Fundamentals of Machine Learning

A. Vinoth and Shubhabrata Datta

Abstract This introductory chapter describes the techniques of machine learning. Primarily the concept of machine learning in the context of artificial intelligence and data analytics is explained. The application process of the above to big data is introduced. Classification of machine learning approaches is described. Then some of the variedly used statistical and artificial intelligence-based machine learning techniques are described in brief. The techniques discussed include decision tree, linear regression, least square method, artificial neural network, clustering techniques. The concepts of deep learning are also introduced.

Keywords Machine learning · Artificial intelligence · Data analytics · Supervised learning · And unsupervised learning

1 Introduction

Machine learning (ML) is a subdivision of computational science which is progressed from the learning of data classification based on the gained understanding and also from the learning gained on computational-based principles of Artificial Intelligence (AI). In simple, machine learning is training the computers to learn automatically through the inputs deprived of being explicitly programmed [1]. The term learning evolved from the humans and animals. Animal and machine learning have quite a few matches. Indeed, a lot of methods in machine learning originate to mark principles of animal and human learning by computational models. For instance, habituation is a basic scholarly conduct where an animal step by step quits reacting to a rehashed stimulus. Dogs are considered to be a perfect example for animal learning where it is capable of substantial learning if it is trained to perform various activities like rolling over, sitting and picking up the things, etc.

A. Vinoth · S. Datta (✉)
Department of Mechanical Engineering, SRM Institute of Science and Technology,
Kattankulathur, Chennai 603203, Tamil Nadu, India
e-mail: shubhabp@srmist.edu.in

S. Datta and J. P. Davim (eds.), *Machine Learning in Industry*,
Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-75847-9_1

With regard to the former example of effective learning, there are few examples that could demonstrate machine learning where we use in our day-to-day life of modern era. Virtual personal assistants, traffic predictions using GPS navigation, surveillance of multiple cameras by AI to detect the crime or unusual behaviour of people, social media uses ML for face recognition and news feed personalization, search engine result refinement, e-mail spam filtering where a machine memorize all the earlier labeled spam e-mails by the user, and lot more applications where ML is widely in use. Through all these applications, it is understood that the incorporation of prior knowledge will preference the mechanism of learning. ML is also closely interconnected to computational statistics where it familiarizes the prediction making [2]. Anyone might wonder 'why a machine must learn something?' There are few aims why ML is essential. Obviously we have just referenced that the accomplishment of learning in machines may assist us with seeing how creatures and people learn. Yet there are few essential engineering details that persist and some of these are

- Certain tasks cannot be clearly explained without example; i.e. we may have the option to identify input/output sets however not a brief correlation between inputs and preferred outputs.
- It is probable that there are unseen relationships between inputs and outputs among huge loads of data. Machine learning methods can repeatedly be utilized to reveal these relationships.

When do we want machine learning instead of straight away program our computers to perform a task? Two characteristics of a certain problem may demand for the usage of programs that learn and develop on the basis of their experience/understanding, i.e. the complexity of problem and the want for adaptivity. There are tasks that are complex to program, for instance human activities like driving, understanding of images and voice recognition of a person, etc., where the art of ML works on the principle of learning through experience that could yield reasonable results [3]. One restraining feature of automated tools is their inflexibility, i.e. once the coding has been formulated and installed, it remains unchanged. Still, many tasks change over period or from one end user to another. For such problems, the utilization of ML which has coding that decode the earlier written program adapting a fixed program to check the variations among the styles of different users.

2 Artificial Intelligence

Artificial intelligence (AI) denotes the replication of human intellect in machines that are encoded to imitate human activities. The term may likewise be applied to any machine that exhibits human qualities, for instance, learning and critical thinking [4]. A more elaborate definition describes AI as 'a system's capacity to effectively decipher outside information, to gain from such information, and to utilize those learning to accomplish explicit objectives and assignments through adaptable transformation'. As innovation progresses, earlier standards that branded AI become outdated. For example, machines that establish necessary capabilities or identify text through model character identification are not, at this juncture said to represent AI, because this purpose is presently underrated as a built-in function of a computer. Artificial intelligence is ceaselessly evolving for the benefit of various enterprises. Machines are wired using a cross-disciplinary approach that includes arithmetic, software engineering, semantics, brain science and a lot more with specialized fields like artificial study of the mind. The objectives of AI incorporate learning, thinking, communication and recognition.

AI is exceptionally focused, and is intensely divided into subfields that are very different from one another [5]. A part of the classification is because of social and cultural elements: subfields have developed over specific foundations and contributions of various researchers. AI is additionally isolated by limited specific topics. Some subfields emphasize on the solution of explicit issues. Others center around one of a few potential procedures or on the utilization of a specific tool or toward the accomplishment of particular applications. AI has been the subject of good faith however has withstood alluring difficulties. Now, it has become a basic aspect of the innovation business, giving the truly difficult work to a significant amount of the major testing disputes in software work.

In the early nineteenth century, AI research is evolved in different ways like digital computer's formal thinking that could mimic any possible demonstration of numerical derivation in 1943, writing simple programs/algorithms to solve problems in algebra, theorems and speaking English in 1956 [6]. US government has started investing on AI research in 1960 on developing various laboratories around the world. Due to the lot of failure that has occurred in the research till 1974, finance for AI projects was difficult to obtain. During 1980s, with the help of few professionals AI research was rejuvenated by the profitable achievement of expert systems. In 1990s and the early twenty-first century, AI attained its great feats where it is utilized for logistics, data mining, clinical findings and numerous different regions all through the innovation business. The quest for more proficient critical thinking algorithms for solving problems in sequential steps is a high need for AI research. AI has made some progress on mimicking these sorts of processes which highlight on the need of good reasoning skills, neural net exploration endeavors to recreate the structures inside the cerebrum that offer ascent to this ability; measurable ways to deal with AI copy the probabilistic nature of the human capacity to predict. AI frequently spins around the utilization of algorithms where plenty of clear-cut directions that a computer can perform. An unpredictable algorithm is regularly based on other easier algorithms which basically cover deduction, reasoning and problem-solving.

Key researches of AI are knowledge depiction and knowledge engineering [7]. Huge numbers of the subjects machines are trusted upon to illuminate will need extensive information about the world. Effort on the development of AI research works is based on the commonsensical knowledge involve huge extents of lengthy ontological engineering as they should be worked, by hand, each convoluted idea in turn. The common traits of an AI system involve the following viz planning, learning, communication, perception, motion and manipulation. Pertaining to planning, an intelligent agent can imagine the future to make predictions in a better way that would change the world and will be able to utilize the available choices to the maximum. Periodic checks on the predictions with the actuals to be done and if required an agent can make change on the plan to avoid any uncertainty. Learning involves the machine learning under three different regimes as supervised learning, unsupervised learning and reinforcement learning and the same will be discussed in detail in the latter part. Communicating the machines will be done through natural language processing where a machine can peruse and comprehend the languages that people talk. A typical technique for processing and pull out significance from normal language is done over semantic ordering which increases the processing speed and cut short the cost of large data storage. Perception of machine is the capacity to use response from various sensors to presume features of the world. Motion and Manipulation in AI are closely related to the field of robotics to handle different jobs as object management and triangulation through robots.

Long-term goals relating to AI research are (a) Societal Intelligence (b) Creativeness and (c) Common Intelligence [8]. Affective computing is the form of societal intelligence that focuses on the investigation and improvement of systems and gadgets that can perceive, decipher, measure and recreate human effects. A branch AI tends to imagination both theoretically (from a philosophical and mental viewpoint) and essentially (by way of explicit usage of charters that yield results that can be regarded as inventive, or frameworks that recognize and survey creativity). Associated zones of computational assessment are Artificial instinct and Artificial reasoning. Numerous analysts believe that their effort will eventually be merged into a machine with general intelligence (known as solid AI), all the aptitudes above and exceeding human abilities all things considered or each one of them. A couple accepts that human highlights comparable to fake alertness or an artificial brain might be necessary for such an undertaking.

Different approaches of AI are classified broadly as '1. Cybernetics and mind simulation' which connects the nerve system, theory of information and automations. '2. Symbolic AI' that would ultimately prosper in building a machine with artificial general intelligence that evolve from 1960s to 1990s as cognitive simulation (based on cognitive and management science), logic-based approach (based on the principle of abstract reasoning and problem solving), knowledge-based approach (based on knowledge revolution into AI applications), sub-symbolic approaches (to definite AI problems), computational intelligence and soft computing (subset of AI that focuses on neural networks, fuzzy systems, evolutionary computation, etc.) and statistical approaches (based on refined mathematical tools to resolve precise sub problems). Integration of the above said approaches are also possible by utilizing the 'Intelligent agent paradigm' and 'Agent and cognitive architectures' [9]. Former one focuses on considering specific glitches and discover resolutions that are valuable, without concurring on one single methodology. Latter focuses on connecting the multiple AI systems as a hybrid system, which has both symbolic and sub-symbolic components.

Different tools of AI involves search algorithm (informed and uninformed search algorithms), mathematical optimization (simulated annealing, random optimization, blind hill climbing and beam search), evolutionary algorithms (ant colony and particle

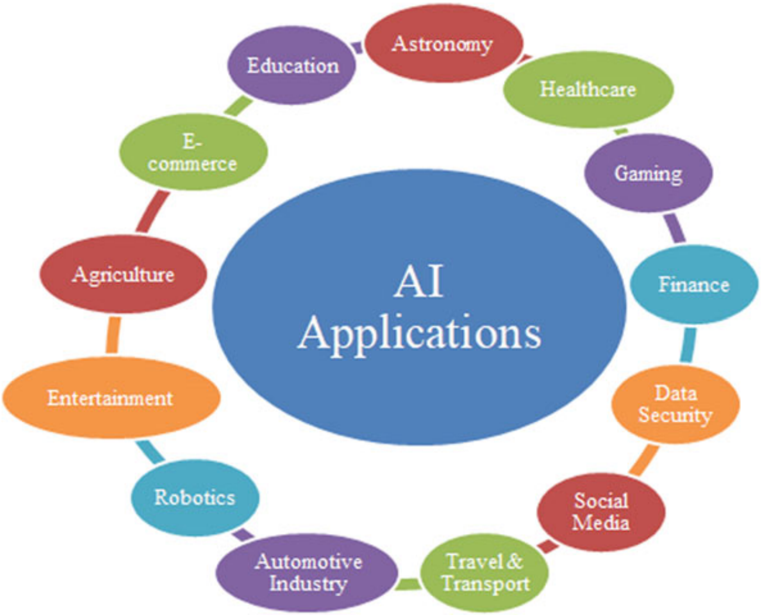


Fig. 1 Applications of Artificial Intelligence

swarm optimization, genetic algorithms and genetic programming), logic programming and automated reasoning (default logics, non-monotonic logics and circumscription), probabilistic methods for uncertain reasoning (Bayesian inference algorithm, decision networks, probabilistic algorithms, etc.), classifiers and statistical learning methods (Neural networks, Gaussian mixture model, decision tree, etc.). Figure 1 describes the applications of AI in various fields.

3 Data Analytics

One of the mathematical and statistical approaches of analyzing data is data analytics that mainly focuses on what the data can say us outside the proper modeling or testing of hypothesis. Data Analysis practices business intelligence and analytics models. Business intelligence (BI) is a prearrangement of techniques and tools for the change of crude data into significant and useful data for business research drives. BI advancements are fortified for dealing with formless data to distinguish, generate and in any case create fresh key business openings. The aim of BI is to consider the guileless understanding of huge bulks of data. One of the analytic models of data analysis is exploratory data analysis (EDA) for data investigation, and to arrive at a hypotheses that could prompt new data assortment and investigations. EDA is remarkable in relation to initial data analysis (IDA), which hubs around scrutinizing

Table 1 Techniques in EDA

Graphical techniques in EDA	Quantitative techniques in EDA
• Pareto chart • Histogram • Run chart • Stem-and-leaf plot • Box plot • Targeted projection pursuit • Multi-vari chart • Parallel coordinates • Scatter plot • Multilinear PCA • Multidimensional scaling • Odds ratio • Principal component analysi	• Ordination • Trimean • Median polish

suppositions essential for model fitting and theory testing, considering missing qualities and changing factors varying. In 1961, Tukey characterized data analysis as procedures for evaluating data, policies for cracking the outputs of such methods, means of positioning the data to simplify analysis exactly, and all the hardware and results of (numerical) statistical data put on to evaluate the data [6]. These statistical developments, all braced by Tukey, were envisioned to add-on to the scientific theory of testing measurable assumptions, chiefly the Laplacian convention's prominence on exponential families [10]. The objectives of EDA are to propose hypothesis, evaluate the expectations statistically, choosing suitable statistical tools/techniques and to pave a way for further data gathering via studies or experimentations. Some of the graphical and quantitative techniques in EDA are listed as given in Table 1.

3.1 Types of Data Analytics

Data analytics is a wide arena of study. The four essential classes of data analytics as descriptive, diagnostic, predictive and prescriptive analytics. Each of it has an alternate objective and a better position in the course of the evaluation of data. These are also the key data analytics in commercial applications. *Descriptive analytics* supports answering studies concerning the whereabouts. These methods total up enormous datasets to portray outcomes to associates. By creating key performance indicators (KPIs), these methodologies can support path achievements or dissatisfactions. Measurements, for example, return on investment (ROI) are utilized in numerous ventures. Exacting dimensions are fashioned to trail accomplishment in unambiguous ventures. This cycle needs the collection of substantial data, organizing of the data, investigation of data and conception of data. This cycle gives basic knowledge into past accomplishments.

Diagnostic analytics supports answering inquiries regarding why things occurred. These methods complement more fundamental descriptive analytics. They contemplate the discoveries from descriptive analytics and burrow further to discover the reason. The performance pointers are additionally explored to find the reason for improvement. This happens mostly in three stages:

- Recognize irregularities in the data. The sudden variations in a quantity or a definite market.
- Gathering of data connected to such irregularities.
- Statistical methods are employed to ascertain networks and designs that simplify such irregularities.

Predictive analytics supports answering inquiries concerning the later events. Such methods employ verifiable data to distinguish drifts and resolve if they are possibly going to recur. Predictive analytical tools provide vital understanding into future events and it measures integrate a collection of numerical and AI procedures, for example, neural networks, decision trees and regression. The prediction of data is crucial in data analytics hence a detailed study on this is highly needed. The types of predictive analytics are 1. Predictive modelling 2. Descriptive modeling and 3. Decision modelling. Predictive models will be replicas of the link between the explicit output of a unit in a model and known credits or highlights of the unit. The goal of the model is to appraise the prospect that a comparative unit in a substitute model will display the specific output. This model is used in wide area, such as marketing, carrying out calculations in live businesses to guide a decision, crime scene investigation [11] and due to its computing quickness it can simulate behaviour or reactions of people for particular situations. Descriptive models measure influences in data to arrange clients or forecasts into collections. Not all predictive models emphasize on forestalling a self-contained client conduct, (for example, credit hazard), descriptive models distinguish an array of connections amongst clients or items. It categorizes the clients by their preferences of items and life phase rather than by the probability of clients on considering a specific task as in predictive models. Decision models depict the connection amongst all the components of a decision, the identified information (accounting outputs of predictive models), the decision, and the predicted outcomes of the decision so as to anticipate the outputs of decisions including numerous factors. Such models may be utilized in optimization, boosting definite results while limiting others. Figure 2 depicts the applications of predictive analytics in various fields broadly as businesses, science and industrial applications. Also, the methods and procedures employed to perform predictive analytics can generally be clustered into regression techniques and machine learning techniques. Further classifications of regression and machine learning practices are listed as follows in Table 2. A few of the listed techniques will be discussed further in detail.

Numerous predictive analytics tools are available both as an open-source tools (KNIME, Open NN, Orange and GNU Octave, etc.) and commercial tools (MATLAB, Minitab, STATA, SAP, Oracle data mining, etc.) [12] that are useful in processes decision making and integrating it into different operations.

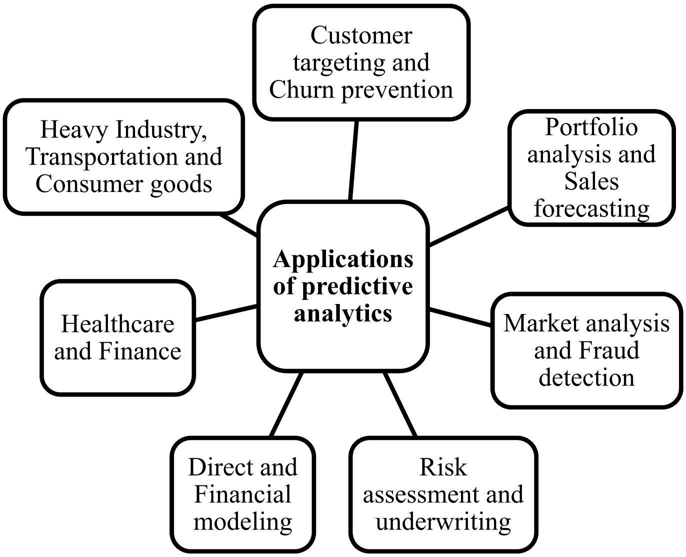

Customer targeting and Churn prevention
Heavy Industry, Transportation and Consumer goods
Portfolio analysis and Sales forecasting
Applications of predictive analytics
Healthcare and Finance
Market analysis and Fraud detection
Direct and Financial modeling
Risk assessment and underwriting

Fig. 2 Applications of predictive analytics

Table 2 Classifications of regression and machine learning

Regression techniques	ML techniques
• Linear regression model • Discrete choice model • Logistic regression • Multinomial logistic regression • Logit vs probit • Time series models • Probit regression • Classification and regression trees • Survival or duration analysis • Multivariate addaptive regression splines	• Radial basis functions • Multilayer perceptron (MLP) • Other Neural Networks • K-nearest neighbours • Naive bayes • Geospatial predictive modeling • Support vector machines

Prescriptive analytics supports answering inquiries regarding what has to be completed. By making use of bits of information from predictive analytics, that ended with data-driven selections. This licenses organizations to stick to choices in the face concerning susceptibility. Prescriptive analytics procedures rely on AI techniques that can ascertain proposals in huge datasets. By exploring previous selections and cases, the prospects of various results can be evaluated.

3.2 Data Mining

One of the activities of data analysis is ‘Data mining’. Data mining (the investigation phase of the ‘Knowledge Discovery in Databases—KDD’ process), a multidisciplinary subspecialty of software engineering nothing but the computational method of determining patterns in massive data sets comprise approaches at the connection of artificial intelligence, machine learning, statistics and database systems [13]. One can confuse data analysis with data mining. The main difference between these two are as follows:

- Data mining recognizes and finds a shrouded design in enormous datasets whereas data Analysis gives bits of knowledge or testing of hypothesis or model from a dataset.
- Data mining is one of the events in data analysis. Data analysis is a comprehensive set of events that deals with the assortment, planning and displaying of data for mining expressive understandings or information. Both are at times comprised as a subdivision of Business Intelligence.
- Data mining educations are typically on organized data. Data analysis should be possible on organized, semi-organized or unorganized data.
- The objective of data mining is to create data more practical while data analysis aids in demonstrating a theory or taking business choices.
- Data mining need not bother with any biased theory to distinguish the example or pattern in the information. Then again, data analysis tests a given theory.
- Data mining depends on numerical and logical methods to recognize patterns or drifts wherein data analysis utilizes business intelligence and analytics models.

Data mining includes six collective modules of tasks as Anomaly detection (finding unfamiliar data sets), Association rule learning (search of correlation between variables), Clustering (act of determining sets and assemblies in data), Classification (act of simplifying known structure to new data), Regression (identifying a task that prototypes the data with the minimum blunder) and Summarization (giving an added solid depiction of the data set, comprising conception and documentation). The wide range applications of data mining focus on human rights, games, science and engineering, medical data mining, sensor data mining, visual data mining, spatial data mining, surveillance, music data mining, pattern mining, knowledge grid, temporal data mining, business and subject-based data mining.

4 Big Data

Big data is an area which breaks down ways, deliberately isolate data from, or in any case achieve data sets that are markedly massive or multifaceted to be accomplished by conventional data-processing application software. Big data encounters apprehending data, storage, data investigation, search, sharing, moving, representation, querying, refreshing, data protection and source. The word big data repeatedly states just to the practice of predictive analytics or further definite innovative approaches to excerpt worth from data, and seldom to a precise size of data set. Precision in big data may possibly pave a way to more assured decision-making and enhanced conclusions can mean better operative efficiency, cost declines and reduced threat. Researchers, business chiefs, clinical experts, publicizing and governments routinely meet troubles with enormous data sets in regions encompassing Internet look, fintech, metropolitan informatics, and business informatics. Scholars face restraints in e-Science work, which includes meteorology, genomics, connectomics, intricate material science models, science and ecological study. The world's innovative per-capita ability to store data has largely grown like clockwork since the 1980s as of 2012; consistently 2.5 Exabytes (2.5×10^{18}) of information [14] were made as seen from Fig. 3. The trial for enormous undertakings is guessing out who should claim big data activities that ride the entire connotation.

Big data is a group of data from several sources, frequently described by the 3Vs namely Volume (quantity of data), Variety (data category) and Velocity (the rate at which the data generation happens). Over time, other Vs namely Veracity (quality of captured data), Value (business worth of the collected data) and Variability (inconsistency that hinders the process) have been added to descriptions of big data. Data management is quite a complex process when huge amount of data reaches from various sources. To offer valuable understanding to the data management and

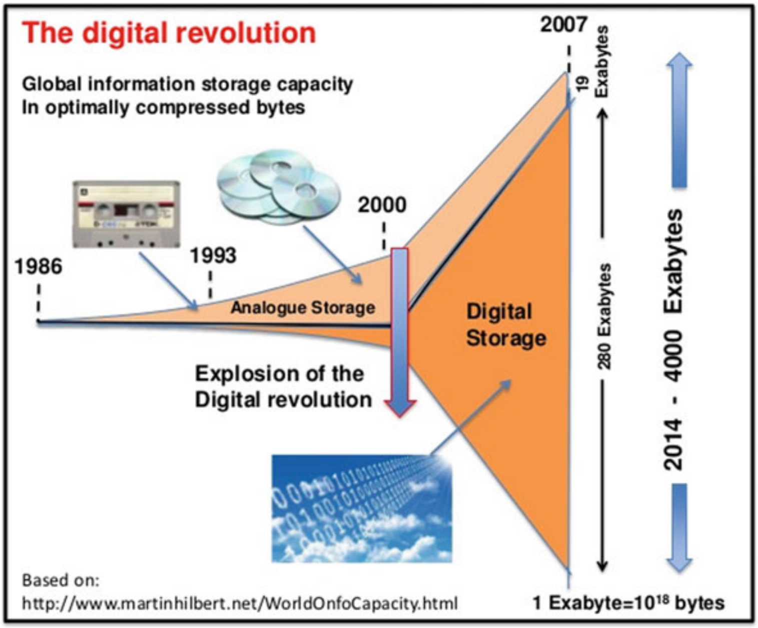


Fig. 3 Evolution and digitization of global data storage capacity (Reproduced from M. Hilbert and P. López, 2011) [14]

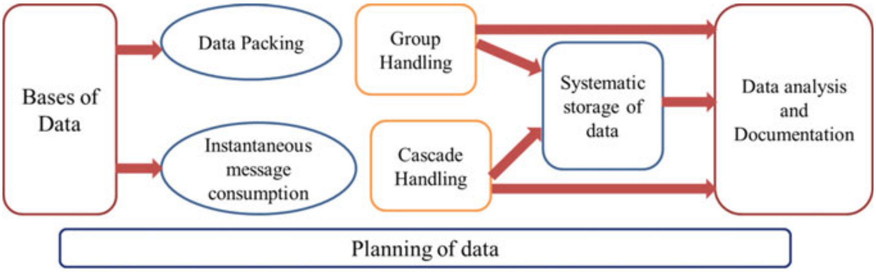

Bases of Data
Data Packing
Instantaneous message consumption
Group Handling
Cascade Handling
Systematic storage of data
Data analysis and Documentation
Planning of data

Fig. 4 Architecture of Big Data

increase correct content, data has to be handled with modern tools (analytics and algorithms) to produce expressive information.

Big data architecture denotes the rational and physical arrangement that commands how an extraordinary amount of data are consumed, processed, stored, accomplished and retrieved. Big data architecture is the base for big data analytics. The architecture mechanisms of big data analytics naturally contain four rational layers and execute four key processes as depicted in Fig. 4. 1. Big Data Sources Layer (managing both batch and real-time processing of big data such as data warehouses, SaaS applications and Internet of Things (IoT) devices), 2. Management & Storage Layer (receiving, converting and storing of data to the suitable format of data analytics tool), 3. Analysis Layer (extraction of business intelligence or BI from the storage layer) and 4. Consumption Layer (collects outputs from the analysis layer and presents them to the appropriate BI layer) [15].

Big data has originated numerous applications in several areas. The key areas where big data is being utilized are as follows. Government and private sectors, Social media analytics, Technology, Fraud detection, Call center Analytics, Banking, Agriculture, Marketing, Smartphones, Education, Manufacturing, Telecom and healthcare.

5 Supervised Learning

Supervised learning involves concluding a function from labeled training data using machine learning activity. The training data contains a set of training samples. In supervised learning, every sample is a duo involving an input item (usually a vector) and a preferred output value (also known as supervisory signal). A supervised learning algorithm examines the training data and creates a contingent function that may be utilized for representing new samples. An ideal situation allows the algorithm to properly define the class labels for hidden occurrences. This needs the learning algorithm to simplify from the training data to hidden occurrences in a ‘realistic’ way. In order to resolve an assigned difficulty of supervised learning, the following steps are to be followed.

1. Identifying the type of training samples
2. Collection of training set
3. Identifying the input feature illustration of learned function
4. Identifying the structure of learned function and the suitable learning algorithm
5. Completion of design on running the algorithm with the collected training set
6. Assessing the correctness of the learned function.

Four key concerns to be considered in supervised learning are (i) *Bias-variance tradeoff* [10]—A learning algorithm with small predisposition should be 'flexible' so as to perfectly fit the data. But if the learning algorithm is excessively supple, it will suit every training data set in a different way, and hence have high variance, (ii) *Function complication and volume of training data*—this issue concerns about the quantity of available training data with the complication of the function (classifier or regression), i.e. simpler the function needs a learning from small quantity of data wherein complex function requires massive quantity of training data, (iii) *dimensionality of the input space*—it depends on the dimension of the input feature vectors since extra dimensions can complicate the learning algorithm that will have more variance and (iv) *Noise in the output values*—this issue concerns about the amount of noise in the preferred output values. If the output values are improper owing to man-made or sensor errors then matching the training samples will not be efficient leads to overfitting. There are numerous algorithms in use to determine the noise in the training samples preceding to the supervised learning algorithm.

In general, all machine learning algorithms have a common principle in which it works that is they are defined as learning a target function (f) that maps the input (X) with the output values (Y) and making it predict Y for a new value of X and the relation is given as follows in Eq. (1).

$$Y = f(X) + e \tag{1}$$

There will also be an error (e) which is independent of X and this error is considered to be an irreducible error no matter how good we get the target function. A supervised learning algorithm also works on this principle. The most extensively used learning algorithms are linear regression, naive Bayes, logistic regression, Support Vector Machines, k-nearest neighbor algorithm, Neural Networks (MLP), decision trees, linear discriminant analysis and Similarity learning.

The various applications of supervised learning are widely in use in major areas such as Bioinformatics, Cheminformatics, Database marketing, Handwriting recognition, Information extraction, Pattern recognition, Speech recognition, Spam detection, Downward causation in biological system and object recognition in computer vision, etc.

6 Unsupervised Learning

Unsupervised learning is a sort of AI that searches for formerly hidden configurations in a data set with no prior labels and with at least manual oversight. Contrary to supervised learning that typically utilizes human-labeled data, unsupervised learning otherwise called as self-association takes into consideration displaying of likelihood densities over data sources [16]. In unsupervised learning, the two fundamental techniques that are utilized namely cluster analysis and principal component analysis. Cluster analysis is utilized in unsupervised learning to gather, or partitioned datasets with common features owing to generalize algorithmic connections. Cluster analysis is a subdivision of machine learning that assembles the data that has not been named, ordered or classified. Rather than reacting to feedback, cluster analysis recognizes unities in the data and responds depending on the existence or nonexistence of such unities in each fresh portion of data. This methodology aids identify abnormal data points that do not apt for any group.

A fundamental use of unsupervised learning is in the area of density estimation in statistics; however unsupervised learning incorporates a lot of areas relating to briefing and clarifying data features. In contrast to supervised learning that uses conditional probability distribution p_x $(x \mid y)$ trained on the y of input data whereas unsupervised learning uses a priori probability distribution p_x (x).

Number of most general algorithms are used in unsupervised learning and each approach practices numerous techniques as follows: 1. Clustering (ex: OPTICS algorithm, k-means, hierarchical clustering, etc.), 2. Irregularity detection (ex: local outlier factor and isolation forest), 3. Neural networks (ex: autoencoders, hebbian learning, deep belief nets, etc.) and 4. Approaches for learning latent variable models (ex: method of moments, expectation-maximization or EM algorithm, blind signal separation techniques, etc.) [6].

7 Reinforcement Learning

Reinforcement learning (RL) is a part of machine learning that dealt with how software agents should take movements in a setting to exploit the idea of accumulative return. It is one amongst the three common machine learning models, along with supervised learning and unsupervised learning. In comparison with supervised learning, RL doesn’t require labeled input or output values and also not in need of sub-optimal activities to be adjusted rather it aids in identifying stability between the investigation of unexplored area and manipulation of existing knowledge. Due to its simplification, reinforcement learning is considered in various disciplines like game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence and statistics. For instance, in operation research and control literature, RL is termed as neuro-dynamic programming. The glitches of attention in RL had been deliberate in the optimal control theory where it mainly focuses on the presence and categorization of optimal solutions and precise computation of algorithms.

Reinforcement learning is mainly compatible to glitches that contain a long-term versus short-term prize trade-off [17]. It is proven to be effective for many problems, comprising robot control, elevator scheduling, telecommunications, backgammon and checkers. Two factors that create reinforcement learning influential, i.e. the usage of samples to enhance performance and the usage of function approximation to deal with huge settings. A simple reinforcement learning model comprises of:

1. A group of environment states;
2. A group of actions;
3. Guidelines for movement between states;
4. Guidelines that define the scalar instant return of a movement; and
5. Guidelines that designate the observation of agent.

8 Decision Tree

A decision tree is a supporting technique that helps to make decisions using a kind of tree-like prototype of choices and their possible significance, comprising chance event results, source costs, and service. It showcases a supervised machine learning algorithm that has only restricted statements. Decision trees are generally utilized in operations research, especially in decision analysis, to support finding an approach most probable to attain an objective, but are also a widely used tool in machine learning and data mining [18]. This technique aims to make a model which forecasts the worth of a target variable/output depending on various input parameters. Das et al. reported hot rolled steel plate classification [19], obtained from CART analysis, for quality check based on chemical composition. A similar decision tree showing composition-based classification of strength of mild steel is given in Fig. 5.

Any tree model will have a *root node* that helps to divide the data into two or more sets. The key attribute of this node is chosen by using attribute selection measure (ASM) technique. *Branch* is the portion of the complete decision tree otherwise named as sub-tree. Arrowheads are used to distribute a node to two or more sub-nodes depending on if-else conditions and the process is called *splitting*. *Decision*

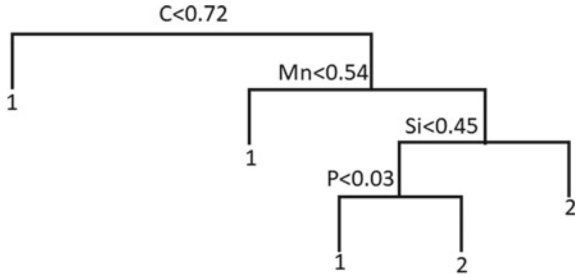


Fig. 5 Optimal decision tree for mild steel plates for classifying low (1) and high strength

node is the one on splitting the sub-nodes into successive sub-nodes. *Leaf or terminal node* is the conclusion of the decision tree in which a sub-node can't be split further. *Pruning* is the process of eliminating a sub-node from a tree.

Decision trees that are utilized in data mining are of two kinds. Tree models in which the target variable can yield a fixed set of values known as *classification trees.* Using these tree models, leaves of the tree signify class labels and branches of the tree signify combinations of sorts that initiate those class labels. Tree models where the target variable can utilize continuous values (usually real numbers) are called *regression trees.* In the analysis of decision, a decision tree can be utilized to visually and clearly denote decisions and decision making. In case of data mining, a decision tree defines data but not decisions; instead for decision making, the subsequent classification tree shall be made use of as an input. The combination of the above two kinds under a single roof is termed as Classification and Regression Tree (CART).

In order to reduce the data used in data mining, ASM technique is widely in practice that helps various algorithms for finding the best attributes. Two key types of ASM techniques are *Gini index and information gain.* Gini Index is the quantity of degree of possibility of a specific variable that is categorized incorrectly. The mathematical relation of a Gini index is given in Eq. (2)

$$Gini = 1 - \sum_{i=1}^{n} (p_i^2) \tag{2}$$

where p_i denotes the probability of classifying an object in a specific class. Normally a feature with a minimum Gini index is chosen if Gini index is used as the condition for an algorithm. Information gain or ID3 algorithm is the one that helps to reduce the level of entropy from root node to leaf node that aid to identify an attribute which yields thorough evidence about a class and the same has been expressed in Eq. (3).

$$E(s) = \sum_{i=1}^{c} (-p_i log_2 p_i) \tag{3}$$

where p_i denotes the probability of entropy 'E(s)'. Normally a feature with a maximum ID3 gain is utilized as the root for splitting. Some of the notable decision tree algorithms under a wide classification are Conditional Inference Trees, ID3 (Iterative Dichotomiser 3), MARS, C4.5 (successor of ID3), CHAID (CHi-squared Automatic Interaction Detector), CART (Classification and Regression Tree), etc.

9 Least Squares

The least squares method is a statistical technique to identify the best fit for a set of data points by reducing the entirety of the squares of the residuals of points from the curve. It is a typical method in regression analysis that predicts the performance of dependent variables in relation to the independent variables. The utmost significant application is in data fitting. The best fit in the least squares limits the entirety of squared residuals being the difference between an observed value and the fitted value provided by a model. At the point when the issue has generous uncertainties in the independent variable (X variable), at that point simple regression and least squares techniques have issues; in such cases, the approach needed for fitting errors-in-factors models are deemed better than that for least squares.

In regression analysis, dependent variables are plotted on the y-axis, while independent variables are plotted on the x-axis. These descriptions will give the equation for the line of best fit as depicted in Fig. 6, which is determined from the least squares method. In contradiction of a linear problem that has a definite solution, a non-linear least squares problem has no definite solution and is usually resolved by iteration on approximating it as a linear one. Polynomial least squares define the difference in a predicted value of the dependent variable as an independent variable function and the deviances from the fitted graph. The least square methods developed in the areas of

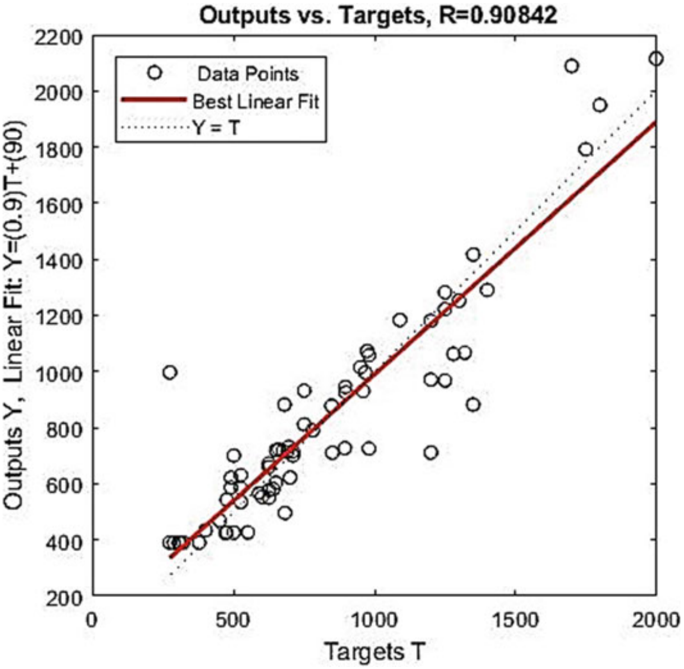


Fig. 6 Best linear fit

astronomy and geodesy through the course of the eighteenth century, where experts and statisticians wanted to deliver answers to the experiments of circumnavigating the Earth's oceans in the Era of Investigation. In 1795, German mathematician Carl Friedrich Gauss revealed the process of the least squares method but only in 1805, it was first printed by a French mathematician Adrien-Marie Legendre who described it as a numerical method for fitting linear equations to data on demonstrating a new procedure on evaluating the similar data as Laplace for the world's shape. However after 1809, Gauss brought in a new development on the method of least squares with the principles of probability, probability density, normal distribution and method of estimation. In 1810, with the base of Gauss's work, Laplace come up with Central limit theorem and in 1822, Gauss formulated Gauss-Markov Theorem [6]. Likewise many researchers have come up with various ways of implementing least squares. A problem will be defined based on an objective function which has 'm' adjusting variables of a model function defined by vector 'β' to best fit a 'n'data set that contains 'x_i' independent variable and 'y_i' dependent variable. The fit of the model is given by residuals 'r_i' as follows which is the difference between the actual values of 'y' to the predicted value of 'y'as given in Eq. (4).

$$r_i = y_i - f(x_i, \beta) \tag{4}$$

The least square methods determine the optimum variable by bringing down the sum 'S' of squared residuals and is shown below in Eq. (5)

$$\mathrm{S} = 1 - \sum_{i=1}^{n} (r_i^2) \tag{5}$$

The formulation of the regression has its limitations as to considering only the observational errors in the dependent variable. There are two relatively dissimilar situations with different inferences as Regression for prediction and Regression for fitting an 'accurate relation'. The common ways for cracking the least square problem are by linear least squares and non-linear least squares [20]. The difference between linear least squares and non-linear least squares are 1. The model function, f, in LLSQ (linear least squares) is a linear arrangement of variables of the form $f = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots$ The prototype may signify a straight line, a parabola or any other linear combination of functions. In NLLSQ (nonlinear least squares) the variables seem as functions, such as β^2, $e^{\beta x}$ and so forth. If the derivatives $\frac{\partial f}{\partial \beta_j}$ are one as constant or the other be influenced by only on the independent variable, the variables shows the model is linear. Else it is a nonlinear model. 2. Want primary values for the variables to identify the result to a NLLSQ problem; LLSQ doesn't necessitate them. 3. In LLSQ the result is distinctive, but in NLLSQ there may be numerous minima in the sum of squares. A distinct set of global least squares termed *weighted least squares* happens when the whole of the off-diagonal entries of the residual's correlation matrix become void; the differences of the observations may even be uneven.

10 Linear Regression

Linear regression is a method for exhibiting the correlation between a dependent variable (y) and one or many independent variables (x) [21]. A model that has only one independent variable is known as simple linear regression and in case of more than one independent variable it is known as multiple linear regression. It is merely different from multivariate linear regression which expects various associated dependent variables instead of single dependent variable. Linear regression emphasizes on the restricted probability distribution of the independent variables given by the function of the model instead of the combined probability distribution of all those variables nothing but the area of multivariate analysis. It has several everyday applications involving both statistics and machine learning due to its models based on the linearly unknown variables that can fit easily rather than the models with non-linear variables and also it is at ease to find the numerical properties of the subsequent estimators. There are several techniques that train the linear models and the most familiar is known as least squares but there are other approaches of fitting the model, for example, Ordinary least squares or Gradient descent or L^1 regularization and L^2 regularization [6]. Hence least squares and linear model are diligently related but not identical in meaning.

A linear regression model is represented by a linear equation that connects a particular set of input variables (x) that gives the results of predicted output (y) for the set of x. A coefficient 'β' is allotted to each of its inputs in a linear equation as a scale factor. Also, an added supplementary coefficient called bias coefficient or intercept that provides the line as in Fig. 6 which is an example for a simple regression line has a degree to move freely on a 2D plot. A typical regression equation with one input and an output is given in Eq. (6)

$$y = \beta_0 + \beta_1 x \tag{6}$$

The complication of a linear regression model depends on the number of coefficients used in it. For example, if a coefficient is zero then it neglects the effect of that input variable and thereafter the prediction of the model. This is common in regularization methods which could modify the algorithm to simplify the complexity of the model on making an entire size of the coefficients to zero.

One has to take a call before interpreting the outcomes of the regression model where each of them may follow the *unique effect* (estimated change in the predicted output for a change in a single input where all the other covariates are held stable) or the *marginal effect* (entire derivative of predicted output relating to input). There are two possibilities that may occur while interpreting the results of regression, i.e. where if the marginal effect is huge then the unique effect is zero or if the unique effect is huge then the marginal is zero. However, there is a chance of failure for multiple regression analysis to have the correlation between the predicted output with the input since the unique effect deals with a multifaceted system where lot of interconnected constituents influences the input variable.

There are lot of additions of linear regression have been established which involves simple and multiple linear regression, general linear models, generalized linear models (GLMs), heteroscedastic models, hierarchical linear models, measurement error models, and so on. It is essential to estimate the parameter and implication in linear regression. Few of the broad estimation approaches are Least squares estimation (ex: ordinary least squares, Generalized least squares, percentage least squares, total least squares, etc.), Maximum likelihood estimation (Ex: ridge regression, lasso regression, adaptive estimation, least absolute deviation) and other miscellaneous estimation approaches (ex: Bayesian linear regression, principal component regression, Quantile regression and Least angle regression). Linear regression has its major applications in the field of finance, economics, environmental science and epidemiology in order to define the suitable correlations between the parameters.

11 Neural Networks

A chain of algorithms that replicate the actions of a human brain to describe the correlation between numerous set of data is called Neural Network (NN). Architecture of neural network is same as that of human brain's which has 'Neurons' may be a biological or artificial neurons acting as a numerical function that gathers and categorizes data in relation to a particular architecture [22]. Since 1943 till late 2000, neural networks have shown tremendous development in artificial intelligence. The evolution of NN follows as right from a computational model called threshold logic on the basis of algorithms and mathematics that focuses on brain's genetic processes and application of NN to AI. Later a Hebbian learning based on hypothesis was created and applying it with B-type machines that follow the unsupervised learning [11]. After which use of calculators as computational machines that mimic the Hebbian network was created. A two layer computer learning network algorithm was created for the recognition of pattern followed by the development of back propagation algorithm in machine learning which sorted out the issue of NN in solving the processing of circuit with mathematical notation and processing power of earlier computers. Other evolution like support vector machines and few easier methods like linear classifiers surpassed NN in machine learning admiration. Later deep learning has transformed a new attention in neural networks. Since 2006 and till date, further the developments of NN is incredible in the new era of digital computing such as feedforward NN, long short-term memory (LSTM) in pattern recognition, traffic sign recognition, molecules identification for new drugs and so on.

In order to solve artificial intelligence (AI) problems, a neural network with artificial neurons called artificial neural network (ANN) is used [23]. Each network has a solid similarity to statistical methods like curve fitting and regression analysis. Layers (input, hidden and output) of interrelated nodes constitute a basic artificial neural network as shown in Fig. 7. Alike multiple linear regression, every node of a network called a perceptron that converts as a nonlinear activation/transfer function by passing the signal given by a multiple linear regression, i.e. a neuron of ANN

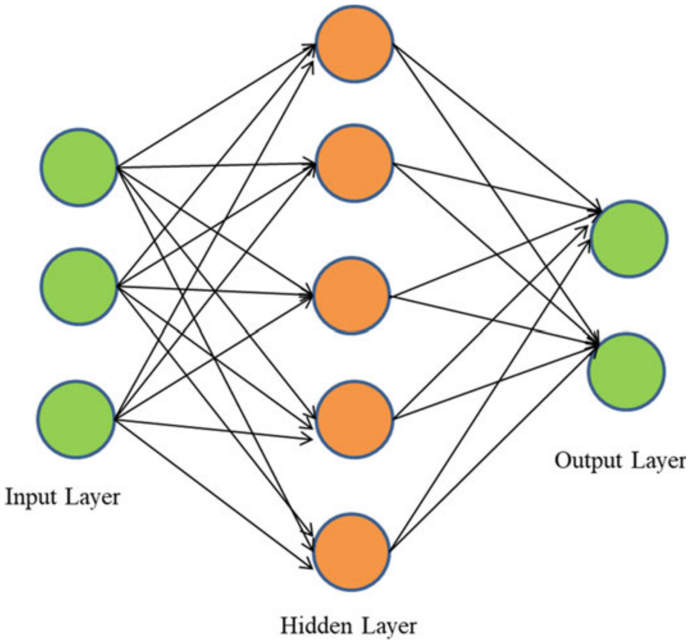


Fig. 7 Architecture of ANN

takes a signal probably a real number then processes the same and gives signal to the neurons attached to it.

In the course of learning process, neurons and the connectors generally have a weight on each input, i.e. product of inputs and weights in a transfer function that increases or decreases the signal's strength with a threshold level and the signal passes only if the cumulative signal crosses the threshold. There are different transfer functions in use and few of them are hard limit transfer function, pure linear transfer function, log-sigmoid and Tan-sigmoid transfer functions, etc. Numerically, f(x) is a function of a neuron which is a structure of another function g(x) and that can further be a structure of other functions which is completely denoted as a structure of a network that shows the relations between the variables. A typically used structure of a function is the weighted sum non-linear function which is given by the relation as in Eq. (7),

$$\mathrm{f(x)} = U\left(\sum_{i=0}^{n} w_i g_i(x)\right) \tag{7}$$

where U is the activation function such as tan-hyperbolic. Learning of ANN is under three key paradigms namely supervised learning, unsupervised learning and reinforcement learning which was explained in detail in the earlier sessions. The training of neural networks are done by utilizing the widely used methods like simulated annealing, particle swarm optimization, expectation maximization, evolutionary methods, genetic programming and non-parametric methods. The applications of ANN are broadly into the following categories: data processing (filtering, clustering, etc.), robotics (guiding prosthesis and manipulators), classification (recognition of sequence and pattern), regression analysis/function approximation (fitness modeling and approximation, prediction of time series), control (process control, computer numeric control and vehicle control) and computational and notional neuroscience.

12 Cluster Analysis

Cluster analysis is an approach that is utilized to arrange set of data/articles into related collections/groups named clusters. It is otherwise known to be classification analysis or clustering or numerical taxonomy. Here, there is no earlier data available about the group or cluster relationship for any of the articles. In clustering, objects are isolated into groups (clusters) with the aim that each object is more like the different objects within the same cluster rather than the objects outside the cluster [6]. Figure 8 depicts the clustering of iris flowers based on petal length and width grouped into different colors. Cluster analysis includes the problem formulation, choosing a separation measure, choosing a clustering method, finalizing the amount of clusters, interpretation of clusters and lastly evaluating the strength of clustering. Therefore cluster analysis can be expressed as a problem of multi-objective optimization that involves an iterative process of detecting knowledge with lot of trials and letdowns rather than the automatic process.

The concept of a ‘cluster’ may not be exactly distinct, which calls in for various clustering algorithms which vary meaningfully in their understanding/properties of

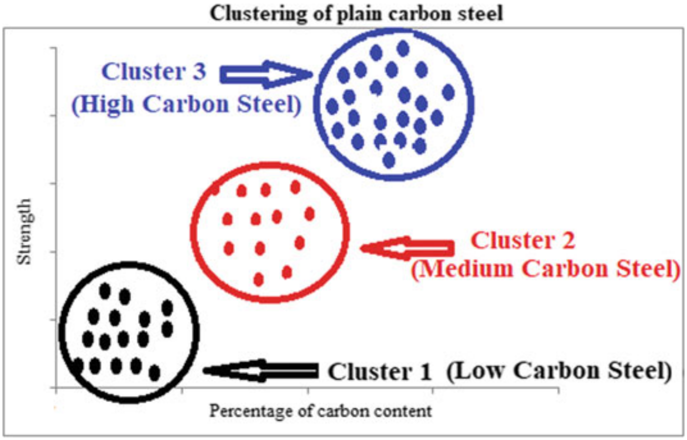

Clustering of plain carbon steel
Cluster 3
(High Carbon Steel)
Cluster 2
(Medium Carbon Steel)
Cluster 1 (Low Carbon Steel)
Strength
Percentage of carbon content

Fig. 8 Typical Cluster analysis of plain carbon steel

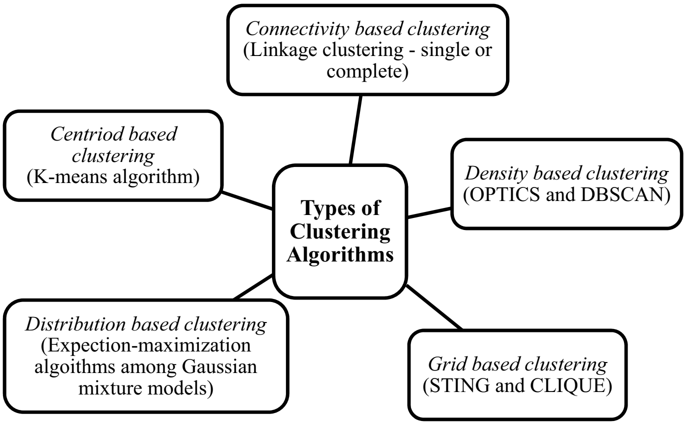

Connectivity based clustering
(Linkage clustering - single or complete)
Centriod based clustering
(K-means algorithm)
Types of Clustering Algorithms
Density based clustering
(OPTICS and DBSCAN)
Distribution based clustering
(Expection-maximization algoithms among Gaussian mixture models)
Grid based clustering
(STING and CLIQUE)

Fig. 9 Types of clustering algorithms

a cluster establishment and the process involving how to identify them competently. Hence it is required to study the differences between these algorithms on considering a classic cluster model that involves neural models, signed graph models, density models, group models and subspace models [24]. The classification of clustering are broadly as *hard clustering* (verifying all objects fits into a cluster or not) *and soft clustering* (verifying all objects fits into all the clusters to a definite level). It is further classified as hierarchical clustering (data of a child group also matches with parent), Strict portioning clustering with (each data matches with only one group) and without outliers (data matches with no groups), subspace clustering (overlying of data not possible within a subspace) and overlapping clustering (data matches with more than one group).

As stated above, clustering algorithms are classified according to the cluster models and the various types are depicted in Fig. 9. Connectivity-based clustering is according to the distance at which the objects are connected to form clusters more likely being linked to closer objects rather than the farther one. The distances are calculated by using the linkage principle. Clustering the objects with lesser distance is known as single-linkage clustering and clustering the objects with higher distance are known as complete linkage clustering. Centroid-based clustering is denoted by a principal vector called cluster vector that need not be associated with the data set. A specific approach called K-means clustering algorithm where a numerous amount of clusters are fixed to ‘k’ which is to be indicated in advance that provides a definite solution to the optimization problem by finding the k cluster and allot the objects to the nearby cluster center wherein it minimizes the square of distances from the cluster. A cluster model more relevant to statistics is done by distribution model based clustering that has the objects from the same distribution. It looks like how the artificial data sets are developed by selecting arbitrary objects from the distribution. The major issue in this method is overfitting and it is taken care by a method called Gaussian mixture models especially expectation minimization algorithm where modelling of dataset is done by fixing the number of Gaussian distributions that are adjusted in random and the variables are optimized in iteration to have a best fit of data which will meet a local optima. Density-based clustering have clusters stated as parts of greater density than the rest of the data set. Objects lies in scarce areas which are needed to isolate clusters are taken as noise and margin points. The most familiar method of density-based clustering is DBSCAN which is same as linkage-based clustering where it is according to the distance of connecting points within the threshold conversely it relates the points that fulfill the density criteria stated as a lesser amount of further objects within that radius.

Another generalized method of DBSCAN is OPTICS which neglects the choice to select a suitable value for the range variable and develops a hierarchical output based on linkage clustering. Grid-based clustering algorithm is utilized for multifaceted data set that develops a grid-like structure and comparing the same by means of grids or cells. It is a quite quicker method with lesser complex in computing. It involves this sequence of operations: initially splits the data set into number of determinate cells, chooses a cell at random, and finding the density of that cell. If the cell's density is higher than the threshold then marking such cell as a new cluster, calculating the density of neighbor cells and if the neighbor cells are higher than the threshold then keep the cell in the cluster and this step is repeated until there are no neighbor cells with higher density than the threshold. This process is performed repeatedly till every cell is passing through.

The cluster analysis approach is applied extensively in the field of biology, bioinformatics, medical imaging, business and marketing, computer science, World Wide Web, social science, robotics, finance, petroleum geology, and lot more.

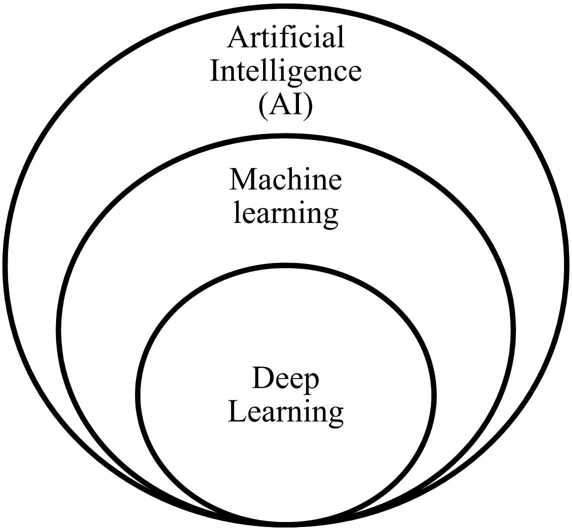
13 Deep Learning

Deep learning is one of the general techniques related to the artificial intelligence (AI) of supervised or unsupervised machine learning of data which is unstructured that mimics the handling of data by the human brain. Many of the deep learning models are built using artificial neural network (ANN). It is otherwise known as deep neural network. Deep learning is a course of machine learning algorithms that utilizes multiple layers in a network to predict correlation from the actual inputs with the target/output parameters that allow solving optimization problems in several practical applications. The architectures of deep learning shall be built layer by layer which aids to separate notions and choose the features that enhance the performance. Some of the architectures are deep belief networks (DBN), recurrent neural networks (RNN), convolutional neural networks (CNN) and deep neural networks (DNN). The term 'deep' addresses the number of layers that transform data from raw (Input) data to the target (Output) data using depth of credit assignment path (CAP) which defines the relation between the raw and target data [25]. For example, the CAP's depth in feedforward neural network is only one in addition to the number of hidden layers whereas it is merely limitless in CNN since a signal may pass over a layer more than once [6].

Most of the deep learning algorithms are structured as problems of unsupervised learning where such algorithms utilize the unlabeled data rather than the supervising learning. The best example of an unsupervised trained deep structure is deep belief network. Figure 10 depicts the revolution of deep learning that shows by what means deep learning is a subdivision of machine learning and in turn it is a subdivision of AI. Since 2012 till date, deep learning in ANN has evolved widely from various works of different researchers as predicting target of bio-molecular drug, detection of deadly effects of environmental chemicals and household goods, recognition of image and object, computer vision, recognition of speech and classification of images using CNN and long short term memory (LSTM) methods [26, 27].

An ANN with several layers between the layers of input and output is called as deep neural network (DNN). Intricate non-linear relations can be modeled using DNN. Figure 11 shows the difference between number of layers in a typical feedforward ANN and DNN. The working of DNN is very similar to ANN which was described in detail in the earlier sessions except that DNN as 'n' number of hidden layers between the input and output layers. For example, in computer chess game, the learning of different moves or tactics can be learned by a computer from several people and the same can be stored in its database and those tactics are determined by various algorithms and that is why it can be termed as deep neural network where the learning is deeper where ANN is not an imaginative method where it can draw a single result while DNN will be able to solve the issues universally and can predict or conclude based on the input and the desired output. Similar to ANN, DNN also has two major issues of computational time and overfitting if it is not trained thoroughly.

Fig. 10 Revolution of Deep learning



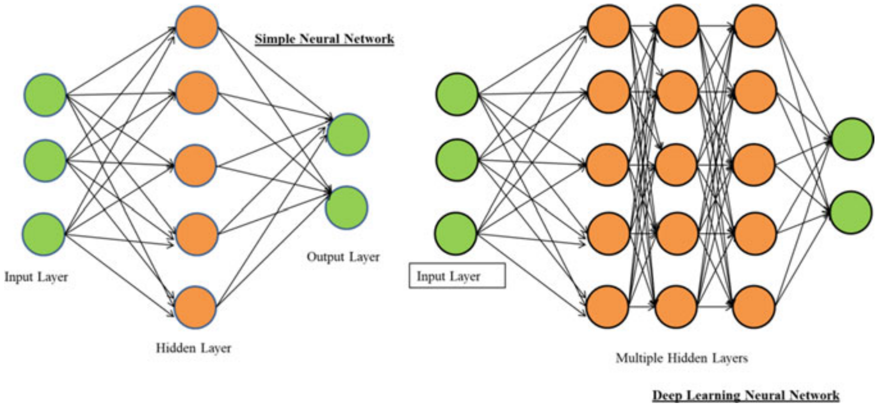


Fig. 11 A typical ANN and DNN architecture

The computation of several layers of a DNN with ‘n’ hidden layers is given by a relation as in Eq. (8)

$$f(x) = f\left[a^{(n+1)}(h^{n}(a^{n}(\ldots(h^{2}\left(a^{2}\left(h^{1}\left(a^{1}(x)\right)\right)\right)\right] \tag{8}$$

$a^{(n)}(x)$ is the Pre-activation function shown in Eq. (9) that is a linear process with the weighted matrix $W^{(n)}$ and $b^{(n)}$ as bias which shall be merged to a variable θ

$$a^{(n)}(x) = W^{(n)}x + b^{(n)} \tag{9}$$

The bar notation $\bar{x}$d enotes that ‘n’ attached to the vector x and $h^{(l)}(x)$ is the hidden layer activation/transfer function and is given by Eqs. (10) and (11).

$$a^{(n)}(\bar{x}) = \theta^{(n)}\bar{x}ifn = 1 \tag{10}$$

$$a^{(n)}\left(\bar{h}^{(n-1)}\right) = \theta^{(n)}\bar{h}^{(n-1)}ifn > 1 \tag{11}$$

A neural network where the data shall direct in any ways is categorized as recurrent neural networks (RNNs), a division of ANN in which the nodal linkages create a focused graph beside a temporary arrangement that shows a temporary lively performance and RNN is utilized in an application like modelling of language. In specific, an active algorithm that is in use for this purpose is long short-term memory. A neural network that is being utilized in computer vision application for evaluating pictorial imagery is convolutional deep neural networks (CNNs) which basically depends on the shared architecture and features of constant translation and it is also used for the recognition of automatic speech by modeling good acoustics. CNNs are generally kinds of multilayer perceptrons that typically of fully connected networks in which every neuron in a single layer is attached to all the neurons in the consecutive layer.

This leads to a chance of overfitting of data that can be sorted out by including some form of weight measurements method to the functional loss. It is not that extreme due to its connectivity of complex patterns with tiny and easier patterns to various regularization approaches.

There are several applications in which deep learning concepts are utilized. They are drug discovery and toxicology, bioinformatics [28], customer relationship management, recognition of electromyography (EMG) and images, processing of natural language and visual arts, mobile advertising, military applications, and detection of financial frauds, etc. [6].

14 Summary

ML is a method of learning from the data based on the principles of statistics and AI. AI replicates human intellect in computers, like learning and critical thinking. Basically, AI deals with knowledge depiction and knowledge engineering. AI involves different types of approaches, like search algorithms, mathematical optimization, evolutionary algorithms, logic programming and automated reasoning, probabilistic methods for uncertain reasoning, classifiers and statistical learning methods.

Data analytics which practices business intelligence and analytics models is one of the major application areas of ML. Data analytics has four classes, viz. descriptive, diagnostic, predictive and prescriptive analytics. Descriptive analytics supports answering studies concerning the whereabouts. Diagnostic analytics supports answering inquiries regarding why things occurred. Predictive analytics supports answering inquiries concerning the later events. Prescriptive analytics supports answering inquiries regarding what has to be completed.

ML can be classified into three major classes. Supervised learning can be used if every sample in the data has an input item and a preferred output value. Unsupervised learning searches hidden configurations in a data set with no prior labels or outputs. Reinforcement learning movements of the software agents to exploit the best accumulative return.

Deep learning is the newest form of machine learning, which utilizes multiple layers in a network to select the features and to find the correlation among them to develop classification as well as predictive models of highly complex systems.

References

1. Davy Cielen, M. A., & Meysman, A. (2016). *Introducing data science: Big data, machine learning, and more, using python tools*. United States: Manning Publications.
2. Langley, P. (2011). The changing science of machine learning. *Machine Learning, 82*(3), 275–279.
3. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, *1*, 39–48.

4. Shabbir, J., & Anwer, T. (2018). Artificial intelligence and its role in near future, *14*(8), 1–11.
5. Ginsberg, M. (2012). *Essentials of artificial intelligence.* San Francisco, CA, United States: Morgan Kaufmann Publishers Inc.
6. Dönmez, P. (2013). Introduction to machine learning. *Natural Language Engineering, 19*(2), 285–288.
7. Luger, W. (2004). George; stubblefield, *artificial intelligence: Structures and strategies for complex problem solving*, 5th ed. Benjamin/Cummings.
8. Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures, 90,* 46–60.
9. Johnston, J. (2010). *The allure of machinic life: cybernetics, artificial life, and the new AI*. Cambridge, Massachusetts London, England: The MIT Press.
10. Provost, R. K. F. (1998). Glossary of terms. *Machine Learning*, *30.* Springer US.
11. Le Roux, A., Bengio, N., & Fitzgibbon, N. (2012). Improving first and second-order methods by modeling uncertainty. In *Optimization for Machine Learning*, S. In Sra, Suvrit; Nowozin and S. J. Wright, Eds. MIT Press, 2012, p. 404.
12. Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*, 1st ed. Wiley.
13. Hand, D. J., & Adams, N. M. (2015). *Data mining in Wiley StatsRef: Statistics reference online* (pp. 1–7). Chichester, UK: John Wiley & Sons Ltd.
14. Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science (80), 332*(6025), 60–65, 2011.
15. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of 'big data' on cloud computing: Review and open research issues. *Information systems, 47*, 98–115.
16. Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1*(3), 295–311.
17. van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and markov decision processes. In *Reinforcement Learning. Adaptation, Learning, and Optimization*, van O. M. Wiering M., Ed. Springer, Berlin, Heidelberg, pp. 3–42.
18. Nilsson, N. J. (2005). Introduction to Machine Learning—an early draft of a proposed textbook. *Machine Learning, 56*(2), 387–399.
19. Das, P., Bhattacharyay, B. K., & Datta, S. (2006). A comparative study for modeling of hot-rolled steel plate classification using a statistical approach and neural-net systems. *Materials and Manufacturing Processes, 21*(8), 747–755.
20. Mannila, H. (1996). Data mining: machine learning, statistics, and databases. In *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management*, pp. 2–9.
21. Montgomery, G. G. V. D. C., & Peck, E. A. (2012). *Introduction to linear regression analysis*, 5th ed. Wiley.
22. Perry, S. W. (2002). Handbook of neural network signal processing. *Journal of the Acoustic Society of America, 111*(6), 2525–2526.
23. Prajapati, D. K., & Tiwari, M. (2017). Use of Artificial Neural Network (ANN) to determining surface parameters, friction and wear during pin-on-disc tribotesting. *Key Engineering Materials, 739,* 87–95.
24. Estivill-Castro, V. (2002). Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter, 4*(1), 65–75.
25. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
26. Li, X. & Wu, X. (2015). Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4520–4524.
27. Sze, V., Member, S., Chen, Y., Member, S., & Yang, T., Efficient Processing of deep neural networks: A tutorial and survey, pp. 1–32.
28. Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association, 24*(2), 361–370.