Machine Learning Beyond Function Approximation I: Nondeterministic PAC-learning

Xiyu Zhai, Alexander Rakhlin

Contents

1 Introduction

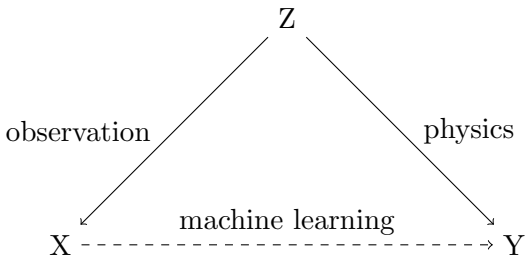
Recently deep learning has been a thing. But no theories. And deep learning is still no match for humans. What's missing?

Good applied theories should satisfy the following criterions,

- simple. Convoluted descriptions lead to no clean theories;
- relevant. Machine learning theories are not studied for its own sake, so it's important to have relevance to
- insightful. Good theories can make predictions that could guide future experiments, instead of just explaining the known empirical success. A good set of assumptions should make it possible to draw insights and make predictions about future directions of experiments and engineering.

Theoretical Physics has been extremely successful in the light of these criterions, and we have yet to see the same level of success happening for theoretical machine learning. As Feifei Li commented, we are in a pre-Newton era of AI. If we compare physics with machine learning, machine learning is more difficult to establish elegant and effective theories due to

- medium scaling. Scaling in machine learning is neither too big nor too small to allow simplification. Statistical Physics or astronomical Physics deals with a gigantic system with the number of particles being around the order of 10^{23} and things can be simplified with renormalization. Atomic Physics deals with a tiny system with only a few number of particles and things are simplified to be equations with several variables. Anything between is notoriously difficult in Physics. However, the scaling of deep learning is not as nice to simplify things to adapt either renormalization or precise analysis.
- incomplete information. Physics studies from a God's perspective, no loss of information when doing the calculation. However, machine learning deals with incomplete information on a daily basis. Informally, one can summarize using the following diagram, which shall reappear with more details,



The solid arrows are well-defined and easily represented by analytic functions, but the dashed arrow is often not well-defined and hard to be analytically represented as $Z \to X$ is not invertible. In computer vision, Z is the complete physical world and X is the projection of Z to a certain camera and Y is a certain well-defined quantity of

the physical world. In natural language processing, Z can be the meaning one wants to convey and X is the actual word one speaks to convey the meaning and Y is a certain aspect of the meaning Z.

1.1 In this paper

We extend the PAC-learning framework.

1.2 Contributions

hello

1.3 Notations

In machine learning, people tend to use Y^X for the set of maps from X to Y. Here we use $X \to Y$ instead for consistency. Our paper address problems lying the intersection of computer science, mathematics, and statistics. In computer science, the curry notion is $X \to Y$ and one use ':' for type annotation. In mathematics, one tends to write $f : X \to Y$. So it seems that $X \to Y$ would be a notation that comes close to all sides.

2 Related Work

Bayesian.

Compression. Ilya talks

3 Limitations of PAC-learning Framework

In this section, we point out that the assumptions in the PAC-learning framework are too restrictive for many important applications. In fact, it might be too restrictive for most important applications.

Let $\mathcal{X}$ be the input space, and $\mathcal{Y}$ be the output space, and we wish to learn a function f_* from $\mathcal{X}$ to $\mathcal{Y}$, called the ground truth. We are given $x_1, \ldots, x_n$ sampled according to a distribution $\mathcal{P}$, and also $y_1, \ldots, y_n$ with $y_i = f_*(x_i)$, and we want to get a function f such that

$$\mathbb{E}_{x \sim \mathcal{P}} l(f_*(x), f(x))$$

is as small as possible.

Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ to $\mathcal{Y}$, called the hypothesis class, we say that $\mathcal{H}$ is PAC-learnable if there exists an algorithm $\mathcal{A}$ that takes in a training set $S \in (\mathcal{X} \times \mathcal{Y})^n$ and returns $\mathcal{A}_S \in \mathcal{H}$ in polynomial time such that

$$\mathbb{E}_{x\sim\mathcal{P}} l(f_*(x), \mathcal{A}_S(x)) < \epsilon$$

Note that the hidden assumption is that functions in $\mathcal{H}$ can be computed in a "straightforward" manner. By "straightforward", we mean that the mathematical description gives directly a feasible way of computation, just like the description of neural networks gives directly a way to compute them in an acceptable amount of time, although it might not be optimal.

The problem with this setup is that it's too simplistic. In a sense, it is as naive as assuming all differential equations with analytic representation have analytic solutions. Many important problems are naturally represented by a hypothesis class containing functions not straightforward to evaluate. Not taking these problems into consideration is somehow like not considering equations without an analytic solution, leading to the unavoidable large gaps between theories and practices.

Here are some examples that show that such problems do exist across different AI domains.

Example (Computer Vision, Image Recognition, Deformable Template Matching, Informal) Image classification in computer vision is about using computers to tell the class of an image, i.e. $\mathcal{X}$ being the space of images, and $\mathcal{Y}$ being a small set of categories. A reasonable simplification for many cases is that the process can be realized through deformable template matching. A deformable template is a map ϕ from a deformation space Γ to the same image space $\mathcal{X}$. Let $d_{\mathcal{X}}$ be the natural metric on image space $\mathcal{X}$, then we can construct a measure of the degree of an image fitting a template through

$$s(\phi, x) := \inf_{\gamma\in\Gamma} d_{\mathcal{X}}(x, \phi(\gamma))$$

It's not as easy to compute as to define. Γ can be of very high dimensionality making it nearly impossible to evaluate $s(\phi, x)$ faithfully.

Now suppose that we have a list of pairs of deformable template and category, say, $(\phi_1, y_1), \ldots, (\phi_n, y_n)$, the ground truth can be given by

$$f_*(x) = y_i \text{ where } i = \operatorname*{argmax}_i s(\phi_i, x).$$

Evaluating f_* is thus not straightforwardly easy.

For more concrete examples, we give a full mathematical characterization of the MNIST dataset in the appendix.

Example (Natural Language Processing, Word, Informal)

Definition.

4 Nondeterministic PAC-learning (NPAC-learning)

The nondeterministic PAC-learning is trying to be an extension of PAC that deals with more cases, but not everything. We would like to point out there are more beyond the scope of this theory.

Let Γ be a set called the configuration space, and $\mathcal{F} \subseteq (\mathcal{X} \times \mathcal{Y} \times \Gamma \to \mathbb{R})$ be a set called the class of structure functions. We shall define the nondeterministic hypothesis class $\mathcal{H}(\Gamma, \mathcal{F})$ as follows.

For each $f \in \mathcal{F}$ we define an energy function $e_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as

$$e_f(x, y) := \sup_{\gamma \in \Gamma} f(x, y, \gamma), \tag{1}$$

and we define a hypothesis function $h_f : \mathcal{X} \to \mathcal{Y}$ as

$$h_f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} e_f(x, y) \tag{2}$$

For the above to be well-defined, one assumes that $\mathcal{Y}$ is finite and has a default ordering and argmax returns y with the largest ordering if the global maxima are not unique.
$y \in \mathcal{Y}$

Finally we define the nondeterministic hypothesis class as

$$\mathcal{H}(\Gamma, \mathcal{F}) := \{h_f : f \in \mathcal{F}\} \tag{3}$$

Remark. *We are still call it hypothesis class instead of concept class because there is no guarantee that the ground truth lies in this hypothesis class.*

then

Remark (Equivalence with Bayesian to the Limit)**.** *h_w is equivalent to*

Let $\mu = \mu_{\mathcal{X}} \times \mu_{\mathcal{Y}} \times \mu_\Gamma$ be a standard measure over $\mathcal{X} \times \mathcal{Y} \times \Gamma$, let P_ε be another distribution defined by

$$p_\varepsilon(x, y, \gamma) := \left.\frac{dP_\varepsilon}{d\mu}\right|_{x,y,\gamma} = Ce^{f(x,y,\gamma;w)/\varepsilon}$$

where C is a normalization constant.

Given only $x \in \mathcal{X}$, we have

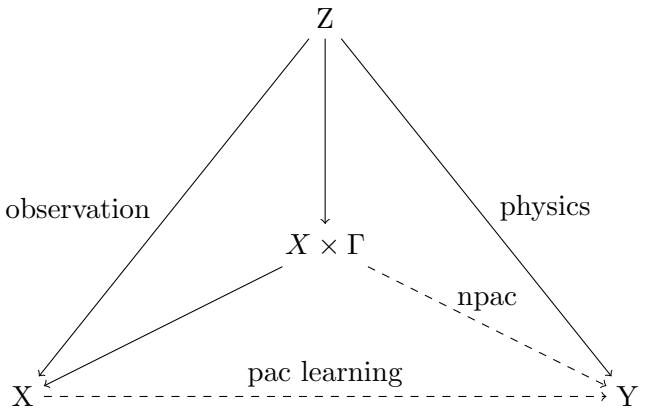
$$p_\varepsilon(y|x) = \frac{\int_\Gamma p_\varepsilon(x, y, \gamma) d\mu_\Gamma(\gamma)}{\int_{\mathcal{Y}} \left(\int_\Gamma p_\varepsilon(x, y, \gamma) d\mu_\Gamma(\gamma)\right) d\mu_{\mathcal{Y}}(y)}. \tag{4}$$

As $\varepsilon \to 0$, and assuming that with ϵ-margin, then

$$p_\epsilon(y|x) \to C1_{y=h_w(x)} \tag{5}$$

where C might be as large as $\delta(0)$.

Consider the following diagram



The diagram suggests that it can be easier to construct a function from $X \times \Gamma$ to Y then from X to Y because of additional information.

5 Basic Function Properties

Proposition (Lipschitz)**.** Suppose that f is L-Lipschitz w.r.t $\mathcal{X}$ then e_w is L-Lipschitz w.r.t $\mathcal{X}$.

Proof. Obvious by unravelling the definitions and note that sup keeps the Lipschitzness. □

6 Generalization Bound

7 Select and Verify

7.1 Differentiable

7.2 Neural Network

8 Shape Theory

A Description of Function Classes

In machine learning theory, people just refer to function classes as mathematical sets, without further structures. For our purposes, this is not enough. So

B Mnist

Here we describe briefly a function in mathematical terms which we believe is the ground truth for the MNIST dataset. Details can be seen in appendix.

We take the convention that the fill of the digits is white and the background is black. The image is represented by a $[0,1]$-valued 28×28 matrix $I = (I_{ij})_{0 \le i \le 27, 0 \le j \le 27}$.

- **digit one of the simplest kind**, which constitutes 95% of all images of digit one.

  A typical image looks like this:

  [an image here]

  Think about how it's drawn. The person when writing down a digit one like this has an ideal version in mind, a straight line that is almost vertical. So take Γ_1 to be the set of straight lines with slopes satisfying some easy constraint. Then we should define $\mathfrak{s}(x;\gamma)$ for $\gamma \in \Gamma_1$ such that

  - for "most" points over γ, it's surrounded by white pixels;
  - for "most" non-white pixels, it's away from γ.

  One choice could be

  $$\mathfrak{s}(x;\gamma) = a_1 \int_0^1 \max_{(i,j)\in[27]\times[27]} 1_{\|\gamma(t)-(i,j)\|_2<\epsilon} I_{ij} dt - a_2 \sum_{(i,j)\in[27]\times[27]} 1_{I_{ij}<0.5} \mathrm{dist}((i,j),\gamma) \tag{6}$$

  where ϵ is an appropriate small number and $a_1, a_2 > 0$ are appropriate coefficients.

  In fact the function applies when γ is any path, not necessarily a straight line. Formally it is defined over the path space (without basepoint) $\Gamma = M([0,1],[0,1]^2)$.

  The dimensionality of the configuration space is 6, which can possibly be reduced to 5.

- **digit seven of the simplest kind**.

  A typical image looks like this:

  [an image here]

  Here we consider all continuous $\gamma : [0,2] \to [0,1]^2$ such that $\gamma|_{[0,1]}, \gamma|_{[1,2]}$ are straight line segments. Additionally, there should be some constraint on the positions of $\gamma(0), \gamma(1), \gamma(2)$ such that $\overline{\gamma(0)\gamma(1)}$ is very close to being horizontal, and $\overline{\gamma(1)\gamma(2)}$ should be roughly vertical downward.

  The score function $\mathfrak{s}$ is the same.

  The dimensionality of the configuration space is 6, which can actually be reduced to 5.

- **digit zero**.

  We consider smooth curves $\gamma : [0, L] \to [0, 1]^2$ with arc length parametrization such that the mean curvature is always nonnegative, i.e.

  $$\|\gamma'(t)\| \equiv 1 \tag{7}$$

  and

  $$\gamma''(t) \times \gamma'(t) \geq 0 \tag{8}$$

  Additionally, we require that $\gamma(0)$ is very close to $\gamma(L)$.

  We should also require that γ is nondegenerate, which can be characterized by isoperimetric inequality.

  All these γ form Γ_0.

  And we still use the same score funtion $\mathfrak{s}$.

- **general case**. Fix a graph $G = (V, E)$. Give it a natural topological structure and identify each e with $[0, 1]$. For each $e \in E$, we give assign a σ_e which is one of the following classes of curves:

  - nonconvex but not straight
  - nonconcave but not straight
  - straight.

  We define the total space as

  $$\Omega_G = \left\{\gamma \in M(G, [0, 1]^2) : \forall e \in E, \gamma|_e \in \sigma_e\right\}. \tag{9}$$

  Then the configuration space Γ_G is a subset of Ω_G such that it is given by a boolean function s in the sense that

  $$1_{\Gamma_G}(\gamma) = s((\gamma(v))_{v\in V}, ((\gamma|'_e(0), \gamma|'_e(1), \mathrm{dist}(\gamma|_e, \overline{\gamma|_e(0)\gamma|_e(1)})))_{e\in E}) \tag{10}$$