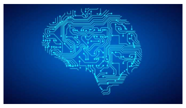
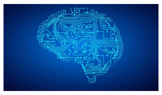
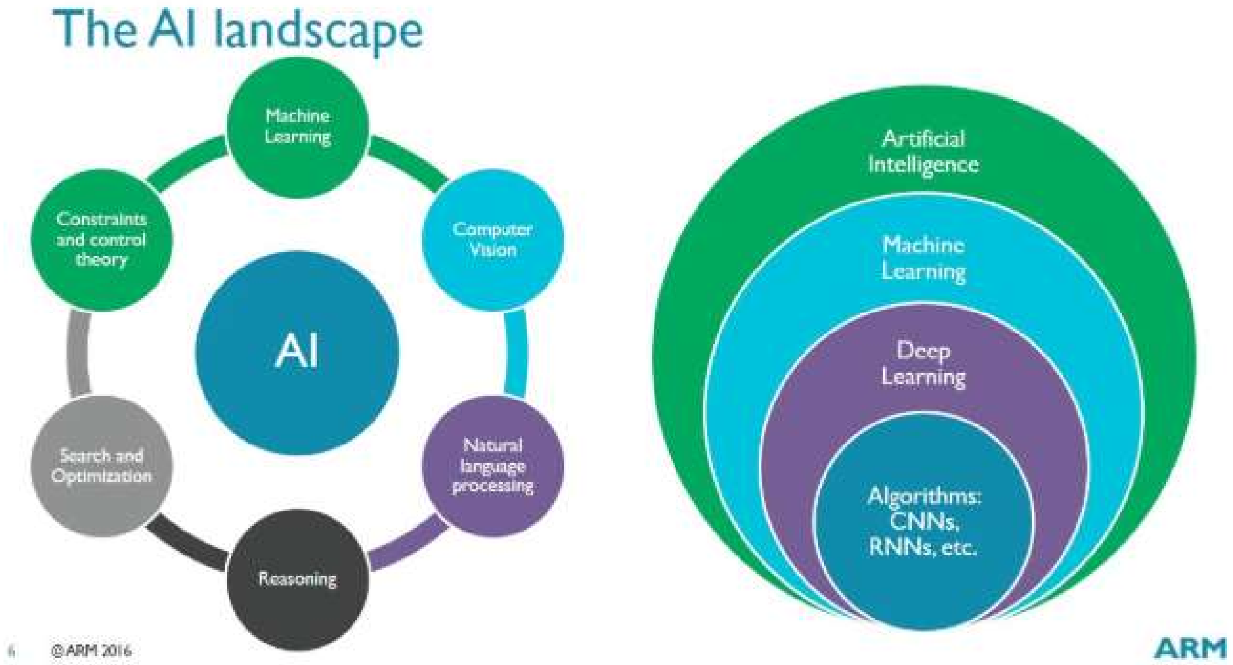


On Machine Learning

Blas G. Moros





The AI landscape

6 © ARM 2016

ARM

Introduction

1. Machine Learning is a subset of Artificial Intelligence and can be summarized as the field of study that gives computers the ability to learn without being explicitly programmed. Recent leaps forward due to massive amounts of readily available data and unparalleled computing power which is becoming cheaper every year has made the prospects for ML and AI in general greater than ever before.

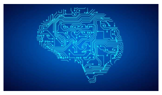
Key Takeaways

1. Artificial Intelligence
    1. Artificial intelligence is the study of agents that perceive the world around them, form plans, and make decisions to achieve their goals. Its foundations include mathematics, logic, philosophy, probability, linguistics, neuroscience, and decision theory. Many fields fall under the umbrella of AI, such as computer vision, robotics, machine learning, and natural language processing.
    2. Artificial narrow intelligence (ANI) – AI which can effectively perform a narrowly defined task.
    3. Artificial general intelligence (AGI) – also known as strong AI. The definition of an AGI is an artificial intelligence that can successfully perform any intellectual task that a human being can, including learning, planning and decision-making under uncertainty, communicating in natural language, making jokes, manipulating people, trading stocks, or… reprogramming itself
2. Machine Learning
    1. Machine learning is a method used to devise complex models and algorithms that lend themselves to prediction. These analytical models allow researchers, data scientists, engineers, and analysts to produce reliable, repeatable decisions and results and uncover "hidden insights" through learning from historical relationships and trends in the data.
    2. Supervised Learning – When you teach the computer how to do something by providing the "right" answers (labelled answers) – teaching by example. We are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output and the goal is to learn a general rule that maps inputs to outputs.
    3. Unsupervised Learning – no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning). Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.
    4. Reinforcement learning – data (in form of rewards and punishments) are given only as feedback to the program's actions in a dynamic environment, such

as driving a vehicle or playing a game against an opponent (exploration and exploitation).

   5. Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. In neural networks, it can be used to minimize the error term by changing each weight in proportion to the derivative of the error with respect to that weight, provided the non-linear activation functions are differentiable.
   6. Bias and Variance – Bias is the amount of error introduced by approximating real-world phenomena with a simplified model. Variance is how much your model's test error changes based on variation in the training data. It reflects the model's sensitivity to the idiosyncrasies of the data set it was trained on. As a model increases in complexity and it becomes more wiggly (flexible), its bias decreases (it does a good job of explaining the training data), but variance increases (it doesn't generalize as well). Ultimately, in order to have a good model, you need one with low bias and low variance.

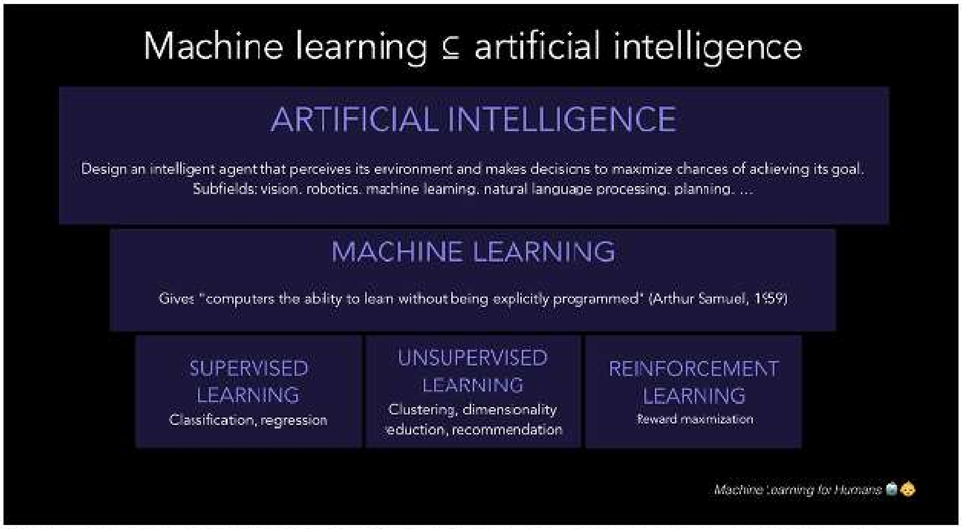
3. Deep Learning
   1. Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.
4. Algorithms
   1. An unambiguous specification of how to solve a class of problems. Algorithms can perform calculation, data processing and automated reasoning tasks. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, eventually producing "output" and terminating at a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate random input.
5. Artificial Neural Network (ANNs)
   1. Computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve their ability) to do tasks by considering examples, generally without task-specific programming.
   2. Convolutional Neural Networks – CNNs were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.
6. Deep Neural Networks (DNNs)
   1. A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship. The network moves through the layers calculating the probability of each output.
7. Recurrent Neural Networks (RNNs)



1. A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

Artificial Intelligence

1. Artificial intelligence is the study of agents that perceive the world around them, form plans, and make decisions to achieve their goals. Its foundations include mathematics, logic, philosophy, probability, linguistics, neuroscience, and decision theory. Many fields fall under the umbrella of AI, such as computer vision, robotics, machine learning, and natural language processing.
2. Artificial narrow intelligence (ANI), which can effectively perform a narrowly defined task.
3. Artificial general intelligence (AGI), also known as strong AI. The definition of an AGI is an artificial intelligence that can successfully perform any intellectual task that a human being can, including learning, planning and decision-making under uncertainty, communicating in natural language, making jokes, manipulating people, trading stocks, or… reprogramming itself. And this last one is a big deal. If we create an AI that can improve itself, it would unlock a cycle of recursive self-improvement that could lead to an intelligence explosion over some unknown time period, ranging from many decades to a single day. You may have heard this point referred to as the singularity. The term is borrowed from the gravitational singularity that occurs at the center of a black hole, an infinitely dense one-dimensional point where the laws of physics as we understand them start to break down.
4. Artificial intelligence will shape our future more powerfully than any other innovation this century. Anyone who does not understand it will soon find themselves feeling left behind, waking up in a world full of technology that feels more and more like magic. The rate of acceleration is already astounding. After a couple of AI winters and periods of false hope over the past four decades, rapid advances in data storage and computer processing power have dramatically changed the game in recent years. Today AI is used to design evidence-based treatment plans for cancer patients, instantly analyze results from medical tests to escalate to the appropriate specialist immediately, and conduct scientific research for drug discovery.



Machine learning ⊆ artificial intelligence

ARTIFICIAL INTELLIGENCE

Design an intelligent agent that perceives its environment and makes decisions to maximize chances of achieving its goal.
Subfields: vision. robotics. machine learning. natural language processing. planning. …

MACHINE LEARNING

Gives "computers the ability to lean without being explicitly programmed" (Arthur Samuel, 1959)

SUPERVISED LEARNING
Classification, regression

UNSUPERVISED LEARNING
Clustering, dimensionality eduction, recommendation

REINFORCEMENT LEARNING
Reward maximization

Machine Learning for Humans 🤖👶

Machine learning is one of many subfields of artificial intelligence, concerning the ways that computers learn from experience to improve their ability to think, plan, decide, and act.

5. “The semantic tree: artificial intelligence and machine learning. One bit of advice: it is important to view knowledge as sort of a semantic tree — make sure you understand the fundamental principles, i.e. the trunk and big branches, before you get into the leaves/details or there is nothing for them to hang on to.” — Elon Musk
6. Tesler’s Theorem – AI is whatever hasn’t been done yet, AI is a “moving target”
7. For most of its history, AI research has been divided into subfields that often fail to communicate with each other
8. The traditional problems (or goals) of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and the ability to move and manipulate objects. General intelligence is among the field's long-term goals. Approaches include statistical methods, computational intelligence, and traditional symbolic AI. Many tools are used in AI, including versions of search and mathematical optimization, artificial neural networks, and methods based on statistics, probability and economics. The AI field draws upon computer science, information engineering, mathematics, psychology, linguistics, philosophy, and many others.
9. In the twenty-first century, AI techniques have experienced a resurgence following concurrent advances in computer power, large amounts of data, and theoretical understanding; and AI techniques have become an essential part of the technology industry, helping to solve many challenging problems in computer science, software engineering and operations research.
10. AI often revolves around the use of algorithms. An algorithm is a set of unambiguous instructions that a mechanical computer can execute. A complex algorithm is often built on top of other, simpler, algorithms.
11. AI, like electricity or the steam engine, is a general purpose technology. There is no consensus on how to characterize which tasks AI tends to excel at. While projects such as AlphaZero have succeeded in generating their own knowledge from scratch, many other machine learning projects require large training datasets. Researcher Andrew Ng has suggested, as a "highly imperfect rule of thumb", that "almost anything a typical human can do with less than one second of mental thought, we can probably now or in the near future automate using AI." Moravec's paradox suggests that AI lags humans at many tasks that the human brain has specifically evolved to perform well.
12. The number of atomic facts that the average person knows is very large. Research projects that attempt to build a complete knowledge base of commonsense knowledge (e.g., Cyc) require enormous amounts of laborious ontological engineering—they must be built, by hand, one complicated concept at a time
13. Much of what people know is not represented as "facts" or "statements" that they could express verbally. For example, a chess master will avoid a particular chess position because it "feels too exposed" or an art critic can take one look at a statue and realize that it is a fake. These are non-conscious and sub-symbolic intuitions or tendencies in the human brain. Knowledge like this informs, supports and provides a context for symbolic, conscious knowledge. As with the related problem of sub-symbolic reasoning, it is hoped that situated AI, computational intelligence, or statistical AI will provide ways to represent this kind of knowledge
14. Philosophy & Ethics
    a. There are three philosophical questions related to AI:

i. Is artificial general intelligence possible? Can a machine solve any problem that a human being can solve using intelligence? Or are there hard limits to what a machine can accomplish?

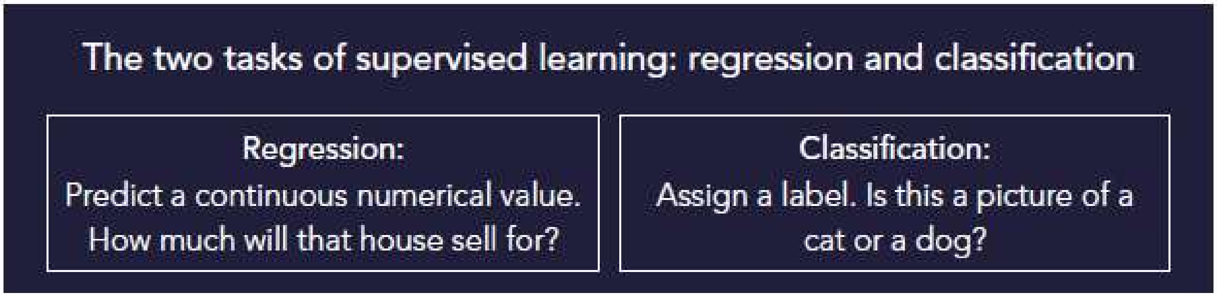
ii. Are intelligent machines dangerous? How can we ensure that machines behave ethically and that they are used ethically?

iii. Can a machine have a mind, consciousness and mental states in exactly the same sense that human beings do? Can a machine be sentient, and thus deserve certain rights? Can a machine intentionally cause harm?

Machine Learning

1. Machine learning is a method used to devise complex models and algorithms that lend themselves to prediction. These analytical models allow researchers, data scientists, engineers, and analysts to produce reliable, repeatable decisions and results and uncover “hidden insights” through learning from historical relationships and trends in the data.
2. Machine learning is a subfield of artificial intelligence. Its goal is to enable computers to learn on their own. A machine’s learning algorithm enables it to identify patterns in observed data, build models that explain the world, and predict things without having explicit pre-programmed rules and models.
3. 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.
4. 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 underfit 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.
5. Machine learning and pattern recognition "can be viewed as two facets of the same field
6. Well-Posed Learning Problem: a computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. – Tom Mitchel
7. “I have heard it said, in fact I believe it is quite a current thought, that we have taken skill out of work. We have not. We have put in skill. We have put a higher skill into planning, management, and tool building, and the results of that skill are enjoyed by the man who is not skilled.” – Henry Ford
8. One of the most powerful uses is micro-segmentation based on behavioral characteristics of individuals. This is changing the fundamentals of competition in many sectors, including education, travel and leisure, media, retail, and advertising.
9. Mechinzation, cognitive and learning capabilities – improving over time as they are trained by their human coworkers on the shop floor, are excellent ways to improve manufacturing by combining AI with human experience and knowledge
10. The goal of automation is to empower and not to displace employees
11. Machine Learning for Humans is a great way to get to 80/20 in this space

12. Intelligent agents: any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals
13. Learners also work on the basis of "Occam's razor": The simplest theory that explains the data is the likeliest. Therefore, to be successful, a learner must be designed such that it prefers simpler theories to complex theories, except in cases where the complex theory is proven substantially better.



The two tasks of supervised learning: regression and classification

Regression:	Classification:
Predict a continuous numerical value. How much will that house sell for?	Assign a label. Is this a picture of a cat or a dog?

14. Supervised learning:

    b. In supervised learning problems, we start with a data set containing training examples with associated correct labels – teaching by example. We are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output and the goal is to learn a general rule that maps inputs to outputs.

    c. Supervised learning problems are categorized into "regression" and "classification" problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

    d. The two tasks of supervised learning: regression and classification. Linear regression, loss functions, and gradient descent.

    e. Supervised machine learning solves this problem by getting the computer to do the work for you. By identifying patterns in the data, the machine is able to form heuristics. The primary difference between this and human learning is that machine learning runs on computer hardware and is best understood through the lens of computer science and statistics, whereas human pattern-matching happens in a biological brain (while accomplishing the same goals). The goal of supervised learning is to predict Y as accurately as possible when given new examples where X is known and Y is unknown. In what follows we'll explore several of the most common approaches to doing so.

    f. Support vector machines (SVMs) 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 or the other.

15. Unsupervised learning

    a. No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning). Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.

b. With unsupervised learning there is no feedback based on the prediction results.

c. Uses – Organize computing clusters, social network analysis, market segmentation, astronomical data analysis

16. Semi-supervised learning

    a. The computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing. Some labelled data and a large amount of unlabeled data to train the system

    b. Huge amounts of computing power and data may be more important for successfully training learning systems than access to large, labelled datasets

17. Active learning

    a. The computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.

18. Reinforcement learning

    a. There's no answer key, but your reinforcement learning agent still has to decide how to act to perform its task. In the absence of existing training data, the agent learns from experience. It collects the training examples ("this action was good, that action was bad") through trial-and-error as it attempts its task, with the goal of maximizing long-term reward.

    b. Experience replay, which learns by randomizing over a longer sequence of previous observations and corresponding reward to avoid overfitting to recent experiences. This idea is inspired by biological brains: rats traversing mazes, for example, "replay" patterns of neural activity during sleep in order to optimize future behavior in the maze.

    c. Recurrent neural networks (RNNs) augmenting DQNs. When an agent can only see its immediate surroundings (e.g. robot-mouse only seeing a certain segment of the maze vs. a birds-eye view of the whole maze), the agent needs to remember the bigger picture so it remembers where things are. This is similar to how human's babies develop object permanence to know things exist even if they leave the baby's visual field. RNNs are "recurrent", i.e. they allow information to persist on a longer-term basis.

19. Overfitting

    a. A common problem in machine learning is overfitting: learning a function that perfectly explains the training data that the model learned from, but doesn't generalize well to unseen test data. Overfitting happens when a model overlearns from the training data to the point that it starts picking up idiosyncrasies that aren't representative of patterns in the real world. This becomes especially problematic as you make your model increasingly complex.

    b. Underfitting is a related issue where your model is not complex enough to capture the underlying trend in the data.

c. Machine learning models need to generalize well to new examples that the model has not seen in practice. We'll introduce regularization, which helps prevent models from overfitting the training data.

d. Options to address overfitting:

    i. Use more training data. The more you have, the harder it is to overfit the data by learning too much from any single training example.

    ii. Reduce the number of features

        1. Manually select which features to keep
        2. Model selection algorithm – automatically selects which features to keep and discard

    iii. Regularization

        1. Add in a penalty in the loss function for building a model that assigns too much explanatory power to any one feature or allows too many features to be taken into account.
        2. Keep all the features, but reduce magnitude/values of parameters
        3. Works well when we have a lot of features, each of which contributes a bit to predicting *y*

20. Another categorization of machine learning tasks arises when one considers the desired *output* of a machine-learned system

    a. In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".

    b. In regression, also a supervised problem, the outputs are continuous rather than discrete.

    c. In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

    d. Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space.

    e. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics

21. Decision tree learning

    a. Decision tree learning uses a decision tree as a predictive model, which maps observations about an item to conclusions about the item's target value.

22. Association rule learning

    a. Association rule learning is a method for discovering interesting relations between variables in large databases.

23. Clustering

    a. 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 some predesignated criterion or 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* (similarity between members of the same cluster) and *separation* between different clusters. Other methods are based on *estimated density* and *graph connectivity*. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis.

24. Similarity and metric learning

    a. In this problem, the learning machine is given pairs of examples that are considered similar and pairs of less similar objects. It then needs to learn a similarity function (or a distance metric function) that can predict if new objects are similar. It is sometimes used in recommendation systems.

25. Genetic algorithms

    a. A search heuristic that mimics the process of natural selection, and uses methods such as mutation and crossover to generate new genotype in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms found some uses in the 1980s and 1990s.Conversely, machine learning techniques have been used to improve the performance of genetic and evolutionary algorithms.

26. Rule-based machine learning

    a. 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 learner 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 learners that commonly identify a singular model that can be universally applied to any instance in order to make a prediction. Rule-based machine learning approaches include learning classifier systems, association rule learning, and artificial immune systems.

27. Learning classifier systems

    a. Learning classifier systems (LCS) are a family of rule-based machine learning algorithms that combine a discovery component (e.g. 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

28. Fathers of Machine Learning
    a. Tom Mitchell
        i. Prolific author on various topics in computer science, including machine learning, AI, robotics, and cognitive neuroscience.
    b. Arthur Samuel
        i. Coined the term machine learning in 1959
        ii. Known for writing articles that made complex subjects easy to understand
        iii. Most known for his groundbreaking work in computer checkers in 1959. Believed this would help with general problems because there was depth to strategy
    c. Herbert Simon

    d. John McCarthy
    e. Marvin Minsky

29. Gradient Descent
    a. Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. In neural networks, it can be used to minimize the error term by changing each weight in proportion to the derivative of the error with respect to that weight, provided the non-linear activation functions are differentiable.

30. Limitations
    a. Although machine learning has been transformative in some fields, effective machine learning is difficult because finding patterns is hard and often not enough training data are available; as a result, many machine-learning programs often fail to deliver the expected value. 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.
    b. Machine learning approaches in particular can suffer from different data biases. A machine learning system trained on your current customers only may not be able to predict the needs of new customer groups that are not represented in the training data. When trained on man-made data, machine learning is likely to pick up the same constitutional and unconscious biases already present in society Language models learned from data have been shown to contain human-like biases. Machine learning systems used for criminal risk assessment have been found to be biased against black people.
    c. Responsible collection of data and documentation of algorithmic rules used by a system thus is a critical part of machine learning.
    d. Bias is the amount of error introduced by approximating real-world phenomena with a simplified model.
    e. Variance is how much your model's test error changes based on variation in the training data. It reflects the model's sensitivity to the idiosyncrasies of the data set it was trained on. As a model increases in complexity and it becomes more wiggly (flexible), its bias decreases (it does a good job of explaining the training data), but variance increases (it doesn't generalize as well).
    f. Ultimately, in order to have a good model, you need one with low bias and low variance.
    g. Remember that the only thing we care about is how the model performs on test data. You want to predict which emails will be marked as spam before they’re marked, not just build a model that is 100% accurate at reclassifying the emails it used to build itself in the first place. Hindsight is 20/20 — the real question is whether the lessons learned will help in the future.

31. Examples of ML
    a. Database mining – large datasets from growth of automation/web (web click data, medical records)
    b. Applications which can’t be programmed by hand (Handwriting recognition, autonomous helicopter, natural language processing, computer vision)
    c. Self-customizing programs (Amazon, Netflix product recommendations)

d. Understanding human learning (brain, real AI)
e. Predictive maintenance for machines

Deep Learning

1. Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.
2. Deep learning is any artificial neural network that can learn a long chain of causal links
3. Deep learning is a class of machine learning algorithms that:
    1. Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
    2. Learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners.
    3. Learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.
4. Artificial neural networks have actually been around for a long time. Their application has been historically referred to as cybernetics (1940s-1960s), connectionism (1980s-1990s), and then came into vogue as deep learning circa 2006 when neural networks started getting, well, “deeper”. There are generally “four separate factors that hold back AI:
    1. Compute (the obvious one: Moore’s Law, GPUs, ASICs),
    2. Data (in a nice form, not just out there somewhere on the internet — e.g. ImageNet),
    3. Algorithms (research and ideas, e.g. backprop, CNN, LSTM), and
    4. Infrastructure (software under you — Linux, TCP/IP, Git, ROS, PR2, AWS, AMT, TensorFlow, etc.)
5. In the past decade or so, the full potential of deep learning is finally being unlocked by advances in (1) and (2), which in turn has led to further breakthroughs in (3) and (4) — and so the cycle continues, with exponentially more humans rallying to the frontlines of deep learning research along the way (just think about what you’re doing right now!)
6. Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts
7. In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation. In an image recognition application, the raw input may be a matrix of pixels; the first representational layer may abstract the pixels and encode edges; the second layer may compose and encode arrangements of edges; the third layer may encode a nose and eyes; and the fourth layer may recognize that the image contains a face. Importantly, a deep learning process can learn which features to optimally place in which level *on its own*. (Of course, this does not completely obviate the need for hand-

tuning; for example, varying numbers of layers and layer sizes can provide different degrees of abstraction.). The "deep" in "deep learning" refers to the number of layers through which the data is transformed. More precisely, deep learning systems have a substantial *credit assignment path* (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connections between input and output. For a feedforward neural network, the depth of the CAPs is that of the network and is the number of hidden layers plus one (as the output layer is also parameterized). For recurrent neural networks, in which a signal may propagate through a layer more than once, the CAP depth is potentially unlimited. No universally agreed upon threshold of depth divides shallow learning from deep learning, but most researchers agree that deep learning involves CAP depth > 2. CAP of depth 2 has been shown to be a universal approximator in the sense that it can emulate any function. Beyond that more layers do not add to the function approximator ability of the network. Deep models (CAP > 2) are able to extract better features than shallow models and hence, extra layers help in learning features.

8. Applications
    1. Computer vision – images and objects
    2. Automatic speech recognition (ASR)
    3. Visual art processing
    4. Natural language processing
    5. Drug discovery and toxicology
    6. Customer relationship management
    7. Recommendation systems
    8. Image restoration
    9. Financial fraud detection
    10. Military (ability to learn new tasks through observation – UT's TAMER. "Using Deep TAMER, a robot learned a task with a human trainer, watching video streams or observing a human perform a task in-person. The robot later practiced the task with the help of some coaching from the trainer, who provided feedback such as "good job" and "bad job"")
9. Criticism
    1. A main criticism concerns the lack of theory surrounding some methods. Learning in the most common deep architectures is implemented using well-understood gradient descent. However, the theory surrounding other algorithms, such as contrastive divergence is less clear. (e.g., does it converge? If so, how fast? What is it approximating?) Deep learning methods are often looked at as a black box, with most confirmations done empirically, rather than theoretically

    2. As deep learning moves from the lab into the world, research and experience shows that artificial neural networks are vulnerable to hacks and deception. By identifying patterns that these systems use to function, attackers can modify inputs to ANNs in such a way that the ANN finds a match that human observers would not recognize. For example, an attacker can make subtle changes to an image such that the ANN finds a match even though the image looks to a human nothing like the search target. Such a manipulation is termed an "adversarial attack."

Algorithms

1. An unambiguous specification of how to solve a class of problems. Algorithms can perform calculation, data processing and automated reasoning tasks. Starting from an initial state and initial input (perhaps empty), the instructions describe a computation that, when executed, proceeds through a finite number of well-defined successive states, eventually producing "output" and terminating at a final ending state. The transition from one state to the next is not necessarily deterministic; some algorithms, known as randomized algorithms, incorporate random input.
2. Because an algorithm is a precise list of precise steps, the order of computation is always crucial to the functioning of the algorithm. Instructions are usually assumed to be listed explicitly, and are described as starting "from the top" and going "down to the bottom", an idea that is described more formally by *flow of control*.
3. Most algorithms are intended to be implemented as computer programs. However, algorithms are also implemented by other means, such as in a biological neural network (for example, the human brain implementing arithmetic or an insect looking for food), in an electrical circuit, or in a mechanical device.
4. An optimal algorithm, even running in old hardware, would produce faster results than a non-optimal (higher time complexity) algorithm for the same purpose, running in more efficient hardware; that is why algorithms, like computer hardware, are considered technology.
5. "Elegant" (compact) programs, "good" (fast) programs : The notion of "simplicity and elegance" appears informally in Knuth and precisely in Chaitin:
    1. Knuth: "We want good algorithms in some loosely defined aesthetic sense. One criterion . . . is the length of time taken to perform the algorithm. Other criteria are adaptability of the algorithm to computers, its simplicity and elegance, etc."
    2. Chaitin: " a program is 'elegant,' by which I mean that it's the smallest possible program for producing the output that it does
6. Typical steps in the development of algorithms:
    1. Problem definition
    2. Development of a model
    3. Specification of the algorithm
    4. Designing an algorithm
    5. Checking the correctness of the algorithm
    6. Analysis of algorithm
    7. Implementation of algorithm
    8. Program testing
    9. Documentation preparation
7. Recursion
    1. A recursive algorithm is one that invokes (makes reference to) itself repeatedly until a certain condition (also known as termination condition) matches, which is a method common to functional programming. Iterative algorithms use repetitive constructs like loops and sometimes additional data structures like stacks to solve the given problems. Some problems are naturally suited for one implementation or the other. For example, towers of Hanoi is well understood using recursive

implementation. Every recursive version has an equivalent (but possibly more or less complex) iterative version, and vice versa.

8. Logical
    1. An algorithm may be viewed as controlled logical deduction. This notion may be expressed as: Algorithm = logic + control. The logic component expresses the axioms that may be used in the computation and the control component determines the way in which deduction is applied to the axioms. This is the basis for the logic programming paradigm. In pure logic programming languages, the control component is fixed and algorithms are specified by supplying only the logic component. The appeal of this approach is the elegant semantics: a change in the axioms produces a well-defined change in the algorithm.
9. Serial, parallel or distributed
    1. Algorithms are usually discussed with the assumption that computers execute one instruction of an algorithm at a time. Those computers are sometimes called serial computers. An algorithm designed for such an environment is called a serial algorithm, as opposed to parallel algorithms or distributed algorithms. Parallel algorithms take advantage of computer architectures where several processors can work on a problem at the same time, whereas distributed algorithms utilize multiple machines connected with a computer network. Parallel or distributed algorithms divide the problem into more symmetrical or asymmetrical subproblems and collect the results back together. The resource consumption in such algorithms is not only processor cycles on each processor but also the communication overhead between the processors. Some sorting algorithms can be parallelized efficiently, but their communication overhead is expensive. Iterative algorithms are generally parallelizable. Some problems have no parallel algorithms and are called inherently serial problems.
10. Deterministic or non-deterministic
    1. Deterministic algorithms solve the problem with exact decision at every step of the algorithm whereas non-deterministic algorithms solve problems via guessing although typical guesses are made more accurate through the use of heuristics.
11. Exact or approximate
    1. While many algorithms reach an exact solution, approximation algorithms seek an approximation that is closer to the true solution. The approximation can be reached by either using a deterministic or a random strategy. Such algorithms have practical value for many hard problems. One of the examples of an approximate algorithm is the Knapsack problem. The Knapsack problem is a problem where there is a set of given items. The goal of the problem is to pack the knapsack to get the maximum total value. Each item has some weight and some value. Total weight that we can carry is no more than some fixed number X. So, we must consider weights of items as well as their value.
12. Quantum algorithm
    1. They run on a realistic model of quantum computation. The term is usually used for those algorithms which seem inherently quantum, or use some essential feature of Quantum computing such as quantum superposition or quantum entanglement.
13. Brute-force or exhaustive search

    1. This is the naive method of trying every possible solution to see which is best.

14. Divide and conquer
    1. A divide and conquer algorithm repeatedly reduces an instance of a problem to one or more smaller instances of the same problem (usually recursively) until the instances are small enough to solve easily. One such example of divide and conquer is merge sorting. Sorting can be done on each segment of data after dividing data into segments and sorting of entire data can be obtained in the conquer phase by merging the segments. A simpler variant of divide and conquer is called a decrease and conquer algorithm, that solves an identical subproblem and uses the solution of this subproblem to solve the bigger problem. Divide and conquer divides the problem into multiple subproblems and so the conquer stage is more complex than decrease and conquer algorithms. An example of the decrease and conquer algorithm is the binary search algorithm.
15. Search and enumeration
    1. Many problems (such as playing chess) can be modeled as problems on graphs. A graph exploration algorithm specifies rules for moving around a graph and is useful for such problems. This category also includes search algorithms, branch and bound enumeration and backtracking.
16. Randomized algorithm
    1. Such algorithms make some choices randomly (or pseudo-randomly). They can be very useful in finding approximate solutions for problems where finding exact solutions can be impractical (see heuristic method below). For some of these problems, it is known that the fastest approximations must involve some randomness. Whether randomized algorithms with polynomial time complexity can be the fastest algorithms for some problems is an open question known as the P versus NP problem. There are two large classes of such algorithms:
        1. Monte Carlo algorithms return a correct answer with high-probability. E.g. RP is the subclass of these that run in polynomial time.
        2. Las Vegas algorithms always return the correct answer, but their running time is only probabilistically bound, e.g. ZPP.
17. Reduction of complexity
    1. This technique involves solving a difficult problem by transforming it into a better-known problem for which we have (hopefully) asymptotically optimal algorithms. The goal is to find a reducing algorithm whose complexity is not dominated by the resulting reduced algorithms. For example, one selection algorithm for finding the median in an unsorted list involves first sorting the list (the expensive portion) and then pulling out the middle element in the sorted list (the cheap portion). This technique is also known as transform and conquer.
18. Back tracking
    1. In this approach, multiple solutions are built incrementally and abandoned when it is determined that they cannot lead to a valid full solution.

Artificial Neural Network

1. Artificial neural networks (ANNs) or connectionist systems are computing systems inspired by the biological neural networks that constitute animal brains. Such systems learn (progressively improve their ability) to do tasks by considering examples, generally without task-specific programming. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the analytic results to identify cats in other images. They have found most use in applications difficult to express with a traditional computer algorithm using rule-based programming.
2. An ANN is based on a collection of connected units called artificial neurons, (analogous to biological neurons in a biological brain). Each connection (synapse) between neurons can transmit a signal to another neuron. The receiving (postsynaptic) neuron can process the signal(s) and then signal downstream neurons connected to it. Neurons may have state, generally represented by real numbers, typically between 0 and 1. Neurons and synapses may also have a weight that varies as learning proceeds, which can increase or decrease the strength of the signal that it sends downstream.
3. Typically, neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first (input), to the last (output) layer, possibly after traversing the layers multiple times.
4. The original goal of the neural network approach was to solve problems in the same way that a human brain would. Over time, attention focused on matching specific mental abilities, leading to deviations from biology such as backpropagation, or passing information in the reverse direction and adjusting the network to reflect that information.
5. 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.
6. A great way to learn complex, non-linear hypotheses
7. There is a hypothesis that the brain has "one learning algorithm"
8. There is researching being done to rewire the brain for those that have damage – i.e., seeing with your tongue through an electric medium placed on the tongue, human echolocation, a haptic belt for directional sense, and even implanting third eyes
9. Convolutional Deep Neural Networks
    a. Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.
    b. A class of deep, feed-forward artificial neural networks, most commonly applied to analyzing visual imagery – image recognition, video analysis, NLP, health risk assessment and biomarkers of aging discovery, drug discovery, checkers, Go

    c. CNNs use relatively little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This independence from prior knowledge and human effort in feature design is a major advantage.

    d. Convolutional layers apply a convolution operation to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli

    e. For many applications, little training data is available. Convolutional neural networks usually require a large amount of training data in order to avoid overfitting. A common technique is to train the network on a larger data set from a related domain. Once the network parameters have converged an additional training step is performed using the in-domain data to fine-tune the network weights. This allows convolutional networks to be successfully applied to problems with small training sets.

10. Human interpretable explanations

    a. End-to-end training and prediction are common practice in computer vision. However, human interpretable explanations are required for critical systems such as a self-driving cars. "Black-box models will not suffice". With recent advances in visual salience, spatial and temporal attention, the most critical spatial regions/temporal instants could be visualized to justify the CNN predictions

Deep Neural Network

1. A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship. The network moves through the layers calculating the probability of each output. For example, a DNN that is trained to recognize dog breeds will go over the given image and calculate the probability that the dog in the image is a certain breed. The user can review the results and select which probabilities the network should display (above a certain threshold, etc.) and return the proposed label. Each mathematical manipulation as such is considered a layer, and complex DNN have many layers, hence the name "deep" networks. The goal is that eventually, the network will be trained to decompose an image into features, identify trends that exist across all samples and classify new images by their similarities without requiring human input.
2. DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives. The extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network.
3. Deep neural networks approach the image classification problem using layers of abstraction. To repeat what we explained earlier in this section: the input layer will take raw pixel brightnesses of an image. The final layer will be an output vector of class probabilities (i.e. the probability of the image being a “cat”, “car”, “horse”, etc.) But instead of learning a simple linear model relating input to output, we’ll instead construct intermediate hidden layers of the network will learn increasingly abstract features, which enables us to not lose all the nuance in the complex data.

Recurrent Neural Networks

1. A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.
2. Data can flow in any direction, are used for applications such as language modeling. Long short-term memory is particularly effective for this use.
3. Long short-term memory (LSTM) is a deep learning system that avoids the vanishing gradient problem. LSTM is normally augmented by recurrent gates called "forget" gates. LSTM prevents back propagated errors from vanishing or exploding. Instead, errors can flow backwards through unlimited numbers of virtual layers unfolded in space. That is, LSTM can learn tasks that require memories of events that happened thousands or even millions of discrete time steps earlier. Problem-specific LSTM-like topologies can be evolved. LSTM works even given long delays between significant events and can handle signals that mix low and high frequency components.
4. LSTM broke records for improved machine translation, Language Modeling and Multilingual Language Processing. LSTM combined with convolutional neural networks (CNNs) improved automatic image captioning.
5. Both finite impulse and infinite impulse recurrent networks can have additional stored state, and the storage can be under direct control by the neural network. The storage can also be replaced by another network or graph, if that incorporates time delays or has feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of long short-term memory networks (LSTMs) and gated recurrent units.
6. Training the weights in a neural network can be modeled as a non-linear global optimization problem. A target function can be formed to evaluate the fitness or error of a particular weight vector as follows: First, the weights in the network are set according to the weight vector. Next, the network is evaluated against the training sequence. Typically, the sum-squared-difference between the predictions and the target values specified in the training sequence is used to represent the error of the current weight vector. Arbitrary global optimization techniques may then be used to minimize this target function.
7. The most common global optimization method for training RNNs is genetic algorithms, especially in unstructured networks.
8. Applications – machine translation, robot control, time series prediction, speech recognition, rhythm learning, grammar learning, handwriting recognition, business process management,

Summary

1. Getting to know the basics in these topics was not only a lot of fun but I think really important as they will likely become pervasive and permeate every area of life. The language and terms used to describe these different topics also gave me some vocabulary and a new way to think about learning, AI, habits, training, and serve as great metaphors to think about other areas not directly related to AI/ML.

Resources

1. Wikipedia - Artificial Intelligence, Machine Learning, Deep Learning, Tom Mitchell, Arthur Samuel
2. Andrew Ng lectures (Coursera)
3. CS231n: Convolutional Neural Networks for Visual Recognition (YouTube Lecture Series)
4. Machine Learning for Humans
5. What's Now and Next in Analytics, AI, and Automation
6. Santa Fe Institute: Fundamentals of ML
7. A Visual Intro to ML
8. The Master Algorithm - Pedro Domingos
9. What is ML? Everything You Need to Know
10. Amazon Web Services ML University Classes for Business Decision Makers
11. Deeplearning.ai
12. Greg Brockman on Deep Learning

Teacher’s Reference Guides

My “teacher’s reference guides” are deep dives into a subject, theme, person, or idea which are then distilled into (hopefully) clear, concise, and helpful resources. My goal is to effectively share what I think are the most actionable, impactful, and noteworthy takeaways of the topic at hand.

There isn’t much rhyme or reason to how I choose these teacher’s reference guides. Sometimes I want to dive deep on a specific concept such as complexity and spend months reading about that and sometimes I simply stumble across a person or topic randomly which captures my attention – trying to balance serendipity and chaos with routine and order.

You can find a full sampling of my teacher’s reference guides (blas.com/teachers-reference-guides/) on my blog, blas.com.

If any of this is of interest, you can subscribe to the monthly newsletter (http://blas.com/newsletter/) and you can always reach out to me directly at rabbithole@blas.us

Amor Fati.

Blas