A perspective on human and artificial intelligence

Amitabh Basu*

May 25, 2020

Abstract

We try to explain the glaring discrepancy between human abilities and the abilities of state-of-the-art machine learning or artificial intelligence systems in fundamental learning tasks. In particular, for many basic tasks (e.g., vision), humans manage to learn with a far more limited data set, compared to the most sophisticated computer systems. We argue, with rigorous arguments, that the key to this could be the phenomenon of evolution.

The problem of learning from examples, a.k.a. the problem of *induction*, is a very old one and has received attention from scholars working in diverse fields, ranging from philosophy, natural and social sciences to statistics and computer science. How is it possible to detect patterns from a finite amount of data when there exist infinitely many different ways to extrapolate it (or at the very least, when the number of possible extrapolations is (super) exponential in the size of the data)? The most common response is some variant of "Occam's razor". We will focus on the particular formalization of this concept within statistical/machine learning theory.

Let us set the stage by focusing on the concrete task of *classification*, say within the field of human/computer vision. The learning theory perspective is that one has an *instance space* $\mathcal{X}$ (say, images of cats and dogs) and there is a "true" labeling of all instances with labels 0 or 1 (e.g., "cats" versus "dogs"). Formally, there is a "true" labeling function $f^\star : \mathcal{X} \to \{0, 1\}$. One observes a finite subset $S \subseteq \mathcal{X}$ with corresponding labels and from this one must extrapolate what $f^\star$ is. Formally, one observes pairs $(x_1, f^\star(x_1)), \ldots, (x_m, f^\star(x_m))$ and must make an estimate of what $f^\star$ is, or operationally speaking, given any $x \in \mathcal{X}$ (not necessarily in S), estimate what $f^\star(x)$ is. The difficulty in the task lies in the fact that $\mathcal{X}$ is typically much larger than S, and possibly even infinite in size. So how can one hope to figure out what $f^\star$ is by observing its values on a limited sample?

The answer afforded by learning theory is that one needs to make some assumptions about $f^\star$, otherwise the task is impossible in a mathematically precise sense. This is, in our opinion, a particular formalization of the "Occam's razor" principle. Let us make this more concrete.

Basic tenets of learning theory. Let $\mathcal{X}$ be described by d features, i.e., $\mathcal{X} \subseteq \mathbb{R}^d$ (a typical assumption made in learning tasks). In fact, for our purposes, it will suffice to consider the special case where each feature can take only two different values, say 0 or 1. In practice, this

*Department of Applied Mathematics and Statistics, The Johns Hopkins University.

is actually not much of a restriction because we represent things with floating point numbers and one can represent everything in binary. Thus, $\mathcal{X}$ is a finite set of size 2^d and can be represented by the vertices of the d-dimensional 0/1 hypercube. The notion that we observe a "limited" or "small" sample is formalized by saying that the size of the observed sample S is bounded by some polynomial in d (in many situations even linear in d). A basic question then arises: by observing such a small fraction of the values of $f^\star$, how can we make any intelligent estimates of its values on $\mathcal{X} \setminus S$? Stated another way, since there are 2^{2^d} (doubly exponentially many) possible labeling functions $f : \mathcal{X} \to \{0, 1\}$, how can a polynomial sized sample help us narrow down to the correct $f^\star$?

Learning theory says that in practice the true labeling function $f^\star$ is not any arbitrary function. We know it is one of a "controlled collection" of possible labelings. This is the precise form that "Occam's razor" takes within learning theory. For instance, a classical case is when $f^\star$ is assumed to be representable by a linear function; more precisely, there exists $a \in \mathbb{R}^d$ and $\delta \in \mathbb{R}$ such that for any $x \in \mathcal{X}$, $f^\star(x) = 0$ if $\langle a, x\rangle \leq \delta$ and $f^\star(x) = 1$ otherwise. This setting is called *halfspace learning* and goes back Rosenblatt's idea of a *perceptron* [3]. A basic theorem in discrete geometry says that the set of possible labeling functions that can be represented by linear functions is much smaller than the set of all possible 2^{2^d} labelings. In particular, there are only $2^{O(d^2)}$ labeling functions that can arise in this manner; for instance, see [1]. Now, a general theorem in learning theory (to be formally stated below as Theorem 0.1) says that if one knows *a priori* that $f^\star$ belongs to some "small" set $\mathcal{H}$, then one can "learn" $f^\star$ with only $O(\log(|\mathcal{H}|))$ many samples. Thus, in our halfspace learning scenario, one would need at most $O(d^2)$ observed data points (this can actually be improved to $O(d)$ with more careful reasoning).

To make this formal, one must allow for some small errors in our predictions. For example, consider two labeling functions $f_1, f_2 : \mathcal{X} \to \{0, 1\}$ that differ on a single point $\hat{x} \in \mathcal{X}$. Suppose further that $f^\star = f_1$ and we told that $f^\star$ is one f_1 or f_2, but not told which one, i.e., $\mathcal{H} = \{f_1, f_2\}$. By just looking at a single data point (or even $O(1)$ many data points), unless we happen to observe $\hat{x}$, we will not be able to distinguish between f_1 and f_2. On the other hand, if we arbitrarily choose f_1 or f_2 as our estimate of $f^\star$, then we make an error only on $\frac{1}{2^d}$ fraction of the points in $\mathcal{X}$, which is an acceptable error rate.

THEOREM 0.1. *[4, Corollary 2.3] Let $\mathcal{H}$ be any finite subset of 0/1 labeling functions on $\mathcal{X}$. Then, there exists an algorithm $\mathcal{A}$ that takes as input a sample $S \subseteq \mathcal{X}$ with labels and reports $f \in \mathcal{H}$ with the following property. For any $f^\star \in \mathcal{H}$, any probability distribution $\mathcal{D}$ on $\mathcal{X}$, any $\varepsilon > 0$ and any $\delta \in (0, 1)$, if S is a sample drawn i.i.d from $\mathcal{X}$ such that*

$$|S| \geq \frac{1}{\varepsilon} \log\left(\frac{|\mathcal{H}|}{\delta}\right),$$

then $\mathcal{A}$ applied to the data $\{(x, f^\star(x)) : x \in S\}$ reports $\hat{f} \in \mathcal{H}$ such that with probability at least $1 - \delta$ (over the sampling of S) one is guaranteed that $\mathbb{P}_{x\sim\mathcal{D}}[\hat{f}(x) \neq f^\star(x)] \leq \varepsilon$.

Theorem 0.1 can be extended to the setting where $\mathcal{X}$ is infinite, and even the family $\mathcal{H}$ of labeling functions is infinite, as long as there is enough "structure" in the family $\mathcal{H}$. This is based on the concept of *VC-dimension* and can be seen as a refinement of Theorem 0.1 towards formalizing the "Occam's razor" principle [4, Chapter 6] and [2, Chapter 12]. However, we

will not need these more sophisticated versions and the basic Theorem 0.1 will suffice for our purposes since we restrict ourselves to the finite (but exponential in d) instance space $\mathcal{X}$, and thus the set of all possible labelings is finite with 2^{2^d} possibilities.

Observe that if we take $\mathcal{H}$ to be the set of all possible labelings of $\mathcal{X}$, then Theorem 0.1 says that if one has $O(\log(2^{2^d})) = O(2^d)$ labeled data points, then learning is possible. In other words, we get a trivial bound, since $\mathcal{X}$ itself is of size 2^d. So this seems to say that we need to observe all the points in $\mathcal{X}$. Well, not quite, since Theorem 0.1 is only a sufficient condition for learning (in fact, as we pointed out learning can happen with infinite sized $\mathcal{H}$ as well, where the bound from Theorem 0.1 would be the trivial ∞). It is certainly possible that the bound in Theorem 0.1 is loose. Unfortunately, not really. Learning theory has so-called "no-free-lunch" theorems that show the impossibility of learning without assumptions on $\mathcal{H}$; in particular, in the setting where $\mathcal{X} = \{0,1\}^d$ and $\mathcal{H}$ is the set of all possible labelings.

THEOREM 0.2. *[4, Theorem 5.1] Let $m \in \mathbb{N}$ be such that $m \leq \frac{|\mathcal{X}|}{2}$. Then for any learning procedure $\mathcal{A}$, i.e., a map from learning samples $S \subseteq \mathcal{X}$ of size m to the space of $0/1$ labelings of $\mathcal{X}$, there exists a labeling function $f^\star : \mathcal{X} \to \{0,1\}$ such that for $1/7$ fraction of all possible samples $S \subseteq \mathcal{X}$, $\mathcal{A}(S)$ differs from $f^\star$ on at least $1/8$ fraction of data points in $\mathcal{X}$.*

Thus, for a constant fraction of all possible labeled samples, the learning procedure makes an error on a constant fraction of all data points, which is not an acceptable error rate.

COROLLARY 0.3. *In our particular setting, since $\mathcal{X} = 2^d$, Theorem 0.2 says that if we have no prior knowledge about $f^\star$, then* no learning procedure *can learn accurately unless it has access to $\Omega(2^d)$ labeled data points. Note that this is an information-theoretic bound; no assumption is made about whether the learning procedure is even computable (in the Turing machine sense) for general d, let alone efficiently computable.*

There are more refined versions of Theorem 0.2 even when we do not allow all possible labelings, but these will not be important for our main discussion [4, Chapter 6]. The trade-off unveiled by the combination of results like Theorems 0.1 and 0.2 is one version of the so-called *bias-variance trade-off* and is sometimes called the *bias-complexity trade-off*.

The paradox of human learning. We now come to the central paradox of human learning. Focusing on classification tasks in vision, there are several basic examples where artificial vision systems fail miserably and sometimes in a confounding manner. A recent high profile example includes so-called adverserial attacks on neural network based vision systems, where a humanly imperceptible change causes drastic misclassification [5].

This failure may be explained by appealing to Corollary 0.3. The "true" labeling function is a very complicated one that is sufficiently far removed from the classes $\mathcal{H}$ that are used in the learning process of these systems. Since good hypotheses about the structure of these true labels are not known, expanding the hypothesis class $\mathcal{H}$ to include very complicated functions runs into the barrier posed by Corollary 0.3; to find $f^\star$, one needs a huge number (exponential in d) many data points.

But then how does the human vision system perform these same tasks effortlessly, and with a seemingly limited data set? Why does this not contradict Corollary 0.3 (which make no assumptions like the Church-Turing hypothesis that the human vision system is computable, or complexity assumptions like $P \neq NP$)?

The resolution. Our answer to this paradox is *evolution*. We believe that the process of evolution has distilled a small class $\mathcal{H}$ that includes the "complicated" target function $f^\star$ of these vision tasks. A child's vision system, working with this "hard-wired", preset class $\mathcal{H}$ represented in the brain, can find $f^\star$ without violating Corollary 0.3 by appealing to Theorem 0.1.

In other words, evolution has taken millions of years to perform most of the learning task to distill out a manageable class $\mathcal{H}$ that includes the real $f^\star$, and leaves only the very last bit of learning for a particular human brain to perform within a span of a few years.

AI versus human intelligence. Note that our arguments above were not specific to vision, even though that was the setting we chose to illustrate the point. If our hypothesis about evolution being a major factor in the human learning problem is correct, this suggests that artificial intelligence systems have to overcome the headstart of millions of years afforded to humans by evolution by some means. It seem unlikely, at least to us, that this can be achieved by current state-of-the-art statistical/machine learning techniques alone. Insights from neuroscience and other fields are needed to "hone in" on the small class of "true" labeling mechanisms in nature, which currently are very "complicated" judging from the standard hypothesis classes $\mathcal{H}$ used in artificial learning systems. As we try to argue above, this is just a matter of perspective: the true labeling function $f^\star$ is "complicated" only because we dont know its structure and our standard hypotheses classes $\mathcal{H}$ are too far away from it, and blindly expanding our hypothesis classes in the hopes of catching $f^\star$ runs into the exponential data problem of Corollary 0.3; if we are told it is one of a small set of labeling mechanisms, its "complexity" disappears from the viewpoint of learning because of results like Theorem 0.1. It seems to us that this "prior" knowledge can only come from domain expertise like neuroscience, unless we are extremely lucky. In some instances we have indeed lucked out, as is amply illustrated by the magnificent successes of some artificial intelligence systems. But, in our opinion, these successes are perhaps more an exception than the rule.

References

[1] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[2] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[3] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[4] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.