The Degenerate Science of Machine Learning

Bat Gangly GANGLY@DME.UCSV.EDU
Department of Machine Earning, University of California, Silicon Valley, California 94000 USA

Abstract

This essay reviews the degenerate state of machine learning and its prospects for future debauchery. Metaphors are mixed liberally, in the hope that the resulting concoction will reveal insights hitherto obscured, without the hangover normally associated with such revels and revelations.

1. Machine Learning Comes of Age

The first workshop on machine learning was held in 1980, meaning our field has existed some 20 years, which is not quite drinking age but probably enough to avoid getting carded. This seems especially significant, given that researchers have long realized a central goal of machine learning is to overcome the *knowledge acquisition bottleneck*. This generic problem, depicted in Figure 1, concerns the difficulty of getting content into knowledge-based systems, with the funnel of machine learning providing a means to speed the process.

Researchers have explored many approaches to this task, many of them revolving around the simple idea of increasing the number of bottles, and thus the number of necks through which knowledge can enter. For example, work in computational learning theory has led to algorithms that, with high probability, acquire accurate knowledge with only a polynomial number of bottles, as shown in Figure 2. Typical methods involve polynomials with small constants, which has led to this approach being known as *six-PAC* learning.

Another, quite different, approach simply stores unopened bottles in memory, bypassing the need to fill them at all. As Figure 3 illustrates, these are sometimes organized into cases, so such techniques are often referred to as *case-based* learning. Some researchers have attempted to go even further, using *stacked* generalizations to include even more bottles, but the resulting knowledge structures can be unstable and have a disappointing tendency to fall over.

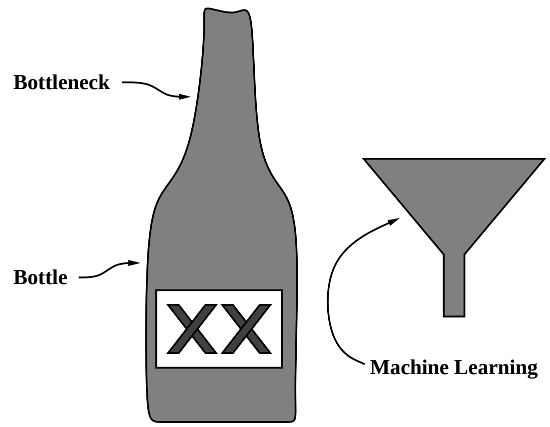

Figure 1. Machine learning as a response to the knowledge acquisition bottleneck.

In principle, the bottleneck should also come into play during knowledge extraction, but this seldom seems an issue at annual conferences, where the imbibing process has little need of mechanical aids. Indeed, this activity has itself inspired some important ideas in machine learning. In particular, the loss of brain cells during receptions originally led to ideas for pruning learned knowledge structures (e.g., Le Cun et al., 1989), and the resulting loss of motor control has motivated research on autonomous driving. Similar insights have produced psychological models of ambulatory drift like STAGGER (Schlimmer, 1986) and control policies like wall following, sometimes known as traversing a *support vector*.

Other popular methods rely on combining the overconfident predictions of different learned classifiers. One such technique gives different weights to these classifiers, then declares loudly to the world how much its accuracy has improved. Another takes a simple majority vote but also announces its improvement widely. For this reason, they are sometimes known as *boasting* and *bragging*. Not surprisingly, this tendency increases with the number of bottles at hand.

Machine learning has developed or incorporated a wide range of other methods, each with advantages but also

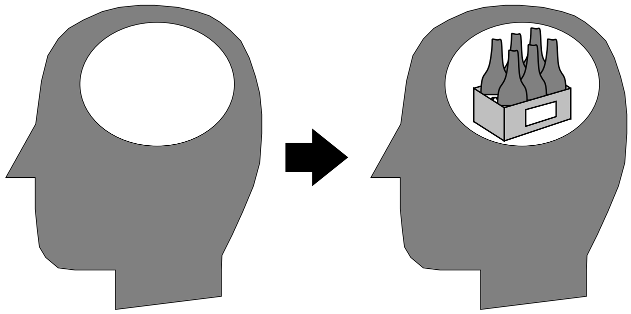

Figure 2. An example of six-PAC learning with a polynomial number of bottles.

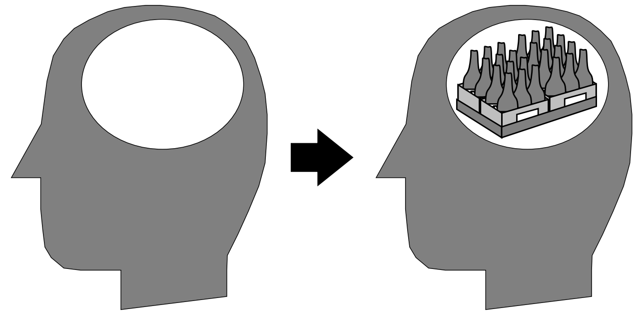

Figure 3. An example of case-based learning, with bottles organized into cases.

with drawbacks. For example, public-domain software for hidden Markov models tends to be so well concealed that researchers cannot find it. Techniques for lazy learning can be quite accurate, but may simply refuse to run unless they are in the mood. Other approaches operate smoothly but lead to problems in presentation. For instance, talks on learning monotone functions are delivered in such a flat voice that listeners often nod off. Similarly, papers on minimum description length, although concise, cannot be interpreted unless one has the right code. Finally, papers on large-margin methods require many pages to describe, and seldom meet the formatting constraints imposed by zealous program chairs.

One consistent trend has been the field's continuing incorporation of new techniques. Any computational method that improves its performance with experience is now considered part of machine learning, whatever its origin and whether its underlying metaphor involves search, memory, evolution, neurobiology, or clothing fashions. This attitude exists partly because the field has consciously defined itself in terms of learning tasks rather than methods. However, another factor has been political movements that encourage multistrategy diversity and equal treatment whatever a technique's ethnic origins. This pressure has also led many researchers to abandon traditional workstations for their development efforts and instead to use more PC devices.

2. Retaining Parental Respect

Although machine learning has now incorporated ideas from pattern recognition and statistics, researchers continue to explore the role of learning in other facets its parent field, artificial intelligence. Some work along these lines focuses on planning, where the existence of large problem spaces suggests learning heuristics to direct search. Since a common aim is to make problem solvers more efficient, this approach is often referred to as *speedup learning*. One generic finding from this paradigm is that problem solvers typically run faster in orbit, where they can operate with no weight state. This approach has also been successful in the domain of poker playing, where systems have been able to cut their losses by learning when to cache in their chips.

More recently, researchers have used machine learning to improve scheduling algorithms, where the typical concern is quality rather than efficiency. An important application has been determining the times for talks at annual conferences, where it has become especially difficult to avoid conflicts, since so many talks focus on only a few topics. Learning methods have also been applied to scheduling classes at universities, where, understandably, a popular representational scheme has been course coding.

Another promising area is robotics, where learning algorithms can improve the strategies that control the behavior of robotic agents. In this context, a popular approach involves using training data to update posterior probabilities, which is especially effective in controlling robotic end effectors. Another active paradigm in robotics involves learning from delayed rewards. Although sometimes discounted as impractical, this approach has been quite successful in the development of pool-playing robots, where it is known as *cue* learning. This application has also proved to be an excellent source of research funds.

Machine learning has also contributed to computer vision, where it has been used for problems ranging from object recognition to image segmentation. Early studies focused on finding objects like coffee cups in complex scenes, a task that made close contact with results in explanation-based learning and, besides, is especially useful the morning after conference receptions. Learning techniques have also been utilized in color interpretation, which lends itself especially to *Beigian* learning methods, since they assume high prior probabilities for certain colors. More recently, military funding has encouraged research on recognizing the rooftops of buildings in aerial photographs (Aloof & Gangly, 1997); unfortunately, most empirical results on this topic have been obscured by ceiling effects.

Another shift has been the increasing effort on learning in domains that involve natural language. For example, the popular area of *text learning* looks at tasks like filtering news stories and autofiling[1] electronic mail. Most work on this topic represents text as a bag of words, which makes them easy to carry but also sensitive to the types of bags used. For example, paper bags are biodegradable, making them more appropriate for methods that invoke singular value decomposition.

One important use of such methods involves automating the review process for conferences. Rather than taking scientists' time away from their research to review papers, we can use decisions from previous conferences to train a text-learning system that classifies submissions into accepts and rejects. This approach also reduces the variance due to different reviewers and, most important, reduces author complaints because they cannot blame the program chair for decisions. Moreover, initial studies with feature selection suggest that retaining only a few popular phrases and mathematical symbols lets one predict accurately the recommendations of many reviewers.

There also exists research on learning in dialogue systems, which carry out a conversation with the user.[2] Such systems must not only understand speech, but must generate utterances. This focus has been a main reason developers have moved away from implementing software in Lisp, which has a speech impediment, toward other programming languages. However, these have their own drawbacks; for instance, the popular language Java can only be used for learning tasks in which ground truth is available.

3. Grade Point Average

The field has also seen methodological changes, especially an increasing emphasis on evaluation. This often takes the form of 'bake offs', where each developer prepares his best algorithm and the community engages in a taste test to determine which method will be the flavor of the year. This produces some tasty bits, but more often than not, many algorithms get burned. Of course, these wares must be displayed in some manner, and a typical scheme is to lay them out on a table. Unfortunately, these tables are often so large that they collapse under their own weight, sometimes crushing those who come too close.

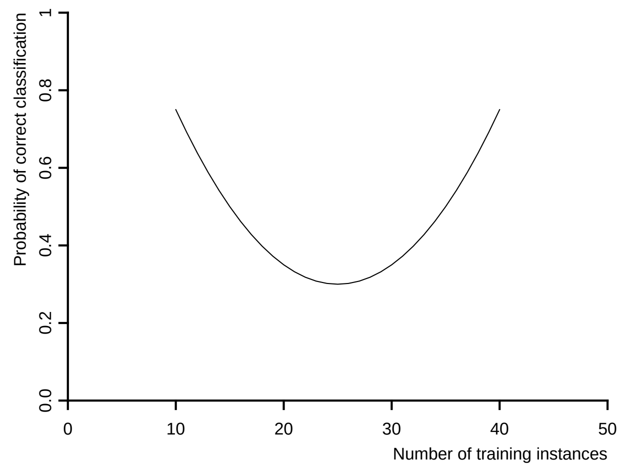

Figure 4. An experimental result that many researchers in machine learning find visually pleasing.

A safer approach involves summarizing experimental results in figures,[3] which are especially useful for describing the behavior of a single method under different conditions. These can reflect environmental variables, like the time of day, or internal parameters, like the system's mood. One varies systematically such factors and plots against them some performance measure, such as accuracy, precision, or, uh, whatever that other one is called. Other researchers can then decide which curves, and thus which algorithms, they find most visually pleasing. Figure 4 shows the most desirable type of result, which G. Dejong (personal communication, 1987) has called a 'happy graph'.

Of course, when evaluating a learning method, it is essential to measure performance on a test set that differs from the training set. Early studies split the available data haphazardly into different partitions, then averaged results over multiple splits. A more recent approach, known as *cross validation*, measures behavior when the system is in an especially bad mood. One version of this scheme, known as *leave one out*, is seldom used because the cases reserved for the test set consistently complained about being lonely. Future

[1] This activity should not be confused with the actions of car thieves when they are removing serial numbers from stolen vehicles.

[2] One exciting application of such dialogue systems lies in the area of phone sex, where learning about the user's personal tastes is paramount.

[3] A few researchers have also explored the acoustic presentation of experimental results. This has led to the increasing use of ROC curves and, more recently, Rap and even Country Western curves.

evaluation schemes will need to become ever less intrusive, as learning algorithms become more intelligent and subject to hurt feelings.

4. Toward Financial Independence

As it has matured, machine learning has devoted increasing attention to commercial applications and their most important byproduct: *machine earning*. Early applications focused on reducing costs in traditional industries like aircraft manufacturing, where gradient ascent methods are popular, and power generation, where techniques that rely on electromagnetic induction are still current. Another important area has been detecting fraud in the use of charge cards, where, naturally, credit assignment is a central issue.

More recently, the explosion of data on consumer behavior has led many companies to invoke machine learning methods to mine those data. Typically, this produces a knowledge base that is nearly as voluminous as the original data set and, thus, is almost guaranteed to include something the company finds interesting.[4] Occasionally, this strategy uncovers some important item that saves money or generates income, but, more often, simply finds the equivalent of used bottles from last year's receptions.

In the past few years, the World Wide Web has opened the way for other applications of machine learning. Many such systems rely on the Web's two-way nature to collect trace data about user interests, after which they use supervised learning methods to construct user profiles. However, a more lucrative approach instead uses ideas from reinforcement learning, not to reward and train the system, but rather to reward and train the *user*, thus ensuring his return to the Web site. We refer to such systems as *addictive user interfaces* (Gangly, 1999). Other variants of reinforcement learning, more appropriate for personalized decision aids that run on hand-held devices, are known as *palm DPs*.

Another important trend has been the incorporation of machine learning technology into Internet startups. Surprisingly, many researchers from the theoretical learning community have done quite well in raising funds for such ventures. Closer analysis suggests that this comes from their deep understanding of the VC dimension. However, they can suffer from a strong bias against exponential growth, which is not a liability in the new Internet economy. Nevertheless, many startups that build on ideas from machine learning seem destined for success, and we look forward to the day when they replace government agencies as the primary source of funding for the field.

[4] This approach is often marketed under the term *data mining* and is sometimes claimed as the illegitimate child of machine learning and databases, giving further evidence of our field's maturity and fertility.

Teknowledgements

The ideas presented in this paper were developed on a grant from the Kellogue Foundation, the conditions of which required us to run our experiments on cereal processors.

References

Aha, D. W. (1997). *Insights about lazy learning*. Unfinished manuscript, NCARAI, Naval Research Laboratory, Washington DC.

Aloof, M. A., & Gangly, B. (1997). Learning to detect rooftops from a distance. *Proceedings of the Image Understanding Workshop*. New Orleans: Morgan Kaufmann.

Gangly, B. (1987). The devolving science of machine learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 398–401). Irvine, CA: Morgan Kaufmann.

Gangly, B. (1995). *Elements, compounds, and polymers of machine learning*. San Francisco: Morgan Kaufmann.

Gangly, B. (1995). Applications of rule induction to machine earning. *Financial Communications of the ACM*, *38*, November, 55–64.

Gangly, B. (1999). User modeling in addictive interfaces. *Proceedings of the Seventh International Conference on User Manipulation* (pp. 357–370). Banff, Alberta: Springer.

Le Cun, Y., Denker, J. S., & Solla, S. A. (1989). Optimal brain damage. *Advances in Neural Information Processing Systems* (pp. 598–605). Denver, CO.

Schlimmer, J. C. (1986). Walking the line: Tracking ambulatory drift. *Proceedings of the Fifth Annual Meeting on Sobriety Testing* (pp. 502–507). Philadelphia, PA.