The Lack of A Priori Distinctions Between Learning Algorithms aka No Free Lunch Theorems for Learning

Nikita Kazeev, based on David H. Wolpert

September 17, 2019

Abstract

The objective of supervised learning is generalization, i. e. learning a relation and predicting on yet unseen data. In this paper we show that this problem can not be solved in general, for all target relations. If there are no restrictions on the structure of the problem, then for any two algorithms there are "as many" targets on which each outperforms the other. This hold true even for random guessing.

1 Introduction

This is a streamlined and simplified representation of some results from *Wolpert, David H. "The lack of a priori distinctions between learning algorithms." Neural computation 8.7 (1996): 1341-1390.*

The objective of supervised learning is generalization — "learning" information about a process from a set of samples and then using it to predict the outcome for examples yet unseen. It is widely advertised as an assumption-free, "data-driven" approach, in contrast to explicit statistical models — see the famous paper by Breiman [1].

In this paper we recite several results obtained by David H. Wolpert, which show the impossibility of a machine learning algorithm, that would work for all targets. The paper does in no way argue that all algorithms are equivalent *in practice*. There are of course algorithms that perform well over the targets we see in the real life. But, as we show here, for any such algorithm there are many targets at which it gets confused by the data and preforms worse than *random guessing*.

2 Formalism

Begin with two finite sets $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{X}$ is the input set, $\mathbf{Y}$ is the output set. Define a metric (loss function): $L(y_1, y_2) \in \mathbb{R}$, $y_1, y_2 \in \mathbf{Y}$. Introduce the target function $f(x, y), x \in \mathbf{X}, y \in \mathbf{Y}$ – an $\mathbf{X}$-conditioned distribution over $\mathbf{Y}$. Select a training set d of m $\mathbf{X} - \mathbf{Y}$ pairs, according to some distribution $P(d|f)$. Let d_X be its $\mathbf{X}$ component and d_Y be the $\mathbf{Y}$ component. Select a test point $q \in \mathbf{X}, q \notin d_X$ – we are interested in the generalization power. Such selection

is called off-training set (OTS). Take a classifier, train it on d, use it to predict on q. Let y_H be the prediction. Any classifier is completely described by its behavior, $P(y_H|q,d)$. Also sample the target distribution f at point q, let y_F be the result. Define loss $c = L(y_H, y_F)$.

In this paper we determine several averages of conditinal probabilities for loss values over all valid f. f is uniquely specified by an $|\mathbf{X}| \times |\mathbf{Y}|$ matrix of real numbers, so we can write a multidimensional integral $\int A(f)df$ and average $E_f A(f) = \int A(f)df / \int 1df$. All integrals over targets f in this paper are implicitly restricted to the valid $\mathbf{X}$-conditioned distributions over $\mathbf{Y}$. We do not evaluate the integrals explicitly, but for the sake of clarity, it is worth to discuss them.

$\sum_y f(x,y) = 1$. Therefore, f is a mapping from $\mathbf{X}$ to an $|\mathbf{Y}|$-dimensional unit simplex. The integration volume F is a Cartesian product of unit simplices, which can be expressed using a combination of Dirac delta functions and Heaviside step functions.

In this paper we consider *homogeneous loss*, meaning that

$$\exists \Lambda[\mathbb{R} \to \mathbb{R}] : \forall c \in \mathbb{R}, \forall y_H \in \mathbf{Y} : \sum_{y_F \in \mathbf{Y}} \delta\left[c, L(y_H, y_F)\right] = \Lambda(c), \tag{1}$$

where δ is the Kronecker delta function. Intuitively, such L have no a priori preference for one $\mathbf{Y}$ value over another. For example, zero-one loss ($L(a,b) = 1$ if $a \neq b, 0$ otherwise) is homogeneous, and quadratic ($L(a,b) = (a-b)^2$; $a, b, \in \mathbb{R}$) is not. A weaker version of No Free Lunch Theorem still holds for non-homogeneous loss, it is discussed in [2]

Likelihood $P(d|f)$ determines how d was generated from f. It is *vertical* if $P(d|f)$ is independent of the values $f(x, y_F)$ for $x \notin d_X$. We do not require that the $\mathbf{Y}$ components for the test and train sets are generated from the same distribution. Verticality is needed to prevent leakage of the test set labels into the train set. For example, the conventional procedure, where d is created by repeatedly and independently choosing its $\mathbf{X}$ component d_X by sampling some distribution $\pi(x)$, and then choosing the associated d_Y value by sampling $f\left(d_x(i), y\right)$, results in a vertical independent and identically distributed (IID) likelihood

$$P(d|f) = \prod_{i=1}^{m} \pi(d_X(i)) f(d_X(i), d_Y(i)). \tag{2}$$

3 No Free Lunch Theorem

The general idea behind the No Free Lunch theorems is calculating the uniform average over f of the distribution of classifier performance (loss c) conditioned on various variables.

3.1 Example

Before writing the formal theorems, let us illustrate the counter-intuitive idea of No Free Lunch on a simple example.

Take $\mathbf{X} = \{0,1,2,3,4\}$, $\mathbf{Y} = \{0,1\}$, a uniform sampling distribution $\pi(x)$, zero-one loss L. For clarity we will consider only deterministic f, i. e. $f : \mathbf{X} \times \mathbf{Y} \to \{0,1\}$. Set the number of distinct elements in the training set $m' = 4$. Let

algorithm A always predict the label most popular in the training set, algorithm B the least popular. In case the numbers of labels are equal, the algorithms choose randomly.

Let $x_i \in \mathbf{X}$ be the feature vector and $y_i \in \mathbf{Y}$ the label for i-th object. Let c_A be the loss on the test element for the algorithm A, c_B for B. We show that $E_f(c|f, m')$ is the same for A and B.

1. There is only one f for which for all $\mathbf{X}$ values, $\mathbf{Y} = 0$. In this case algorithm A works perfectly, $c_A = 0$, algorithm B always misses, $c_B = 1$.

2. There are $C_5^1 = 5$ fs with only one $y = 1$, the rest being 0. For each such f, the probability that the training set has all zeros is 0.2. For these training sets, the true test label is 1, A predicts 0, B predicts one, $c_A = 1$, $c_B = 0$. For the other 4 training sets, $c_A = 0$, $c_B = 1$. Therefore, the expected value of $Ec_A = 0.2 \times 1 + 0.8 \times 0 = 0.2$ and $Ec_B = 0.2 \times 0 + 0.8 \times 1 = 0.8$

3. There are $C_5^2 = 10$ fs with two $y_i = 1$. There is a 0.4 probability, that the training set has one 1. Therefore, the other 1 is in the test set, and $c_A = 1$, $c_B = 0$. There is a 0.6 probability that the train set has two 1s. In that case both algorithms guess randomly and $Ec_A = Ec_B = 0.5$. So for each f, $Ec_A = 0.4 \times 1 + 0.6 \times 0.5 = 0.7$, $Ec_B = 0.4 \times 0 + 0.6 \times 0.5 = 0.3$. Note that B outperforms A.

4. The cases with three, four and five 1s are equivalent to the already described.

5. Averaging over f, we have $E_f c_A = \frac{1 \times 0 + 5 \times 0.2 + 10 \times 0.7}{1+5+10} = 0.5$, $E_f c_B = \frac{1 \times 1 + 5 \times 0.8 + 10 \times 0.3}{1+5+10} = 0.5$

3.2 Theorems

Lemma 1.

$$P(c|d, f) = \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q, d\right) P\left(y_F|q, f\right) P\left(q|d\right) \tag{3}$$

Proof.

$$c = L\left(y_H, y_F\right) \tag{4}$$

$$\begin{aligned} P\left(c|q, d, f\right) &= \sum_{y_H, y_F} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H, y_F|q, d, f\right) \\ P\left(c|d, f\right) &= \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H, y_F|q, d, f\right) P\left(q|d\right) \\ &= \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q, d\right) P\left(y_F|q, f\right) P\left(q|d\right) \end{aligned} \tag{5}$$

□

Lemma 2. *For homogeneous loss L, the uniform average over all f of $P\left(c|d, f\right)$ equals $\Lambda\left(c\right)/r$.*

For any training set, any $|\mathbf{Y}|$, any homogeneous loss L, any OTS method $P(q|d)$ of selecting the test point, including sampling the same $\pi(x)$, that was used to select d_X, for any learning algorithm, the average performance over all possible targets is a constant.

This result ignores the relationship between d and f. In other words, the $\mathbf{Y}$ values for train and test sets are generated from different distributions. Thus it is not particularly interesting in itself, but will rather serve us a base for further inquiries.

Proof. Using Lemma 1, the uniform average over all targets f of $P(c|d,f)$ can be written as

$$E_f\left[P\left(c|d,f\right)\right] = \sum_{y_H,y_F,q} \delta\left[c, L\left(y_H,y_F\right)\right] P\left(y_H|q,d\right) E_f\left[P\left(y_F|q,f\right)\right] P\left(q|d\right) \quad (6)$$

$$E_f\left[P\left(y_F|q,f\right)\right] = E_f f(q,y_F) \quad (7)$$

The integration is over all possible values of f. The space of all possible values is the same for all components, thus the average is a constant that does not depend on q and y_F. Also,

$$\sum_{y_F} E_f\left[f(q,y_F)\right] = E_f\left[\sum_{y_F} f(q,y_F)\right] = 1, \quad (8)$$

therefore

$$E_f\left[f(q,y_F)\right] = 1/r. \quad (9)$$

Using the homogeneity property of L:

$$E_f P\left(c|d,f\right) = \sum_{y_H,q} \Lambda(c) P\left(y_H|q,d\right) P\left(q|d\right)/r = \Lambda(c)/r \quad (10)$$

□

Theorem 3.1. *For vertical $P(d|f)$, and a homogeneous loss L, the uniform average over all targets f of $P(c|f,m) = \Lambda(c)/r$*

For any $|\mathbf{Y}|$, any homogeneous loss L, any fixed training set size m, any vertical method of training set generation, including the conventional IID-generated, for any OTS method $P(q|d)$ of selecting the test point, including sampling the same $\pi(x)$, that was used to select d_X, for any learning algorithm, the average performance over all possible targets is a constant.

This is the result advertised in the beginning. If an algorithm "beats" some other, including the random guess, on some f's, it will necessary lose on the rest, so that the average losses will be the same.

Proof.

$$P\left(c|f,m\right) = \sum_{d:|d|=m} P\left(c|d,f\right) P\left(d|f\right) \quad (11)$$

From Lemma 1:

$$P(c|d,f) = \sum_{y_H,y_F,q} \delta\left[c, L\left(y_H,y_F\right)\right] P\left(y_H|q,d\right) P\left(y_F|q,f\right) P\left(q|d\right). \quad (12)$$

Because we consider OTS error, $P(q|d)$ will be non-zero only for $q \notin d_X$, so $P(y_F|q,f)$ only depends on components of $f(x,y)$ that correspond to $x \notin d_X$.

We also know that $P(d|f)$ is vertical, so it is independent of the values $f(x,y_F)$ for $x \notin d_X$.

Therefore the integral can be split into two parts, over dimensions corresponding to d_X and $\mathbf{X} \setminus d_X$:

$$E_f\left[P\left(c|f,m\right)\right] = \sum_{d:|d|=m}\left[\frac{\int P\left(c|d,f\right)df_{x\notin d_X}\int P\left(d|f\right)df_{x\in d_X}}{\int 1df_{x\notin d_X}df_{x\in d_X}}\right] \tag{13}$$

Again using $P(c|d,f)$ independence from $x \in d_X$ and Lemma 2:

$$E_{f_{x\notin d_X}}\left[P\left(c|d,f\right)\right] = E_f\left[P\left(c|d,f\right)\right] = \Lambda(c)/r \tag{14}$$

$$E_f\left[P\left(c|f,m\right)\right] = \Lambda(c)/r\sum_{d:|d|=m}\left[\frac{\int P\left(d|f\right)df_{x\in d_X}}{\int 1df_{x\in d_X}}\right] = \Lambda(c)/r \tag{15}$$

□

No Free Lunch theorem can also be formulated in Bayesian analysis terms:

Theorem 3.2. *For a vertical $P(d|f)$, uniform $P(f)$, and a homogeneous loss L, $P(c|d) = \Lambda(c)/r$.*

Proof.

$$E_f\left[P(c|d)\right] = \frac{\int P(c|d,f)P(f|d)df}{\int df} \tag{16}$$

Using the Bayes theorem,

$$P(f) = P(f|d)P(d)/P(d|f), \tag{17}$$

and uniformity of $P(f)$:

$$E_f\left[P(c|d)\right] = \frac{\int P(c|d,f)P(f)P(d|f)/P(d)df}{\int 1df} = \alpha(d)\frac{\int P(c|d,f)P(d|f)df}{\int 1df}, \tag{18}$$

where $\alpha(d)$ is some function. Like in Theorem 3.1, the integral can be split into parts that depend on $f\left(x \in d_X\right)$ and $f\left(x \notin d_X\right)$:

$$E_f\left[P(c|d)\right] = \alpha(d)\frac{\int P\left(c|d,f\right)df_{x\notin d_X}\int P\left(d|f\right)df_{x\in d_X}}{\int 1df_{x\notin d_X}df_{x\in d_X}}. \tag{19}$$

The integral $\int P\left(d|f\right)df_{x\in d_X}$ can again be absorbed into the d-dependent constant:

$$E_f\left[P(c|d)\right] = \beta(d)\frac{\int P\left(c|d,f\right)df_{x\notin d_X}}{\int 1df_{x\notin d_X}df_{x\in d_X}} = \frac{\Lambda(c)}{r}\frac{\beta(d)}{\int 1df_{x\in d_X}}. \tag{20}$$

To obtain the value of the constant, we integrate both sides over c:

$$\int E_f\left[P(c|d)\right]dc = 1 = \frac{\beta(d)}{\int 1df_{x\in d_X}}\int\frac{\Lambda(c)}{r}dc. \tag{21}$$

From Theorem 3.1 we know that $\frac{\Lambda(c)}{r}$ is, in fact, a probability, thus $\int \frac{\Lambda(c)}{r} dc = 1$. Therefore $\frac{\beta(d)}{\int 1 df_{x \in d_X}} = 1$ as well. Substituting it back to Formula 20, we obtain:

$$E_f \left[P(c|d) \right] = \frac{\Lambda(c)}{r} \tag{22}$$

□

4 Implications

For learning theory, the No Free Lunch theorems invalidate any formal performance guarantees, that do not make a restriction on the problem class.

The NFL theorems also prove that performance on a "test" or "validation" set T, commonly used in practice, is not an agnostic tool to compare algorithms. That is if we are interested in the error for $x \notin \{d \cup T\}$, in absence of prior assumptions, the error on T is meaningless, no matter how many elements there are in T.

The Wolpert's paper begins with a quote from David Hume: "Even after the observation of the frequent conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience." In some sense, the paper is a refomulation of this thesis in mathmaticl terms. All our experiences, the training set, belong to the past. This includes any "prior knowledge"—that targets tend to be smooth, the Occams's razor, etc. The NFL theorems state, that even if some knowledge and algorithms allowed you to generalize well in the the past (your current training set), there are no formal guarantees about its behavior in the future. So, if we are to subscribe to strict empiricism and do not make any assumptions about the world, we must accept that the world is unknowable. On the other hand, if we are to claim the ability to predict the future, we must also admit that this ability is based on some arbitrary assumptions. And there are infinitely many possible assumptions and the choice can not be based on anything empirical as otherwise it would fall under the prior knowledge.

References

[1] Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical science 16.3 (2001): 199-231.

[2] Wolpert, David H. "The existence of a priori distinctions between learning algorithms." Neural Computation 8.7 (1996): 1391-1420.