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 14, 2017

Abstract

Machine learning is about genralization. The perormance measured on not yet seen data is how we compare algorithms. In this paper we show that a perfect universal learner is impossible. If there are no restrictions on the strucutre of the problem, then for any two algorithms there are "as many" targets on which each outperformes the other.

1 Introduction

TODO paper link

We have a nice array of impossibility theorems. Goedel, halting, Arrow. NFL is one more.

We make practical climes. Empirically ML works) However theory-side, a universal learner is impossible.

2 Formalism

Glossary

F f value space in. 2

OTS error

$$P(q|d) = \frac{\delta(q \notin d_X)\pi(q)}{\sum_q [\delta(q \notin d_X)\pi(q)]}, \tag{1}$$

where $\delta(z) \equiv 1$ if z is true and 0 otherwise. . 3

vertical $P(d|f)$: iff $P(d|f)$ is independent of the values $f(x, y_F)$ for $x \notin d_X$. . 3

Integral

$$\int A(f)df = \int_{\mathbf{R}^{rn}} A(f)df \prod_{i=0}^{n} \left[\delta \left(\sum_{j=i\times r}^{(i+1)\times r-1} f_j - 1 \right) \prod_{j=i\times r}^{(i+1)\times r-1} \theta(f_j) \right] \tag{2}$$

3 No Free Lunch

Lemma 1.

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

- $P\left(q|d\right)$ – conditional probability of test set q given training set d
- $P\left(y_F|q,f\right)$ – conditional probability of given target sample y_F for the given the target distribution f and test set q
- $P\left(y_H|q,d\right)$ – conditional probability of given predicted sample y_H given the test and training sets. Is a function of the learning algorithm.
- $P(c|f,d)$ – conditional probability of given cost c given target f and training set d.

Proof.

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

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

□

CONSIDER example with random guessing.

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

Proof. The uniform average over all targets f of $P\left(c|f,d\right)$ equals to

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

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

Because F is symmetric, the average is a constant that does not depend on q and y_F. Also, $\sum_{y_F} E_f\left[f(y_F, q)\right] = E_f\left[\sum_{y_F} f(y_F, q)\right] = 1$, therefore $E_f\left[f(y_F, q)\right] = 1/r$.

Using the homogeneity property of L:

$$E_f P\left(c|f,d\right) = \sum_{y_H, q} \Lambda(c) P\left(y_H|q,d\right) P\left(q|d\right)/r = \Lambda(c)/r \tag{8}$$

□

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

Proof.

$$P(c|f,m) = \sum_{d:|d|=m} P(c|f,d)\, P(d|f) \tag{9}$$

□

From 3:

$$P(c|d,f) = \sum_{y_H, y_F, q} \delta\left[c, L(y_H, y_F)\right] P(y_H|q,d)\, P(y_F|q,f)\, P(q|d)\,. \tag{10}$$

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 $\mathbf{X} \setminus d_X$:

$$E_f\left[P(c|f,m)\right] = \frac{\sum_{d:|d|=m}\left[\int df_{x\notin d_X} P(c|d,f) \int df_{x\in d_X} P(d|f)\right]}{\int df_{x\notin d_X} df_{x\in d_X} 1} \tag{11}$$

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

$$E_{f_{x\notin d_X}}\left[P(c|d,f)\right] = E_f\left[P(c|d,f)\right] = \Lambda(c)/r \tag{12}$$

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

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.

$$P(c|d) = \frac{\int df P(c|d,f) P(f|d)}{\int df} \tag{14}$$

Using the Bayes theorem:

$$P(f) = P(f|d)P(d)/P(d|f) \tag{15}$$

$$P(c|d) = \frac{\int df P(c|d,f) P(f) P(d|f)/P(d)}{\int df} = \alpha(d) \int df P(c|d,f) P(d|f), \tag{16}$$

where $\alpha(d)$ is some function. □

4 Implications

[mine?] Empirism philosophy?