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 10, 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 Formalizm

3 No Free Lunch

Lemma 3.1.

$$P(c|f,d) = \sum_{y_H, y_F, q} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q,d\right) P\left(y_F|q,f\right) P\left(q|d\right) \tag{1}$$

- $P\left(q|d\right)$ – 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.

$$c = L\left(y_H, y_F\right) \tag{2}$$

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

□

CONSIDER example with random guessing.

Theorem 3.2. *For homogenious loss L, the uniform average over all f of $P\left(C|f,d\right)$ equals $\Lambda\left(c\right)/r$*

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

$$\begin{aligned} &\frac{\sum_f P\left(c|f,d\right)}{n \times r} \\ &= \sum_{y_H,y_F,f} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q,d\right) P\left(y_F|q,f\right) P\left(q|d\right) / \left(n \times r\right) \\ &= \sum_{y_H,y_F,q,f} \delta\left[c, L\left(y_H, y_F\right)\right] P\left(y_H|q,d\right) P\left(y_F|q,f\right) P\left(q|d\right) / \left(n \times r\right) \end{aligned} \tag{4}$$

The nature of f. $f(y_F, q) = P\left(y_F|q,f\right)$. Therefore $\sum_{y_F} f(y_F, q) = 1$ Therefore the set of values of f fall into a a unit simplex of dimension r. So all f's are all possible mappings $\mathbf{X} \to S_r$.

Volume integral of the coordinate corresponding to y_F over the S_r: $E\left[f\left(y_F, q\right)\right] = \int_{s_r} s_{r_{y_F}} ds$. Given that S_r is symmetrical, the integral doesn't depend on y_F.

TODO - formalism □

4 Implications

[mine?] Empirism philosophy?