Digital libraries: the systems analysis perspective machine erudition

Robert Fox
University of Notre Dame, Notre Dame, Indiana, USA

Received 3 February 2016
Accepted 3 February 2016

Abstract

Purpose – The purpose of this paper is to explore the concept of machine learning. Current trends in the field are explored, along with the potential impact on information science. Machine learning is both an old and new field. It has been theoretically explored since the 1940s, but advances in technology, statistics and mathematics have recently created conditions, wherein it can be put into practice.

Design/methodology/approach – This is a conceptual column exploring the notion of machine learning and the applications for information science.

Findings – Some of the objections to machine intelligence are common philosophical problems dealing with the nature of thinking, self-awareness, understanding and other human traits that allow us to relate to people, develop intuitions and have situational awareness.

Originality/value – While machine learning is being taken advantage of in the commercial sector, it has not been effectively exploited in the academic sphere. Libraries have traditionally focused on structured analysis and strictly controlled vocabularies to enable information discovery. Machine learning opens up possibilities for unstructured data to be analyzed intelligently. Over 80 per cent of regularly consumed information on the Internet is unstructured, so this field has huge implications for discovery from a library perspective.

Keywords Machine learning, Neural networks, Discovery, Data classification, Machine logic, Reinforcement learning

Paper type Conceptual paper

Following the world wars in the early part of the twentieth century, and the subsequent rise of both physics and mathematics as fields of inquiry, there was a great deal of speculation about the future. By 1950, the world had been transformed by automation, and machines were a part of every day life. These changes had happened so rapidly that it seemed as though automata that would be performing all of our menial tasks in the near future. In that era, a man named Alan Turing incorporated his intellectual ponderings about automation and machine learning into a paper he published in *Mind* called "Computer Machinery and Intelligence" (Turing, 1950). When this paper was published, many of the controversies that we still struggle with today were raised and debated. Some of the objections to machine intelligence are common philosophical problems dealing with the nature of thinking, self-awareness, understanding and other human traits that allow us to relate to people, develop intuitions and have situational awareness.

Turing understood the importance of these problems, and in his paper, he did not avoid them, but he did side step them to a certain degree by shifting the nature of the question. For example, a prevalent religious objection to machine intelligence is that the

Digital Library Perspectives
Vol. 32 No. 2, 2016
pp. 62-67
© Emerald Group Publishing Limited
2059-5816
DOI 10.1108/DLP-02-2016-0006

mind has an immaterial nature and that it is related very closely to the human soul. There can be no true understanding or comprehension without that immaterial aspect. There are also objections that stem from the study of human consciousness, which to this day is not well understood (Stalker, 1978). Is it possible for a machine to develop inferences or have original thoughts that are not derivative of human intelligence? Can machines ever become self-aware? Turing's engagement in this area shifted the focus from the more philosophical questions and instead asks the question: Is it possible to devise a learning machine that is under normal circumstances indistinguishable from a human being? This forgoes the questions about subjectivity and self-awareness and instead gives space to the notion that while machines are just machines, they might possibly be able to mimic human understanding to such a degree that it is difficult, if not impossible, for a human to tell the difference.

Asking the question in this manner eventually became canonized in computer science lore as the "Turing test"[1]. A simple scenario is the basis for this test. Let us say three people are playing a game that involves interrogation of two of the people by an interrogator. The "judge" or interrogator cannot see the two people he or she is asking questions of. If we were to swap out one of the humans with a computer, then the computer would effectively take the place of the second human in the test. However, the judge is not aware of which of the two people was switched out. The judges objective is to determine which is the person and which the computer (or "machine"). If the judge cannot make that determination and cannot tell the difference, then the computer "wins" the game. In his article, though, Turing makes it clear via his replies to the standard objections that the goal would be not simply to deceive a human regarding the identity of the computer. Rather, turning the focus to the machine he asks: Is the machine capable of *imitating* a human? If a machine is capable of doing so, then the questions of consciousness and self-awareness become moot because it is just as impossible for a human to know the internal cogitations of a machine as it is another human. If the machine is successful enough, then how do we know whether or not the machine is exhibiting "consciousness"?

Academic androids

The impact of machine learning on twenty-first-century life is becoming more evident every day. This is happening in subtle and not so subtle ways. While we might not think of artificial intelligence when we are purchasing things in the e-commerce context, rest assured that it is present. We see it manifested whenever we purchase an item on a site such as Amazon or when we perform a search in Google[2]. When items are mysteriously recommended to us based on our past purchasing decisions or when results in a search engine seem to match our exact need, machine learning is at work. The Internet powerhouse companies have invested a great deal of time and money in this field because it translates directly into profit. This field is not monolithic, however. The methods and the applications are quite numerous. A mention of just a few areas where machine learning A.I. is used will suffice to demonstrate how ubiquitous it is becoming in modern culture.

If you have ever seen a demonstration of an "autonomous" or self-driving car, then you have seen machine learning at work[3]. These vehicles house more than $150,000 of equipment in them that includes infrared, ultraviolet and laser technology to dynamically map the environment, as well as inch-precision maps of areas where the car may travel. Machine learning comes into play in this scenario when engineers are "training" a car how to drive under various conditions. There are so many factors that humans deal with when driving that are not apparent to the conscious mind because they are taken for granted such as: contextual acceleration, braking speed, steering precision, obstacles, other drivers and road conditions just to name a few. Algorithms are created that allow the vehicle's computer system to "train" using complex mathematical formulas and an enormous amount of data gathered during training sessions. For certain driving conditions, engineers and data scientists will drive for thousands of miles to train the vehicle to handle just that one condition (e.g. muddy roads).

There are two aspects of driverless vehicle training, which apply to any machine-learning situation. The first is data. To train an algorithm to respond like a human, large quantities of data are required. The super computer Deep Blue that was pitted against then reigning chess champion Gary Kasparov was able to generate up to 200m positions per second which were then fed into an evaluative function to determine the best move at a given moment[4]. Deep Blue won a second match against Kasparov in 1997 after IBM engineers had significantly enhanced Deep Blue's database of situational chess moves. Quantities of discrete data points of this size are hard to imagine, but that is what it takes to provide enough information for a computer to perform judgments that are similar to a human. The other aspect of driverless vehicle training that is true in most machine learning contexts is the predictive nature of artificial intelligence algorithms. To perform an effective evaluation of a decision *now*, the computer needs to generate a fairly high certainty of what will happen *in the future* based on that decision. Autonomous vehicles need to make thousands of calculations in a very short time span to make second to second evaluations while driving. Potential collision detection alone would require an algorithm to make very accurate predictive measures.

Facial recognition, voice recognition and other forms of complex pattern recognition are another vast area being explored using machine learning. Facial recognition is so difficult that a new computational model had to be developed that mimics the activity of the human brain. This technique is called *convolutional neural networking*, and the basis for the algorithm is similar to the neural links the human brain makes between neurons during learning (Fortune article, 2015). The stronger the link, the tighter the association, and as those links get stronger, learning takes place. The method for training an A.I. to recognize the relevant aspects of a photo that correspond to the human face is a painstaking process. The computer needs to build up enough data, at the pixel level, so that it can quickly measure the importance of image aspects at the smallest level. This precision is required to not only recognize a face against a background but also identify that unique face against a database of potentially millions of other faces.

Advances in machine learning have some obvious benefits in the commercial sector. Advertising and marketing are huge beneficiaries, but companies are also exploiting this technology to do predictive analysis for market trends. However, the benefits within academia are potentially more significant because they can lead to expedited research methods in areas such as the sciences, medicine, economics, psychology and disciplines that have a direct humanitarian impact. Research being done in desperate areas can be correlated and analyzed much more rapidly than was ever possible before. For some time now, grant funding agencies have required that research data be archived and stored in such a way that it can be retrieved and utilized by other researchers. In some cases, universities now have petabytes of research data from all areas of study. A very good example of this was the human genome project, which created a virtual roadmap to the human genetic structure. A huge network of computers utilized vast quantities of data to map the relevant gene sequences (Libbrecht and Noble, 2015). When the research was concluded, 20-25,000 genes were mapped and correlated and research continues to be done to determine the significance of the sequencing and the impact on human life. Even with the massive amount of data now available to scientists, a tremendous amount of work needs to be done to correlate specific gene sequences and anomalies with diseases such as cancer or predispositions to serious physical conditions, such as diabetes, anemia, ALS, etc.

Regarding the predictive aspect of machine learning, the use cases are somewhat standard fare. For example, recommender services have been in the development for quite some time, and the commercial world has taken advantage of the related algorithms in profitable fashion. In recent years, library discovery service vendors have taken to imitate this trend to a smaller degree. The relative success of any recommender service is dependent on large data sets of robust, correlative information that is buttressed by crowdsourcing "expert" opinions concerning relevancy. This is the lifeblood of Google, not only for the success of the core search engine but also for their core revenue line which is advertising. A "self-taught" algorithm is necessary for success in this arena because of the rapidly changing nature of Web content. What is relevant today might not be relevant tomorrow or a month from now. The data must be constantly analyzed, but effective analysis depends on the algorithm being regularly trained and retrained.

This is one of the reasons why natural language processing is still in the nascent stage. Human language needs to be interpreted. Interpretation is something that the human mind has been "pre-programmed" for, but for a computer whose basis of "thought" is a binary language, the subtleties must be learned. The outcome of any interpretation on the part of a computer is dependent largely on degrees of predictive certainty. The English language alone is mired with ambiguities that become apparent when a non-English speaker is learning the language. The spelling is not consistent, words take on specific contextual meanings and idioms run rampant. This is even more the case with auditory language processing. There is a scene from the movie 2010: The Year We Make Contact that demonstrates the difficulty. In this scene, the character Dr Chandra, a computer scientist, is restarting the Discovery's onboard computer system HAL, and he needs to tell the others with him that they need to be very careful when speaking to HAL:

Dr Chandra: Understand, nobody can talk. The accents will confuse him. He can understand me, so if you have any questions, please let me ask them. Good Morning, HAL.

HAL: Good morning, Dr Chandra.

Dr Chandra: Do you feel capable of resuming your duties?

HAL: Of Course. I am completely operational and all of my circuits are functioning perfectly.

[…]

HAL: Who are these people? I can only identify you […] although I compute a 65 per cent probability that the man behind you is Dr Floyd.

Inflection, tone and cadence can be calculated mathematically, and those patterns can assist a computer in recognizing words. However, finding meaning in words is an entirely different question. Following the success of Deep Blue with the victory over the world reigning chess champion in hand, IBM did not rest on their laurels. They continued to experiment with the super computer processing model, and today, they have in place a suite of services built on a technology platform that they call Watson[5]. The simplest explanation of what Watson represents is that it is a learning machine that has two primary functions. First, it can perform natural language processing with a high degree of accuracy, so questions can be posed to it that match the linguistic model of the researcher. Second, Watson can process unstructured data to reveal insights. IBM estimates that today over 80 per cent of all data is in some unstructured format. Consider for a moment how much time has been spent by library professionals classifying and organizing metadata to provide a relatively useful platform for discovering information. In all cases, that data, used to describe things such as books or other media, is ensconced in some serialized format which is highly structured. And even with that vast amount of data, it is still *metadata* – it is not the thing itself but rather a description of the thing that is ultimately useful to the student, professor or library patron.

Watson was designed to ingest information, process it comprehensively and then *interpret* it (Shah, 2011). And the information being referenced is not metadata but full text articles, news feeds, reports, social media and other forms of institutional data. This competes with traditional models, such as controlled vocabularies on several levels. The closer the computers come to being able to discern *meaning*, the less important specific keywords are because relevancy is no longer tied to textual patterns but rather meaning for meaning.

Logicae ex machina

IBM has begun partnering not only with other businesses to market their new machine learning services but also with academia. They offer services under several different categories that include medical research, discovery, pharmaceuticals and a category mysteriously labeled "engagement". From a business standpoint, engagement could refer to customer relations, concierge services, a knowledge base or a help desk. In the academic context, IBM is testing a student advice service. This is currently in "beta" at Deakin University in Australia, and the service is branded as *DeakInSync*. In partnership with the University, IBM has placed the service directly in the center of most student activities. When they use any major campus digital service, DeakInSync is available. The ubiquitous placement allows students to query the service using natural language in relation to almost any academic need. For every question asked, Watson is designed to learn in an evolutionary manner. Over time, it is able to more accurately respond to queries.

Machine learning is becoming an ever present, transparent part of modern life. It will gradually have a transformative effect upon the way information science is practiced and the manner in which libraries offer services. There are many practical ways in which libraries can contribute to this evolution using the vast array of data available. The traditional practice of using data to describe things using structured analysis and controlled vocabularies can enhance autonomous algorithms, so that they can more effectively serve public and academic needs through human-like interaction. As long as computers operate at a fundamental level using binary logic, machine judgments will always be based on probabilities and probabilistic decisions need to be informed through many sources and means. As Thomas Watson Jr, the second president of IBM put it, "Our machines should be nothing more than tools for extending the powers of the human beings who use them".

Notes

1. Turing actually called his test the "Imitation Game", IBID, p. 442.
2. See https://hbr.org/2015/07/what-every-manager-should-know-about-machine-learning
3. See www.theguardian.com/technology/2015/jun/10/baidu-could-beat-google-self-driving-car-bmw
4. See www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/
5. See www.ibm.com/smarterplanet/us/en/ibmwatson/

References

Fortune article (2015), available at: http://fortune.com/2015/06/15/facebook-ai-moments/

Libbrecht, M. and Noble, W. (2015), "Machine learning applications in genetics and genomics", *Nature Reviews Genetics*, Vol. 16 No. 6, pp. 321-332.

Shah, H. (2011), "Turings misunderstood imitation game and IBM's Watson success", *Invited Talk, Society for the Study of Artificial Intelligence and Simulation of Behavior (AISB) 2011 Convention*, University of New York, New York.

Stalker, D. (1978), "Why machines can't think: a reply to james moor", *Philosophical Studies*, Vol. 34 No. 3, pp. 317-320.

Turing, A.M. (1950), "Computing machinery and intelligence", *Mind*, Vol. 59 No. 236, pp. 433-460.

Corresponding author
Robert Fox can be contacted at: rfox2@nd.edu