Universal Learning vs. No Free Lunch results

Shai Ben-David
David R. Cheriton
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
CANADA
shai@cs.uwaterloo.ca

Nathan Srebro
Toyota Technological Institute
at Chicago
6045 S. Kenwood Ave.
Chicago, Illinois 60637
USA
nati@uchicago.edu

Ruth Urner
David R. Cheriton
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
CANADA
rurner@cs.uwaterloo.ca

Abstract

The so called No-Free-Lunch principle is a basic insight of machine learning. It may be viewed as stating that in the lack of prior knowledge (or inductive bias), any learning algorithm may fail on some learnable task. In recent years, several paradigms for ”universal learning” have been proposed and advocated. These range from paradigms of almost science-fictional nature, like ”Automation of science”, through practically oriented Deep Belief Networks, to theoretical constructs like Universal Kernels, Universal Priors and Universal Coding for MDL-based learning.

In this work we investigate this apparent contradiction by examining and analyzing several possible definitions of universal learning, proving a basic no-free-lunch theorem for such notions and discussing how they apply to the above mentioned universal learning paradigms.

1 Introduction

A basic insight of machine learning is that prior knowledge is a necessary requirement for successful learning. This principle is formally reflected by several no-free-lunch theorems, showing that no learning algorithm can be guaranteed to succeed on all learnable tasks. In other words, any learning algorithm has a limited scope of of phenomena that it can capture, or an inherent *inductive bias*, and there can be no universal learner.

Considering animal learning, it is easy to realize that each animal has a clear inductive bias; there are tasks that it is very good at learning, while other types of patterns that are very difficult, if not impossible, for that animal to pick up. Bait shyness in rats is a well known demonstration of such a bias. When rats encounter poisoned food, they learn very fast the causal relationship between the taste and smell of the food and sickness that follows a few hours later. In contrast to that, when the stimulus preceding sickness is sound rather than taste or smell, the rats fail to detect the association and do not avoid future eating when the same warning sound occurs [5].

Of course, one can argue that the inductive bias that the rat's learning mechanism displays must have developed through evolution, a process that can also be viewed as a type of learning. So, maybe nature does exhibit some kind of universal learning, though through a process that requires very large resources (in terms of training data and time).

Several paradigms of learning have been recently emerging. They seem to offer a notion of a general-universal learner that can be applied to *any* learning task without the need for prior knowledge, or inductive bias.

Examples of such paradigms range from paradigms of almost science-fictional nature, like ”Automation of science”, [6] through practically oriented Deep Belief Networks [2] to theoretical constructs like Universal Kernels, Universal Priors [4] and Universal Coding for MDL based learning.

In this work, we focus of classification prediction tasks in the batch statistical setting. We examine several candidate formal notions of universal learning for that model and analyze learnability w.r.t these notions. with respect to the issue of universal learning.

2 Different notions of universal learning

In this work we try to formalize the notion of universal learners, so that one could mathematically analyze their existence and scope. What is the meaning of a ”universal learner”? A naive answer would probably be ”a learning algorithm that can learn any target function”. Or, a learning algorithm that is (almost) as good as any other learning algorithm (on any possible task). It is not hard to see that a leaner cannot be guaranteed to be able to learn *every* labeling rule over some domain, unless it has access to a labeled training sample whose size is comparable to the size of the underlying domain (this is, essentially, the fundamental “no-free-lunch” phenomenon). In other words, universality w.r.t. *all* possible labeling rules is too much to ask for. However, it can be argued that such universality is not really needed, since realistic labeling rules are not completely arbitrary, and should have some kind of regularity. How should such regularity be formalized? We consider two possible answers; requiring that the labeling rules that the a universal learner should be able to handle are computable functions, or defining universal learning by requiring such a learner to be able to compete with every other computable learner. One should note that the second requirement is stronger, since for any computable function there is a (trivial) successful computable learning algorithm that just outputs that function (ignoring its input training sample).

From a theoretical point of view, one wishes to have some kind of performance guarantees for such learners. We will examine several candidate performance requirements that one can expect such universal learners to meet. What kind of performance guarantees can such a learner enjoy?

We examine two resources, computational time and training sample size. Of course, the needed quantities of such resources depend on the task at hand. Ideally, the resources needed for learning (in our context, statistical classification prediction) would depend only on the accuracy and confidence parameters, ϵ and δ, and be independent of the labeling function and of the underlying data distribution. However, the classical statistical learning results show that such performance guarantees can be achieved only if the class of potential labeling functions (or the class of predictors the learner competes with) has a finite VC dimension. This restriction contrasts any reasonable interpretation of universality for learning. We therefore consider two relaxations; allowing the needed resources to vary with the labeling function (or with the learner that the universal learner is compared with), and allowing them to depend on the underlying data distribution.

In other words, we need to relax the PAC notion of learnability, to allow some non-uniformity in the allowed sample sizes.

3 Results

Allowing non-uniformity w.r.t. the data distribution means that the sample size needed to learn some target labeling function within some fixed accuracy parameter, can vary from one data distribution to another. We will discuss why such a notion of leanability is too weak to be useful. As a formal support to this clam, we show that, with such a notion of learnability, the algorithm MEMORIZE (that just memorizes the training sample, predicts the training sample label on every instance that appears in the training sample, and zero as the label for any out of sample instance) is a successful universal leaner with respect to the class of all possible classification functions over any countable domain.

The other parameter along which the definition of PAC learnability may be relaxed, is the dependence of the required training sample size on the labeling (target) function (while being distribution-free, or, in other words, independent of the underlying marginal data distribution). This seems to be a natural notion to study. In contrast with the common notion of PAC learning, such learners are allowed to vary the sample size as a function of the target labeling - learning more complex functions is allowed to require larger sample sizes. For lack of a better name, let us call this notion $\mathcal{D}$-uniform-learnability.

We will analyze, for which classes $\mathcal{H}$ is there a $\mathcal{D}$-uniform universal learning rule w.r.t $\mathcal{H}$.

If one ignores any computational requirements, we can show that there exists a $\mathcal{D}$-uniform universal learning rule w.r.t $\mathcal{H}$ if and only if $\mathcal{H}$ is a countable union of classes of finite VC-dimension. (The negative side of this statement is due to the fact that the class of all predictors over an infinite set cannot be represented as a countable union of classes of finite VC-dimension [1].) In particular, for every countable class $\mathcal{H}$ (e.g., for the class of all computable functions), there exists a learner which is a $\mathcal{D}$-uniform universal learning rule w.r.t $\mathcal{H}$. However, such learners may well be non-computable (as functions from labeled samples to classifiers). Non-computable learners are, of course, useless for any practical application.

The next plausible relaxation of the definition of universal learners is requiring universality only with respect to the class of all computable labeling functions, or with respect to the class of all computable learners. In this case too, following a result of Soloveichik [7], we can show that no computable universal learner exists.

It is interesting to note in this context that indeed, for any universal description language, the minimal description length of a string (or its kolmogorov complexity) is not a computable function. So a strict MDL principle cannot be applied as a learning rule by any computable learner.

Interestingly, one can regain the computability of universal learners w.r.t the class of all computable learners by relaxing the training sample size requirements. This is demonstrated by a result of Goldeich and Ron [3].

Many interesting questions remain to be discussed. Is the class of computable learners (or, the class of computable functions) the 'right' class with respect to which universality should be defined? Maybe one can settle for a weaker class, say of all efficiently computable learners. Is there an efficient leaner that is universal w.r.t. that class? What are the universality claims that are implicitly made by the advocators of deep belief network learning, or of the automation of science research? How useful is a universal learner whose required sample complexity may vary with the learnt task?

In our talk we intend to raise and discuss these questions.

References

[1] Shai Ben-David and Leonid Gurvits. A note on vc-dimension and measure of sets of reals. *Combinatorics, Probability & Computing*, 9(5):391–405, 2000.

[2] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[3] Oded Goldreich and Dana Ron. On universal learning algorithms. Online version, 1996.

[4] Marcus Hutter. Universal learning theory. *CoRR*, abs/1102.2467, 2011.

[5] Linda Philips Brett John Garcia and Kenneth Rusiniak. Limits of darwinian conditioning. In S.B. Klein and R.R. Mowrer, editors, *Contemporary learning theories: instrumental conditioning theory and the impact of biological constraints*, pages 181–204. L. Erlbaum, 1989.

[6] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science 3*, 324 no. 5923 pp. 81-85, 2009.

[7] David Soloveichik. Statistical learning of arbitrary computable classifiers. *CoRR*, abs/0806.3537, 2008.