taking the sample average

$$\ell_{\mathcal{T}'}(g_\tau^{\mathcal{G}}) := \frac{1}{n'} \sum_{i=1}^{n'} \text{Loss}(Y_i', g_\tau^{\mathcal{G}}(\boldsymbol{X}_i')), \tag{2.7}$$

where $\{(\boldsymbol{X}_1', Y_1'), \ldots, (\boldsymbol{X}_{n'}', Y_{n'}')\} =: \mathcal{T}'$ is a so-called *test sample*. The test sample is completely separate from $\mathcal{T}$, but is drawn in the same way as $\mathcal{T}$; that is, via independent draws from $f(\boldsymbol{x}, y)$, for some sample size n'. We call the estimator (2.7) the *test loss*. For a random training set $\mathcal{T}$ we can define $\ell_{\mathcal{T}'}(g_{\mathcal{T}}^{\mathcal{G}})$ similarly. It is then crucial to assume that $\mathcal{T}$ is independent of $\mathcal{T}'$. Table 2.1 summarizes the main definitions and notation for supervised learning.

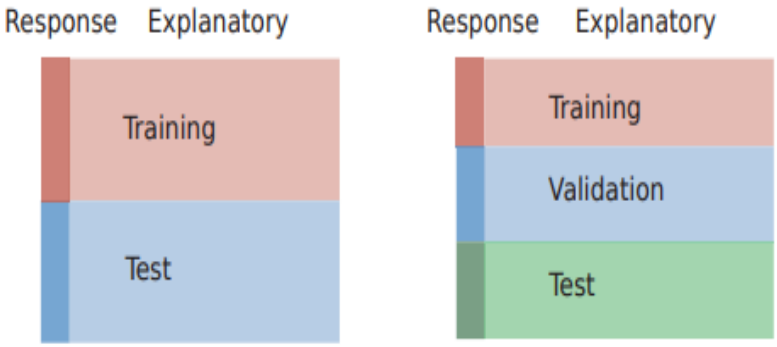


Figure 2.3: Statistical learning algorithms often require the data to be divided into training and test data. If the latter is used for model selection, a third set is needed for testing the performance of the selected model.

TOPIC-7 Tradeoffs in Statistical Learning:

The art of machine learning in the supervised case is to make the generalization risk (2.5) or expected generalization risk (2.6) as small as possible, while using as few computational resources as possible. In pursuing this goal, a suitable class $\mathcal{G}$ of prediction functions has to be chosen. This choice is driven by various factors, such as

- the complexity of the class (e.g., is it rich enough to adequately approximate, or even contain, the optimal prediction function g^*?),
- the ease of training the learner via the optimization program (2.4),
- how accurately the training loss (2.3) estimates the risk (2.1) within class $\mathcal{G}$,
- the feature types (categorical, continuous, etc.).

We can decompose the generalization risk (2.5) into the following three components:

$$\ell(g_\tau^{\mathcal{G}}) = \underbrace{\ell^*}_{\text{irreducible risk}} + \underbrace{\ell(g^{\mathcal{G}}) - \ell^*}_{\text{approximation error}} + \underbrace{\ell(g_\tau^{\mathcal{G}}) - \ell(g^{\mathcal{G}})}_{\text{statistical error}}, \tag{2.16}$$

where $\ell^* := \ell(g^*)$ is the *irreducible risk* and $g^{\mathcal{G}} := \operatorname{argmin}_{g \in \mathcal{G}} \ell(g)$ is the best learner within class $\mathcal{G}$. No learner can predict a new response with a smaller risk than ℓ^*.

The second component is the *approximation error*; it measures the difference between the irreducible risk and the best possible risk that can be obtained by selecting the best prediction function in the selected class of functions $\mathcal{G}$.

The third component is the *statistical (estimation) error*. It depends on the training set τ and, in particular, on how well the learner $g_\tau^{\mathcal{G}}$ estimates the best possible prediction function, $g^{\mathcal{G}}$, within class $\mathcal{G}$. For any sensible estimator this error should decay to zero (in

TOPIC-8 Estimating Risk:

The most straightforward way to quantify the generalization risk (2.5) is to estimate it via the test loss (2.7). However, the generalization risk depends inherently on the training set, and so different training sets may yield significantly different estimates.

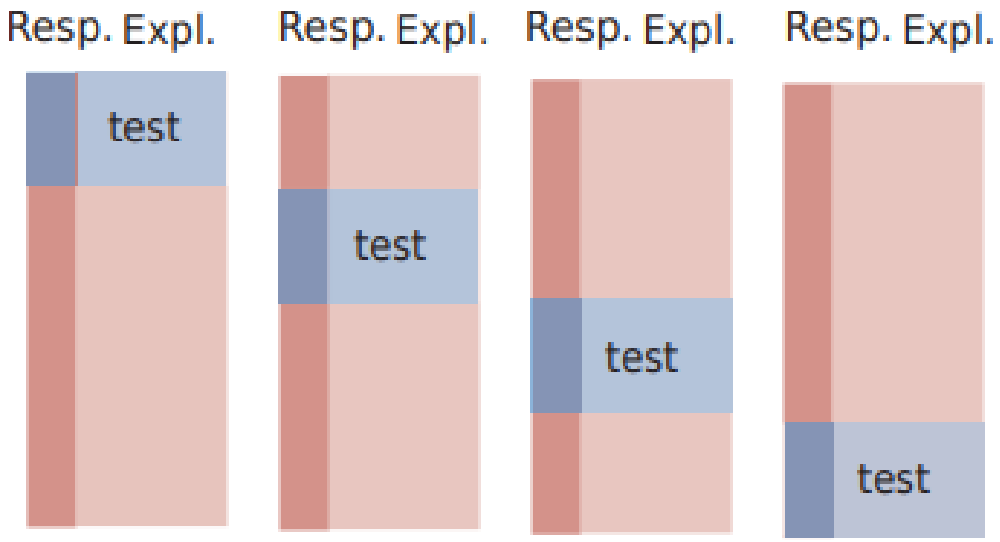
1. IN-SAMPLE RISK:

We mentioned that, due to the phenomenon of overfitting, the training loss of the learner, $\ell_\tau(g_\tau)$ (for simplicity, here we omit $\mathcal{G}$ from $g_\tau^{\mathcal{G}}$), is not a good estimate of the generalization risk $\ell(g_\tau)$ of the learner. One reason for this is that we use the same data for both training the model and assessing its risk. How should we then estimate the generalization risk or expected generalization risk?

To simplify the analysis, suppose that we wish to estimate the average accuracy of the predictions of the learner g_τ at the n feature vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (these are part of the training set τ). In other words, we wish to estimate the *in-sample risk* of the learner g_τ:

$$\ell_{\text{in}}(g_\tau) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\,\text{Loss}(Y_i', g_\tau(\boldsymbol{x}_i)), \tag{2.23}$$

2. CROSS-VALIDATION



TOPIC-9 Sampling distributions of estimators

Since our estimators are statistics (particular functions of random variables), their distribution can be derived from the joint distribution of X1 . . . Xn.

It is called the sampling distribution because it is based on the joint distribution of the random sample.

-Given a sampling distribution, we can – calculate the probability that an estimator will not differ from the parameter θ by more than a specified amount

– obtain interval estimates rather than point estimates after we have a sample

- An interval estimate is a random interval such that the true parameter lies within this interval with a given probability (say 95%).

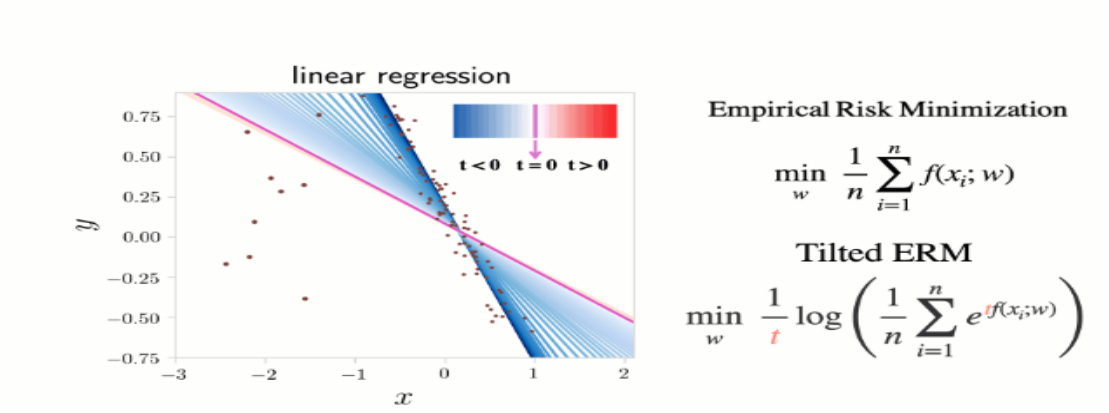
– Choose between to estimators- we can, for instance, calculate the mean-squared error of the estimator, $E\theta[(\theta\hat{} - \theta)\ 2\ ]$ using the distribution of θˆ.

Sampling distributions of estimators depend on sample size, and we want to know exactly how the distribution changes as we change this size so that we can make the right trade-offs between cost and accuracy.

TOPIC-10 Empirical Risk Minimization:

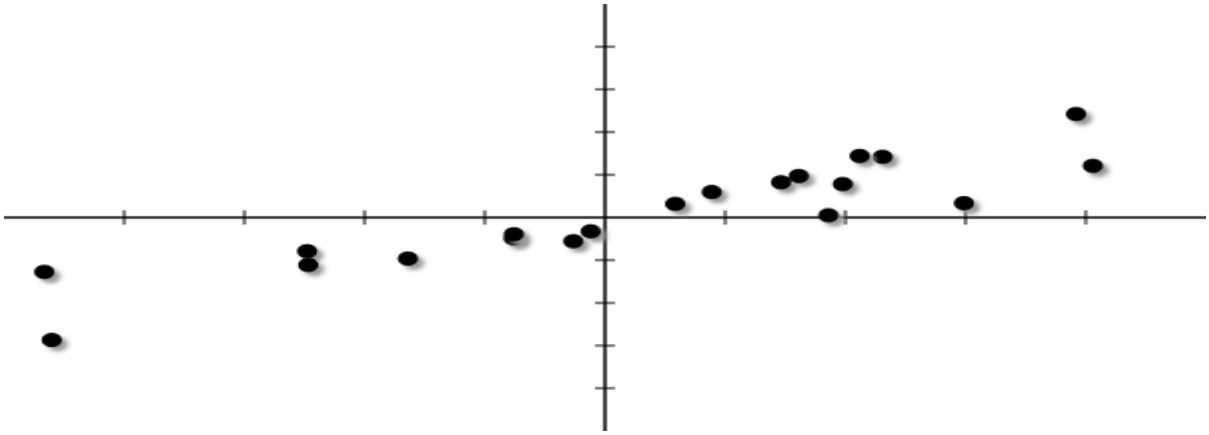
- Empirical Risk Minimization is a fundamental concept in machine learning, yet surprisingly many practitioners are not familiar with it.
- Understanding ERM is essential to understanding the limits of machine learning algorithms and to form a good basis for practical problem-solving skills.
- The theory behind ERM is the theory that explains the VC-dimension, Probably Approximately Correct (PAC) Learning and other fundamental concepts.

Tilted Empirical Risk Minimization



The ERM is a nice idea, if used with care

The plot below shows a regression problem with a training set of 15 points.



The ERM principle is an inference principle which consists in finding the model f^ by minimizing the empirical risk:

$$f\hat{} = \arg\min_{f:X\to Y} R_{emp}(h)$$

where the empirical risk is an estimate of the risk computed as the average of the loss function over the training sample $D=\{(X_i,Y_i)\}_{i=1}^{N}$:

$$R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}\ell(f(X_i),Y_i)$$

with the loss function ℓ.