Machine Learning

Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions.[1] Within a subdiscipline in machine learning, advances in the field of deep learning have allowed neural networks, a class of statistical algorithms, to surpass many previous machine learning approaches in performance.[2]

ML finds application in many fields, including natural language processing, computer vision, speech recognition, email filtering, agriculture, and medicine.[3][4] The application of ML to business problems is known as predictive analytics.

Statistics and mathematical optimization (mathematical programming) methods comprise the foundations of machine learning. Data mining is a related field of study, focusing on exploratory data analysis (EDA) via unsupervised learning.[6][7]

From a theoretical viewpoint, probably approximately correct learning provides a framework for describing machine learning.

The term machine learning was coined in 1959 by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and artificial intelligence.[8][9] The synonym self-teaching computers was also used in this time period.[10][11]

Although the earliest machine learning model was introduced in the 1950s when Arthur Samuel invented a program that calculated the winning chance in checkers for each side, the history of machine learning roots back to decades of human desire and effort to study human cognitive processes.[12] In 1949, Canadian psychologist Donald Hebb published the book The Organization of Behavior, in which he introduced a theoretical neural structure formed by certain interactions among nerve cells.[13] Hebb's model of neurons interacting with one another set a groundwork for how AIs and machine learning algorithms work under nodes, or artificial neurons used by computers to communicate data.[12] Other researchers who have studied human cognitive systems contributed to the modern machine learning technologies as well, including logician Walter Pitts and Warren McCulloch, who proposed the early mathematical models of neural networks to come up with algorithms that mirror human thought processes.[12]

By the early 1960s, an experimental "learning machine" with punched tape memory, called Cybertron, had been developed by Raytheon Company to analyse sonar signals, electrocardiograms, and speech patterns using rudimentary reinforcement learning. It was repetitively "trained" by a human operator/teacher to recognize patterns and equipped with a "goof" button to cause it to reevaluate incorrect decisions.[14] A representative book on research into machine learning during the 1960s was Nilsson's book on Learning Machines, dealing mostly with machine learning for pattern classification.[15] Interest related to pattern recognition continued into the 1970s, as described by Duda and Hart in 1973.[16] In 1981 a report was given on using teaching strategies so that an artificial neural network learns to recognize 40 characters (26 letters, 10 digits, and 4 special symbols) from a computer terminal.[17]

Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."[18] This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turing's proposal in his paper "Computing Machinery and Intelligence", in which the question "Can machines think?" is replaced with the question "Can machines do what we (as thinking entities) can do?".[19]

Modern-day machine learning has two objectives. One is to classify data based on models which have been developed; the other purpose is to make predictions for future outcomes based on these models. A hypothetical algorithm specific to classifying data may use computer vision of moles coupled with supervised learning in order to train it to classify the cancerous moles. A machine learning algorithm for stock trading may inform the trader of future potential predictions.[20]

As a scientific endeavor, machine learning grew out of the quest for artificial intelligence (AI). In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics.[22] Probabilistic reasoning

was also employed, especially in automated medical diagnosis.[23]:488

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation.[23]:488 By 1980, expert systems had come to dominate AI, and statistics was out of favor.[24] Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming(ILP), but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval.[23]:708–710,755 Neural networks research had been abandoned by AI and computer science around the same time. This line, too, was continued outside the AI/CS field, as "connectionism", by researchers from other disciplines including John Hopfield, David Rumelhart, and Geoffrey Hinton. Their main success came in the mid-1980s with the reinvention of backpropagation.[23]:25

Machine learning (ML), reorganized and recognized as its own field, started to flourish in the 1990s. The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature. It shifted focus away from the symbolic approaches it had inherited from AI, and toward methods and models borrowed from statistics, fuzzy logic, and probability theory.[24]

There is a close connection between machine learning and compression. A system that predicts the posterior probabilities of a sequence given its entire history can be used for optimal data compression (by using arithmetic coding on the output distribution).

Conversely, an optimal compressor can be used for prediction (by finding the symbol that compresses best, given the previous history). This equivalence has been used as a justification for using data compression as a benchmark for "general intelligence".[25][26][27]

An alternative view can show compression algorithms implicitly map strings into implicit feature space vectors, and compression-based similarity measures compute similarity within these feature spaces. For each compressor C(.) we define an associated vector space ℵ, such that C(.) maps an input string x, corresponding to the vector norm ||~x||. An exhaustive examination of the feature spaces underlying all compression algorithms is precluded by space; instead, feature vectors chooses to examine three representative lossless compression methods, LZW, LZ77, and PPM.[28]

According to AIXI theory, a connection more directly explained in Hutter Prize, the best possible compression of x is the smallest possible software that generates x. For example, in that model, a zip file's compressed size includes both the zip file and the unzipping software, since you can not unzip it without both, but there may be an even smaller combined form.

Examples of AI-powered audio/video compression software include NVIDIA Maxine, AIVC.[29] Examples of software that can perform AI-powered image compression include OpenCV, TensorFlow, MATLAB's Image Processing Toolbox (IPT) and High-Fidelity Generative Image Compression.[30]

In unsupervised machine learning, k-means clustering can be utilized to compress data by grouping similar data points into clusters. This technique simplifies handling extensive datasets that lack predefined labels and finds widespread use in fields such as image compression.[31]

Data compression aims to reduce the size of data files, enhancing storage efficiency and speeding up data transmission. K-means clustering, an unsupervised machine learning algorithm, is employed to partition a dataset into a specified number of clusters, k, each represented by the centroid of its points. This process condenses extensive datasets into a more compact set of representative points. Particularly beneficial in image and signal processing, k-means clustering aids in data reduction by replacing groups of data points with their centroids, thereby preserving the core information of the original data while significantly decreasing the required storage space.[32]

Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on prediction, based on known properties learned from the training data, data mining focuses on the discovery of (previously) unknown properties in the data (this is the analysis step of knowledge discovery in databases). Data mining uses many machine learning methods, but with different goals; on the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. Much of the confusion between these two research communities (which do often have separate conferences and separate journals, ECML PKDD being a major exception)

comes from the basic assumptions they work with: in machine learning, performance is usually evaluated with respect to the ability to reproduce known knowledge, while in knowledge discovery and data mining (KDD) the key task is the discovery of previously unknown knowledge. Evaluated with respect to known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

Machine learning also has intimate ties to optimization: Many learning problems are formulated as minimization of some loss function on a training set of examples. Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances (for example, in classification, one wants to assign a label to instances, and models are trained to correctly predict the preassigned labels of a set of examples).[35]

Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.

Machine learning and statistics are closely related fields in terms of methods, but distinct in their principal goal: statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns.[36] According to Michael I. Jordan, the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics.[37] He also suggested the term data science as a placeholder to call the overall field.[37]

Conventional statistical analyses require the a priori selection of a model most suitable for the study data set. In addition, only significant or theoretically relevant variables based on previous experience are included for analysis. In contrast, machine learning is not built on a pre-structured model; rather, the data shape the model by detecting underlying patterns. The more variables (input) used to train the model, the more accurate the ultimate model will be.[38]

Leo Breiman distinguished two statistical modeling paradigms: data model and algorithmic model,[39] wherein "algorithmic model" means more or less the machine learning algorithms like Random Forest.

Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning.[40]

Analytical and computational techniques derived from deep-rooted physics of disordered systems can be extended to large-scale problems, including machine learning, e.g., to analyse the weight space of deep neural networks.[41] Statistical physics is thus finding applications in the area of medical diagnostics.[42]

A core objective of a learner is to generalize from its experience.[5][43] Generalization in this context is the ability of a learning machine to perform accurately on new, unseen

examples/tasks after having experienced a learning data set. The training examples come from some generally unknown probability distribution (considered representative of the space of occurrences) and the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases.

The computational analysis of machine learning algorithms and their performance is a branch of theoretical computer science known as computational learning theory via the probably approximately correct learning model. Because training sets are finite and the future is uncertain, learning theory usually does not yield guarantees of the performance of algorithms. Instead, probabilistic bounds on the performance are quite common. The bias–variance decomposition is one way to quantify generalization error.

For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of the function underlying the data. If the hypothesis is less complex than the function, then the model has under fitted the data. If the complexity of the model is increased in response, then the training error decreases. But if the hypothesis is too complex, then the model is subject to overfitting and generalization will be poorer.[44]

In addition to performance bounds, learning theorists study the time complexity and feasibility of learning. In computational learning theory, a computation is considered feasible if it can be done in polynomial time. There are two kinds of time complexity results: Positive results show that a certain class of functions can be learned in polynomial time. Negative results show that certain classes cannot be learned in

polynomial time.

Machine learning approaches are traditionally divided into three broad categories, which correspond to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system:

Although each algorithm has advantages and limitations, no single algorithm works for all problems.[45][46][47]

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs.[48] The data, known as training data, consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs.[49] An optimal function allows the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.[18]

Types of supervised-learning algorithms include active learning, classification and regression.[50] Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. Examples of regression would be predicting the height of a person, or the future temperature.[51]

Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

Unsupervised learning algorithms find structures in data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. Central applications of unsupervised machine learning include clustering, dimensionality reduction,[7] and density estimation.[52]

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more

predesignated criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated, for example, by internal compactness, or the similarity between members of the same cluster, and separation, the difference between clusters. Other methods are based on estimated density and graph connectivity.

A special type of unsupervised learning called, self-supervised learning involves training a model by generating the supervisory signal from the data itself.[53][54]

Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Some of the training examples are missing training labels, yet many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy.

In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.[55]

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines,

such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In reinforcement learning, the environment is typically represented as a Markov decision process (MDP). Many reinforcements learning algorithms use dynamic programming techniques.[56] Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

Dimensionality reduction is a process of reducing the number of random variables under consideration by obtaining a set of principal variables.[57] In other words, it is a process of reducing the dimension of the feature set, also called the "number of features". Most of the dimensionality reduction techniques can be considered as either feature elimination or extraction. One of the popular methods of dimensionality reduction is principal component analysis (PCA). PCA involves changing higher-dimensional data (e.g., 3D) to a smaller space (e.g., 2D).
The manifold hypothesis proposes that high-dimensional data sets lie along low-dimensional manifolds, and many dimensionality reduction techniques make this assumption, leading to the area of manifold learning and manifold regularization.

Other approaches have been developed which do not fit neatly into this three-fold categorization, and sometimes more than one is used by the same machine learning system. For example, topic modeling, meta-learning.[58]

Self-learning, as a machine learning paradigm was introduced in 1982 along with a neural network capable of self-learning, named crossbar adaptive array (CAA).[59][60] It gives a solution to the problem learning without any external reward, by introducing emotion as an internal reward. Emotion is used as state evaluation of a self-learning agent. The CAA self-learning algorithm computes, in a crossbar fashion, both decisions about actions and emotions (feelings) about consequence situations. The system is driven by the interaction between cognition and emotion.[61]

The self-learning algorithm updates a memory matrix W =||w(a,s)|| such that in each iteration executes the following machine learning routine:

It is a system with only one input, situation, and only one output, action (or behavior) a. There is neither a separate reinforcement input nor an advice input from the environment. The backpropagated value (secondary reinforcement) is the emotion toward the consequence situation. The CAA exists in two environments, one is the behavioral environment where it behaves, and the other is the genetic environment, wherefrom it initially and only once receives initial emotions about situations to be encountered in the behavioral environment. After receiving the genome (species) vector from the genetic environment, the CAA learns a goal-seeking behavior, in an environment that contains both desirable and undesirable situations.[62]

Several learning algorithms aim at discovering better representations of the inputs provided during training.[63] Classic examples include principal component analysis and cluster analysis. Feature learning algorithms, also called representation learning algorithms, often attempt to preserve the information in their input but also transform it in a way that makes it useful, often as a pre-processing step before performing

classification or predictions. This technique allows reconstruction of the inputs coming from the unknown data-generating distribution, while not being necessarily faithful to configurations that are implausible under that distribution. This replaces manual feature engineering, and allows a machine to both learn the features and use them to perform a specific task.

Feature learning can be either supervised or unsupervised. In supervised feature learning, features are learned using labeled input data. Examples include artificial neural networks, multilayer perceptrons, and supervised dictionary learning. In unsupervised feature learning, features are learned with unlabeled input data. Examples include dictionary learning, independent component analysis, autoencoders, matrix factorization[64] and various forms of clustering.[65][66][67]

Manifold learning algorithms attempt to do so under the constraint that the learned representation is low-dimensional. Sparse coding algorithms attempt to do so under the constraint that the learned representation is sparse, meaning that the mathematical model has many zeros. Multilinear subspace learning algorithms aim to learn low-dimensional representations directly from tensor representations for multidimensional data, without reshaping them into higher-dimensional vectors.[68] Deep learning algorithms discover multiple levels of representation, or a hierarchy of features, with higher-level, more abstract features defined in terms of (or generating) lower-level features. It has been argued that an intelligent machine is one that learns a representation that disentangles the underlying factors of variation that explain the observed data.[69]

Feature learning is motivated by the fact that machine learning tasks such as classification often require input that is mathematically and computationally convenient to process. However, real-world data such as images, video, and sensory data has not yielded attempts to algorithmically define specific features. An alternative is to discover such features or representations through examination, without relying on explicit algorithms.

Sparse dictionary learning is a feature learning method where a training example is represented as a linear combination of basis functions and assumed to be a sparse matrix. The method is strongly NP-hard and difficult to solve approximately.[70] A popular heuristic method for sparse dictionary learning is the k-SVD algorithm. Sparse dictionary learning has been applied in several contexts. In classification, the problem is to determine the class to which a previously unseen training example belongs. For a dictionary where each class has already been built, a new training example is associated with the class that is best sparsely represented by the corresponding dictionary. Sparse dictionary learning has also been applied in image de-noising. The key idea is that a clean image patch can be sparsely represented by an image dictionary, but the noise cannot.[71]

In data mining, anomaly detection, also known as outlier detection, is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data.[72] Typically, the anomalous items represent an issue such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are referred to as outliers, novelties, noise, deviations and exceptions.[73]

In particular, in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but unexpected bursts of inactivity. This pattern does not adhere to the common statistical definition of an outlier as a rare object. Many outlier detection methods (in particular, unsupervised algorithms) will fail on such data unless aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro-clusters formed by these patterns.[74]

Three broad categories of anomaly detection techniques exist.[75] Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal, by looking for instances that seem to fit the least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference from many other statistical classification problems is the inherently unbalanced nature of outlier detection). Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set and then test the likelihood of a test instance to be generated by the model.

Robot learning is inspired by a multitude of machine learning methods, starting from supervised learning, reinforcement learning,[76][77] and finally meta-learning (e.g. MAML).

Association rule learning is a rule-based machine learning method for discovering relationships between variables in large databases. It is intended to identify strong rules discovered in databases using some measure of "interestingness".[78]

Rule-based machine learning is a general term for any machine learning method that identifies, learns, or evolves "rules" to store, manipulate or apply knowledge. The defining characteristic of a rule-based machine learning algorithm is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learning algorithms that commonly identify a singular model that can be universally applied to any instance in order to make a prediction.[79] Rule-based machine learning approaches include learning classifier systems, association rule learning, and artificial immune systems.

Based on the concept of strong rules, Rakesh Agrawal, Tomasz Imieliński and Arun Swami introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets.[80] For example, the rule

{

o

n

i

o

n

s

,

p

o

t

a

t

o

e

s

}

⇒

{

b

u

r

g

e

r

}

{\displaystyle \{\mathrm {onions,potatoes} \}\Rightarrow \{\mathrm {burger} \}}

found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as promotional pricing or product placements. In addition to market basket analysis, association rules are employed today in application areas including Web usage mining, intrusion detection, continuous production, and bioinformatics. In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

Learning classifier systems (LCS) are a family of rule-based machine learning algorithms that combine a discovery component, typically a genetic algorithm, with a learning component, performing either supervised learning, reinforcement learning, or unsupervised learning. They seek to identify a set of context-dependent rules that collectively store and apply knowledge in a piecewise manner in order to make predictions.[81]

Inductive logic programming (ILP) is an approach to rule learning using logic programming as a uniform representation for input examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program that entails all positive and no negative examples. Inductive programming is a related field that considers any kind of programming language for representing hypotheses (and not only logic programming), such as

functional programs.

Inductive logic programming is particularly useful in bioinformatics and natural language processing. Gordon Plotkin and Ehud Shapiro laid the initial theoretical foundation for inductive machine learning in a logical setting.[82][83][84] Shapiro built their first implementation (Model Inference System) in 1981: a Prolog program that inductively inferred logic programs from positive and negative examples.[85] The term inductive here refers to philosophical induction, suggesting a theory to explain observed facts, rather than mathematical induction, proving a property for all members of a well-ordered set.

A machine learning model is a type of mathematical model that, once "trained" on a given dataset, can be used to make predictions or classifications on new data. During training, a learning algorithm iteratively adjusts the model's internal parameters to minimize errors in its predictions.[86] By extension, the term "model" can refer to several levels of specificity, from a general class of models and their associated learning algorithms to a fully trained model with all its internal parameters tuned.[87]

Various types of models have been used and researched for machine learning systems, picking the best model for a task is called model selection.

Artificial neural networks (ANNs), or connectionist systems, are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such

systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.

An ANN is a model based on a collection of connected units or nodes called "artificial neurons", which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit information, a "signal", from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called "edges". Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers multiple times.

The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.

Deep learning consists of multiple hidden layers in an artificial neural network. This approach tries to model the way the human brain processes light and sound into vision and hearing. Some successful applications of deep learning are computer vision and speech recognition.[88]

Decision tree learning uses a decision tree as a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining, and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data, but the resulting classification tree can be an input for decision-making.

Random forest regression (RFR) falls under umbrella of decision tree-based models. RFR is an ensemble learning method that builds multiple decision trees and averages their predictions to improve accuracy and to avoid overfitting. To build decision trees, RFR uses bootstrapped sampling, for instance each decision tree is trained on random data of from training set. This random selection of RFR for training enables model to reduce bias predictions and achieve accuracy. RFR generates independent decision trees, and it can work on single output data as well multiple regressor task. This makes

RFR compatible to be used in various application. [89][90]

Support-vector machines (SVMs), also known as support-vector networks, are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category.[91] An SVM training algorithm is a non-probabilistic, binary, linear classifier, although methods such as Platt scaling exist to use SVM in a probabilistic classification setting. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Regression analysis encompasses a large variety of statistical methods to estimate the relationship between input variables and their associated features. Its most common form is linear regression, where a single line is drawn to best fit the given data according to a mathematical criterion such as ordinary least squares. The latter is often extended by regularization methods to mitigate overfitting and bias, as in ridge regression. When dealing with non-linear problems, go-to models include polynomial regression (for example, used for trendline fitting in Microsoft Excel[92]), logistic regression (often used in statistical classification) or even kernel regression, which introduces non-linearity by taking advantage of the kernel trick to implicitly map input variables to higher-dimensional space.

Multivariate linear regression extends the concept of linear regression to handle

multiple dependent variables simultaneously. This approach estimates the relationships between a set of input variables and several output variables by fitting a multidimensional linear model. It is particularly useful in scenarios where outputs are interdependent or share underlying patterns, such as predicting multiple economic indicators or reconstructing images,[93] which are inherently multi-dimensional.

A Bayesian network, belief network, or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independence with a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning. Bayesian networks that model sequences of variables, like speech signals or protein sequences, are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.

A Gaussian process is a stochastic process in which every finite collection of the random variables in the process has a multivariate normal distribution, and it relies on a pre-defined covariance function, or kernel, that models how pairs of points relate to each other depending on their locations.

Given a set of observed points, or input–output examples, the distribution of the (unobserved) output of a new point as function of its input data can be directly

computed by looking like the observed points and the covariances between those points and the new, unobserved point.

Gaussian processes are popular surrogate models in Bayesian optimization used to do hyperparameter optimization.

A genetic algorithm (GA) is a search algorithm and heuristic technique that mimics the process of natural selection, using methods such as mutation and crossover to generate new genotypes in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms were used in the 1980s and 1990s.[95][96] Conversely, machine learning techniques have been used to improve the performance of genetic and evolutionary algorithms.[97]

The theory of belief functions, also referred to as evidence theory or Dempster–Shafer theory, is a general framework for reasoning with uncertainty, with understood connections to other frameworks such as probability, possibility and imprecise probability theories. These theoretical frameworks can be thought of as a kind of learner and have some analogous properties of how evidence is combined (e.g., Dempster's rule of combination), just like how in a pmf-based Bayesian approach[clarification needed] would combine probabilities. However, there are many caveats to these beliefs functions when compared to Bayesian approaches in order to incorporate ignorance and uncertainty quantification. These belief function approaches that are implemented within the machine learning domain typically leverage a fusion approach of various ensemble methods to better handle the learner's decision

boundary, low samples, and ambiguous class issues that standard machine learning approach tend to have difficulty resolving.[4][9] However, the computational complexity of these algorithms are dependent on the number of propositions (classes), and can lead to a much higher computation time when compared to other machine learning approaches.

Typically, machine learning models require a high quantity of reliable data to perform accurate predictions. When training a machine learning model, machine learning engineers need to target and collect a large and representative sample of data. Data from the training set can be as varied as a corpus of text, a collection of images, sensor data, and data collected from individual users of a service. Overfitting is something to watch out for when training a machine learning model. Trained models derived from biased or non-evaluated data can result in skewed or undesired predictions. Biased models may result in detrimental outcomes, thereby furthering the negative impacts on society or objectives. Algorithmic bias is a potential result of data not being fully prepared for training. Machine learning ethics is becoming a field of study and notably, becoming integrated within machine learning engineering teams.

Federated learning is an adapted form of distributed artificial intelligence to training machine learning models that decentralizes the training process, allowing for users' privacy to be maintained by not needing to send their data to a centralized server. This also increases efficiency by decentralizing the training process to many devices. For example, Gboard uses federated machine learning to train search query prediction models on users' mobile phones without having to send individual searches back to Google.[98]

There are many applications for machine learning, including:

In 2006, the media-services provider Netflix held the first "Netflix Prize" competition to find a program to better predict user preferences and improve the accuracy of its existing Cinematch movie recommendation algorithm by at least 10%. A joint team made up of researchers from AT&T Labs-Research in collaboration with the teams Big Chaos and Pragmatic Theory built an ensemble model to win the Grand Prize in 2009 for $1 million.[101] Shortly after the prize was awarded, Netflix realized that viewers' ratings were not the best indicators of their viewing patterns ("everything is a recommendation") and they changed their recommendation engine accordingly.[102] In 2010 The Wall Street Journal wrote about the firm Rebellion Research and their use of machine learning to predict the financial crisis.[103] In 2012, co-founder of Sun Microsystems, Vinod Khosla, predicted that 80% of medical doctors jobs would be lost in the next two decades to automated machine learning medical diagnostic software.[104] In 2014, it was reported that a machine learning algorithm had been applied in the field of art history to study fine art paintings and that it may have revealed previously unrecognized influences among artists.[105] In 2019 Springer Nature published the first research book created using machine learning.[106] In 2020, machine learning technology was used to help make diagnoses and aid researchers in developing a cure for COVID-19.[107] Machine learning was recently applied to predict the pro-environmental behavior of travelers.[108] Recently, machine learning technology was also applied to optimize smartphone's performance and thermal behavior based on the user's interaction with the phone.[109][110][111] When applied correctly, machine learning algorithms (MLAs) can utilize a wide range of company

characteristics to predict stock returns without overfitting. By employing effective feature engineering and combining forecasts, MLAs can generate results that far surpass those obtained from basic linear techniques like OLS.[112]

Recent advancements in machine learning have extended into the field of quantum chemistry, where novel algorithms now enable the prediction of solvent effects on chemical reactions, thereby offering new tools for chemists to tailor experimental conditions for optimal outcomes.[113]

Machine Learning is becoming a useful tool to investigate and predict evacuation decision making in large scale and small scale disasters. Different solutions have been tested to predict if and when householders decide to evacuate during wildfires and hurricanes.[114][115][116] Other applications have been focusing on pre evacuation decisions in building fires.[117][118]

Although machine learning has been transformative in some fields, machine-learning programs often fail to deliver expected results.[119][120][121] Reasons for this are numerous: lack of (suitable) data, lack of access to the data, data bias, privacy problems, badly chosen tasks and algorithms, wrong tools and people, lack of resources, and evaluation problems.[122]

The "black box theory" poses another yet significant challenge. Black box refers to a situation where the algorithm or the process of producing an output is entirely opaque,

meaning that even the coders of the algorithm cannot audit the pattern that the machine extracted out of the data.[123] The House of Lords Select Committee, which claimed that such an "intelligence system" that could have a "substantial impact on an individual's life" would not be considered acceptable unless it provided "a full and satisfactory explanation for the decisions" it makes.[123]

In 2018, a self-driving car from Uber failed to detect a pedestrian, who was killed after a collision.[124] Attempts to use machine learning in healthcare with the IBM Watson system failed to deliver even after years of time and billions of dollars invested.[125][126] Microsoft's Bing Chat chatbot has been reported to produce hostile and offensive response against its users.[127]

Machine learning has been used as a strategy to update the evidence related to a systematic review and increased reviewer burden related to the growth of biomedical literature. While it has improved with training sets, it has not yet developed sufficiently to reduce the workload burden without limiting the necessary sensitivity for the findings research themselves.[128]

Explainable AI (XAI), or Interpretable AI, or Explainable Machine Learning (XML), is artificial intelligence (AI) in which humans can understand the decisions or predictions made by the AI.[129] It contrasts with the "black box" concept in machine learning where even its designers cannot explain why an AI arrived at a specific decision.[130] By refining the mental models of users of AI-powered systems and dismantling their misconceptions, XAI promises to help users perform more effectively. XAI may be an

implementation of the social right to explanation.

Settling on a bad, overly complex theory gerrymandered to fit all the past training data is known as overfitting. Many systems attempt to reduce overfitting by rewarding a theory in accordance with how well it fits the data but penalizing the theory in accordance with how complex the theory is.[131]

Learners can also disappoint by "learning the wrong lesson". A toy example is that an image classifier trained only on pictures of brown horses and black cats might conclude that all brown patches are likely to be horses.[132] A real-world example is that, unlike humans, current image classifiers often do not primarily make judgments from the spatial relationship between components of the picture, and they learn relationships between pixels that humans are oblivious to, but that still correlate with images of certain types of real objects. Modifying these patterns on a legitimate image can result in "adversarial" images that the system misclassifies.[133][134]

Adversarial vulnerabilities can also result in nonlinear systems, or from non-pattern perturbations. For some systems, it is possible to change the output by only changing a single adversarially chosen pixel.[135] Machine learning models are often vulnerable to manipulation or evasion via adversarial machine learning.[136]

Researchers have demonstrated how backdoors can be placed undetectably into classifying (e.g., for categories "spam" and well-visible "not spam" of posts) machine

learning models that are often developed or trained by third parties. Parties can change the classification of any input, including in cases for which a type of data/software transparency is provided, possibly including white-box access.[137][138][139]

Classification of machine learning models can be validated by accuracy estimation techniques like the holdout method, which splits the data in a training and test set (conventionally 2/3 training set and 1/3 test set designation) and evaluates the performance of the training model on the test set. In comparison, the K-fold-cross-validation method randomly partitions the data into K subsets and then K experiments are performed each respectively considering 1 subset for evaluation and the remaining K-1 subsets for training the model. In addition to the holdout and cross-validation methods, bootstrap, which samples n instances with replacement from the dataset, can be used to assess model accuracy.[140]

In addition to overall accuracy, investigators frequently report sensitivity and specificity meaning true positive rate (TPR) and true negative rate (TNR) respectively. Similarly, investigators sometimes report the false positive rate (FPR) as well as the false negative rate (FNR). However, these rates are ratios that fail to reveal their numerators and denominators. Receiver operating characteristic (ROC) along with the accompanying Area Under the ROC Curve (AUC) offer additional tools for classification model assessment. Higher AUC is associated with a better performing model.[141]

The ethics of artificial intelligence covers a broad range of topics within the field that are considered to have particular ethical stakes.[142] This includes algorithmic biases, fairness,[143] automated decision-making, accountability, privacy, and regulation.
It also covers various emerging or potential future challenges such as machine ethics (how to make machines that behave ethically), lethal autonomous weapon systems, arms race dynamics, AI safety and alignment, technological unemployment, AI-enabled misinformation, how to treat certain AI systems if they have a moral status (AI welfare and rights), artificial superintelligence and existential risks.[142]

Different machine learning approaches can suffer from different data biases. A machine learning system trained specifically on current customers may not be able to predict the needs of new customer groups that are not represented in the training data. When trained on human-made data, machine learning is likely to pick up the constitutional and unconscious biases already present in society.[144]

Systems that are trained on datasets collected with biases may exhibit these biases upon use (algorithmic bias), thus digitizing cultural prejudices.[145] For example, in 1988, the UK's Commission for Racial Equality found that St. George's Medical School had been using a computer program trained from data of previous admissions staff and that this program had denied nearly 60 candidates who were found to either be women or have non-European sounding names.[144] Using job hiring data from a firm with racist hiring policies may lead to a machine learning system duplicating the bias by scoring job applicants by similarity to previous successful applicants.[146][147] Another example includes predictive policing company Geolitica's predictive algorithm that

resulted in "disproportionately high levels of over-policing in low-income and minority communities" after being trained with historical crime data.[148]

While responsible collection of data and documentation of algorithmic rules used by a system is considered a critical part of machine learning, some researchers blame lack of participation and representation of minority population in the field of AI for machine learning's vulnerability to biases.[149] In fact, according to research carried out by the Computing Research Association (CRA) in 2021, "female faculty merely make up 16.1%" of all faculty members who focus on AI among several universities around the world.[150] Furthermore, among the group of "new U.S. resident AI PhD graduates," 45% identified as white, 22.4% as Asian, 3.2% as Hispanic, and 2.4% as African American, which further demonstrates a lack of diversity in the field of AI.[150]

Language models learned from data have been shown to contain human-like biases.[151][152] Because human languages contain biases, machines trained on language corpora will necessarily also learn these biases.[153][154] In 2016, Microsoft tested Tay, a chatbot that learned from Twitter, and it quickly picked up racist and sexist language.[155]

In an experiment carried out by ProPublica, an investigative journalism organization, a machine learning algorithm's insight into the recidivism rates among prisoners falsely flagged "black defendants high risk twice as often as white defendants".[148] In 2015, Google Photos once tagged a couple of black people as gorillas, which caused controversy. The gorilla label was subsequently removed, and in 2023, it still cannot

recognize gorillas.[156] Similar issues with recognizing non-white people have been found in many other systems.[157]

Because of such challenges, the effective use of machine learning may take longer to be adopted in other domains.[158] Concern for fairness in machine learning, that is, reducing bias in machine learning and propelling its use for human good, is increasingly expressed by artificial intelligence scientists, including Fei-Fei Li, who said that "[t]here's nothing artificial about AI. It's inspired by people, it's created by people, and—most importantly—it impacts people. It is a powerful tool we are only just beginning to understand, and that is a profound responsibility."[159]

There are concerns among health care professionals that these systems might not be designed in the public's interest but as income-generating machines. This is especially true in the United States where there is a long-standing ethical dilemma of improving health care, but also increasing profits. For example, the algorithms could be designed to provide patients with unnecessary tests or medication in which the algorithm's proprietary owners hold stakes. There is potential for machine learning in health care to provide professionals an additional tool to diagnose, medicate, and plan recovery paths for patients, but this requires these biases to be mitigated.[160]

Since the 2010s, advances in both machine learning algorithms and computer hardware have led to more efficient methods for training deep neural networks (a particular narrow subdomain of machine learning) that contain many layers of nonlinear hidden units.[161] By 2019, graphics processing units (GPUs), often with AI-specific

enhancements, had displaced CPUs as the dominant method of training large-scale commercial cloud AI.[162] OpenAI estimated the hardware compute used in the largest deep learning projects from AlexNet (2012) to AlphaZero (2017), and found a 300,000-fold increase in the amount of compute required, with a doubling-time trendline of 3.4 months.[163][164]

Neuromorphic computing refers to a class of computing systems designed to emulate the structure and functionality of biological neural networks. These systems may be implemented through software-based simulations on conventional hardware or through specialized hardware architectures.[165]

A physical neural network is a specific type of neuromorphic hardware that relies on electrically adjustable materials, such as memristors, to emulate the function of neural synapses. The term "physical neural network" highlights the use of physical hardware for computation, as opposed to software-based implementations. It broadly refers to artificial neural networks that use materials with adjustable resistance to replicate neural synapses.[166][167]

Embedded machine learning is a sub-field of machine learning where models are deployed on embedded systems with limited computing resources, such as wearable computers, edge devices and microcontrollers.[168][169][170] Running models directly on these devices eliminates the need to transfer and store data on cloud servers for further processing, thereby reducing the risk of data breaches, privacy leaks and theft of intellectual property, personal data and business secrets. Embedded machine

learning can be achieved through various techniques, such as hardware acceleration,[171][172] approximate computing,[173] and model optimization.[174][175] Common optimization techniques include pruning, quantization, knowledge distillation, low-rank factorization, network architecture search, and parameter sharing.

Software suites containing a variety of machine learning algorithms include the following: