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

Nikita Kazeev, based on David H. Wolpert

March 15, 2017

Abstract

The objective of machine learning is generalization – learning a relation and predicting on yet unseen data. In this paper we show that this problem can not be solved in a general, for all target relations. If there are no restrictions on the strucutre of the problem, than 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 partial rewrite of: 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 in cases 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, that show the impossibility of a machine learning algorithm, that would work across all targets. The paper does in no way argue that all algorithms are equivalent *in practice*. There are of course algorithms that perform better than random over some classes of targets (often the likes of what 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 $P(d|f)$. Select a test point $q \in \mathbf{X}, q \notin d_X$ – we are interested in the generalization power. Such selection is called off the sample (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 $c = L(y_H, y_F)$.

The results in the paper are various averages over f. f is a set of $r \times n$ real numbers, so we can write a multidimensional integral $\int A(f)df$ and average $E_f A(f) = \int df A(f) / \int df 1$. We do not evaluate the integrals explicitly, but for the clarity sake, it is worth to discuss them. $\sum_y f(x,y) = 1$. Therefore, f is a mapping from $\mathbf{X}$ to an r-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:

$$\int A(f)df = \int A(f)df \prod_{i=0}^{n} \left[\delta \left(\sum_{j=i\times r}^{(i+1)\times r-1} f_j - 1 \right) \prod_{j=i\times r}^{(i+1)\times r-1} \theta\left(f_j\right) \right] \quad (1)$$

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

$$\forall c \in \mathbb{R}, \forall y_H \in \mathbf{Y} : \sum_{y_F \in \mathbf{Y}} \delta\left[c, L(y_H, y_F)\right] = \Lambda(c). \quad (2)$$

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 losses, they are 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 $x \notin d_X$. For example, the conventional procedure, where d is created by repeatedly choosing its $\mathbf{X}$ component 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

$$P(d|f) = \prod_{i=1}^{m} \pi(d_x(i)) f(d_X(i), d_Y(i)). \quad (3)$$

3 No Free Lunch

The general idea behind the No Free Lunch theorems is calculating the uniform average over f of the distribution of classifier performace (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 determined f. Set the number of distinct elements in trainig set $m' = 4$. Let algorithm A always predict the label most popular in the training set, algorithm B the least popular. If case the nubmbers of labels are equal, the algorithms choose randomly.

We shall show that $E(c|f, m')$ 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, $c = 0$, algorithm B always misses, $c = 1$.

2. There are 5 fs with one 1. For each such f, the probability that the training set has all zeros is 0.2. For these trainig sets $c_A = 1$, $c_B = 0$. For the other 4 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 10 fs with two 1s. 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 $E_f c_A = \frac{1\times0+5\times0.2+10\times0.7}{1+5+10} = 0.5$, $E_f c_B = \frac{1\times1+5\times0.8+10\times0.3}{1+5+10} = 0.5$

3.2 Theorems

Lemma 1.

$$P(c|d,f) = \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q,d\right) P\left(y_F|q,f\right) P\left(q|d\right) \tag{4}$$

Proof.

$$c = L\left(y_H, y_F\right) \tag{5}$$

$$\begin{aligned} P\left(c|q,d,f\right) &= \sum_{y_H, y_F} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H, y_F|q,d,f\right) \\ P\left(c|d,f\right) &= \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H, y_F|q,d,f\right) P\left(q|d\right) \\ &= \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q,d\right) P\left(y_F|q,f\right) P\left(q|d\right) \end{aligned} \tag{6}$$

□

Theorem 3.1. *For homogenious loss L, the uniform average over all f of $P\left(c|d,f\right)$ equals $\Lambda\left(c\right)/r$.*

For any fixed training set, 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, any homogenious loss L, the average performance over all possible targets is a constant, that only depends on $|\mathbf{Y}|$ and L.

This result ignores the relationship between d and f – in other words, the $\mathbf{Y}$ values for train and test sets are generatied from different distributions. Thus it is not particulary 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

$$E_f\left[P(c|d,f)\right] = \sum_{y_H,y_F,q} \delta\left[c, L(y_H,y_F)\right] P(y_H|q,d)\, E_f\left[P(y_F|q,f)\right] P(q|d) \quad (7)$$

$$E_f\left[P(y_F|q,f)\right] = E_f f(q,y_F) \quad (8)$$

Because F is symmetric, the average is a constant that does not depend on q and y_F. Also,

$$\sum_{y_F} E_f\left[f(q,y_F)\right] = E_f\left[\sum_{y_F} f(q,y_F)\right] = 1, \quad (9)$$

therefore

$$E_f\left[f(q,y_F)\right] = 1/r. \quad (10)$$

Using the homogeneity property of L:

$$E_f P(c|d,f) = \sum_{y_H,q} \Lambda(c) P(y_H|q,d)\, P(q|d)/r = \Lambda(c)/r \quad (11)$$

□

Theorem 3.2. *For OTS error, a 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 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, any homogeneous loss L, the average performance over all possible targets is a constant, that only depends on $|\mathbf{Y}|$ and L.

This is a valid No Free Lunch theorem, as advertised in the beginning. If an algorithm "beats" some other, incuding the random guess, on some f's, it will necessary lose on the rest, so that the averages would be the same.

Proof.

$$P(c|f,m) = \sum_{d:|d|=m} P(c|d,f)\, P(d|f) \quad (12)$$

□

From 1:

$$P(c|d,f) = \sum_{y_H,y_F,q} \delta\left[c, L(y_H,y_F)\right] P(y_H|q,d)\, P(y_F|q,f)\, P(q|d). \quad (13)$$

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 $\mathbf{X} \setminus d_X$:

$$E_f\left[P(c|f,m)\right] = \frac{\sum_{d:|d|=m}\left[\int df_{x\notin d_X} P(c|d,f) \int df_{x\in d_X} P(d|f)\right]}{\int df_{x\notin d_X} df_{x\in d_X} 1} \quad (14)$$

Again using $P(c|d,f)$ independence from $x \in d_X$ and theorem 3.1:

$$E_{f_{x\notin d_X}}\left[P(c|d,f)\right] = E_f\left[P(c|d,f)\right] = \Lambda(c)/r \tag{15}$$

$$E_f\left[P(c|f,m)\right] = \Lambda(c)/r \frac{\sum_{d:|d|=m}\left[\int df_{x\in d_X} P(d|f)\right]}{\int df_{x\in d_X} 1} = \Lambda(c)/r \tag{16}$$

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

Theorem 3.3. *For OTS error, a vertical $P(d|f)$, uniform $P(f)$, and a homogeneous loss L, $P(c|d) = \Lambda(c)/r$.*

Proof.

$$E_f\left[P(c|d)\right] = \frac{\int df P(c|d,f)P(f|d)}{\int df} \tag{17}$$

Using the Bayes theorem,

$$P(f) = P(f|d)P(d)/P(d|f), \tag{18}$$

and unformity of $P(f)$:

$$E_f\left[P(c|d)\right] = \frac{\int df P(c|d,f)P(f)P(d|f)/P(d)}{\int df} = \alpha(d)\frac{\int df P(c|d,f)P(d|f)}{\int df 1}, \tag{19}$$

where $\alpha(d)$ is some function. Like in theorem 3.2, the integral can be split into parts that depend on $f\,(x \in d_X)$ and $f\,(x \notin d_X)$:

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

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

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

To find out the value of the constant, we integrate both sides by c:

$$\int dc E_f\left[P(c|d)\right] = 1 = \frac{\beta(d)}{\int df_{x\in d_X}} \int dc \frac{\Lambda(c)}{r}. \tag{22}$$

From theorem 3.2 we know that $\frac{\Lambda(c)}{r}$ is, in fact, probability, thus $\int dc \frac{\Lambda(c)}{r} = 1$. Therefore $\frac{\beta(d)}{\int df_{x\in d_X}} = 1$ as well. Substituting it back to 21, we obtain:

$$E_f\left[P(c|d)\right] = \frac{\Lambda(c)}{r} \tag{23}$$

□

4 Implications

No Free Lunch theorems pose a philosophical paradox. 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." 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 knowelage and algorithms allowed you to generalize well in the training set (the past), there is are no formal guarantees about its behavior in the future. So, if we are to subscribe to strict empiricism and don't 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 knowelage.

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.