

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

Nikita Kazeev, based on David H. Wolpert

March 18, 2017

Abstract

The objective of supervised 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 structure 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 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 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 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 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

36 predict on q . Let y_H be the prediction. Any classifier is completely described
 37 by its behavior, $P(y_H|q, d)$. Also sample the target distribution f at point q ,
 38 let y_F be the result. Define loss $c = L(y_H, y_F)$.

39 The results in the paper are various averages over f . f is a set of $r \times n$
 40 real numbers, so we can write a multidimensional integral $\int A(f)df$ and aver-
 41 age $E_f A(f) = \int df A(f) / \int df 1$. All integrals over targets f in this paper are
 42 implicitly restricted to the valid \mathbf{X} -conditioned distributions over \mathbf{Y} . We do not
 43 evaluate the integrals explicitly, but for the clarity sake, it is worth to discuss
 44 them.

45 $\sum_y f(x, y) = 1$. Therefore, f is a mapping from \mathbf{X} to an r -dimensional
 46 unit simplex. The integration volume F is a Cartesian product of unit sim-
 47 plices, which can be expressed using a combination of Dirac delta functions and
 48 Heaviside step functions:

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

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

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

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

54 Likelihood $P(d|f)$ determines how d was generated from f . It is *vertical*
 55 if $P(d|f)$ is independent of the values $f(x, y_F)$ for $x \notin d_X$. For example, the
 56 conventional procedure, where d is created by repeatedly choosing its \mathbf{X} compo-
 57 nent by sampling some distribution $\pi(x)$, and then choosing the associated d_Y
 58 value by sampling $f(d_X(i), y)$, results in a vertical independent and identically
 59 distributed (IID) likelihood

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

60 3 No Free Lunch

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

64 3.1 Example

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

67 Take $\mathbf{X} = \{0, 1, 2, 3, 4\}$, $\mathbf{Y} = \{0, 1\}$, a uniform sampling distribution $\pi(x)$,
 68 zero-one loss L . For clarity we will consider only determined f . Set the number
 69 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. If case the numbers 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, $c_A = 0$, algorithm B always misses, $c_B = 1$.
2. There are 5 f s with one 1. For each such f , the probability that the training set has all zeros is 0.2. For these training 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 f s 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 $E_f c_A = \frac{1 \times 0 + 5 \times 0.2 + 10 \times 0.7}{1 + 5 + 10} = 0.5$, $E_f c_B = \frac{1 \times 1 + 5 \times 0.8 + 10 \times 0.3}{1 + 5 + 10} = 0.5$

3.2 Theorems

Lemma 1.

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

Proof.

$$c = L(y_H, y_F) \quad (5)$$

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

□

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

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 homogeneous loss L , the average performance over all possible targets is a constant, that only depends on $|\mathbf{Y}|$ and L .

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

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

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

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

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

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

206 therefore

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

207 Using the homogeneity property of L :

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

208 □

209 **Theorem 3.2.** For OTS error, a vertical $P(d|f)$, and a homogeneous loss L ,
 210 the uniform average over all targets f of $P(c|f, m) = \Lambda(c)/r$

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

217 This is a valid No Free Lunch theorem, as advertised in the beginning. If an
 218 algorithm “beats” some other, including the random guess, on some f ’s, it will
 219 necessary lose on the rest, so that the averages would be the same.

Proof.

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

220 □

221 From 1:

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

222 Because we consider OTS error, $P(q|d)$ will be non-zero only for $q \notin d_X$, so
 223 $P(y_F|q, f)$ only depends on components of $f(x, y)$ that correspond to $x \notin d_X$.

124 We also know that $P(d|f)$ is vertical, so it is independent of the values
 125 $f(x, y_F)$ for $x \notin d_X$.

126 Therefore the integral can be split into two parts, over dimensions corre-
 127 sponding to d_X and $\mathbf{X} \setminus d_X$:

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

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

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

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

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

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

Proof.

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

132 Using the Bayes theorem,

$$P(f) = P(f|d)P(d)/P(d|f), \quad (18)$$

133 and uniformity of $P(f)$:

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

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

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

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

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

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

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

139 From theorem 3.2 we know that $\frac{\Lambda(c)}{r}$ is, in fact, a probability, thus $\int dc \frac{\Lambda(c)}{r} = 1$.

140 Therefore $\frac{\beta(d)}{\int df_{x \in d_X}} = 1$ as well. Substituting it back to 21, we obtain:

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

141 □

142 4 Implications

143 For learning theory, the No Free Lunch theorems invalidate any formal perfor-
144 mance guarantees, that do not make a restriction on the problem class.

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

150 No Free Lunch theorems pose a philosophical paradox. The Wolpert’s paper
151 begins with a quote from David Hume: “Even after the observation of the fre-
152 quent conjunction of objects, we have no reason to draw any inference concerning
153 any object beyond those of which we have had experience.” All our experiences,
154 the training set, belong to the past. This includes any “prior knowledge” – that
155 targets tend to be smooth, the Occams’s razor, etc. The NFL theorems state,
156 that even if some knowledge and algorithms allowed you to generalize well in
157 the training set (the past), there is are no formal guarantees about its behavior
158 in the future. So, if we are to subscribe to strict empiricism and do not make
159 any assumptions about the world, we must accept that the world is unknowable.
160 On the other hand, if we are to claim the ability to predict the future, we must
161 also admit that this ability is based on some arbitrary assumptions. And there
162 are infinitely many possible assumptions and the choice can not be based on
163 anything empirical as otherwise it would fall under the prior knowledge.

164 References

- 165 [1] Breiman, Leo. ”Statistical modeling: The two cultures (with comments and
166 a rejoinder by the author).” Statistical science 16.3 (2001): 199-231.
- 167 [2] Wolpert, David H. ”The existence of a priori distinctions between learning
168 algorithms.” Neural Computation 8.7 (1996): 1391-1420.