CS399 Seminar: No Free Lunch Theorem for Machine Learning

Shikhar Jaiswal
1601CS44

Indian Institute of Technology Patna

14 February 2019

Bibliography

CS399 Seminar: No Free Lunch Theorem for Machine Learning

Papers

Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, **8**(7), 1341–1390.

Wolpert, D. and MacReady, W. (1997). No Free Lunch Theorems for Optimization. *IEEE Transactions on Evolutionary Computation*, **1**, 67–82.

Motivation

- **Mathematical Rigour**
  Much of modern supervised learning theory gives the impression that one can deduce something about the efficiency of a particular learning algorithm without the need for any assumptions about the target input-output relationship one is trying to learn.

CS399 Seminar: No Free Lunch Theorem for Machine Learning

Shikhar Jaiswal
1601CS44

Motivation

CS399 Seminar: No Free Lunch Theorem for Machine Learning

- So naturally one begs the question: Can one get useful, caveat-free theoretical results that link the training set and the learning algorithm to generalization error, without making assumptions concerning the target?

Introduction

CS399 Seminar: No Free Lunch Theorem for Machine Learning

- **No Free Lunch Theorems**
  David Hume (1739–1740) pointed out that 'even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience'.
  Broadly speaking, there are two no free lunch (NFL) theorems, one for supervised machine learning and one for search/optimization.

Introduction

- **No Free Lunch Theorems**
  Wolpert (1996) shows that in a noise-free scenario where the loss function is the mis-classification rate, if one is interested in off-training-set error, then there are no a priori distinctions between learning algorithms.
  In simpler words, this means that there **isn't** a master algorithm that will work the best across all problems.

Introduction

- **Off-Training-Set Error**
  The primary mathematical tool used in these papers is off-training set (OTS) generalization error, i.e., generalization error for test sets that contain no overlap with the training set.
  OTS testing is of importance because it leaves no room for memorization, and is the natural way to investigate whether one can make assumption free statements concerning an algorithm's ability to generalize.

Introduction

CS399 Seminar: No Free Lunch Theorem for Machine Learning

- **Extended Bayesian Formalism**
  Run-off-the-mill conditional probability theory which is applied under the assumption that the training and the testing phases may have different cost functions to optimize. This gives us an additional freedom to test various conditions of NFL theorem.

$$P(a|b) = \Sigma_c P(c|b) P(a|b, c)$$

Methodology

CS399 Seminar: No Free Lunch Theorem for Machine Learning

Definitions

- X be a input distribution.
- Y be a target distribution.
- $f : X \to Y$ be the objective function.
- Dataset $d^m = \{d_X^m, d_Y^m\}$ be any set of m separate pairs $(x \in X, f(x))$.
- Search algorithm α be a function that maps any d^m for any $m \in \{0, 1, ..\}$ to an $x \notin d_X^m$.

Methodology

CS399 Seminar: No Free Lunch Theorem for Machine Learning

Inferences

- By iterating a search algorithm we build successively larger datasets: $d^{m+1} = d^m \cup (\alpha(d_X^m), f[\alpha(d_X^m)])$ for all m.
- If we are given a performance measure $\Phi : d_Y^m \to \mathbb{R}$, then we can evaluate how the performance of a given search algorithm on a given objective function changes as it is run on that function.

Methodology

Derivations

$$P(\phi|\alpha, m) = \Sigma_{d_Y^m} P(d_Y^m|\alpha, m) P(\phi|d_Y^m, \alpha, m)$$

$$P(\phi|\alpha, m) = \Sigma_{d_Y^m} P(d_Y^m|\alpha, m) \delta(\phi, \Phi(d_Y^m)) \quad (1)$$

where $\delta(a, b) = 1$ if $a = b$, else 0.

This shows that the choice of search algorithm affects performance only through the term $P(d_Y^m|\alpha, m)$.

Methodology

Derivations

$$P(d_Y^m|\alpha, m) = \Sigma_f P(d_Y^m|f, \alpha, m) P(f|\alpha, m)$$

$$P(d_Y^m|\alpha, m) = \Sigma_f P(d_Y^m|f, \alpha, m) P(f) \quad (2)$$

The first term gives all the details of how the search algorithm operates, but nothing concerning the space where the algorithm is deployed. The second gives all the details of the world in which one deploys that search algorithm, but nothing concerning the search algorithm itself.

Results

Let B be any subset of the set of all objective functions, Y^X. Also let α be any search algorithm, and let Φ be any performance measure.

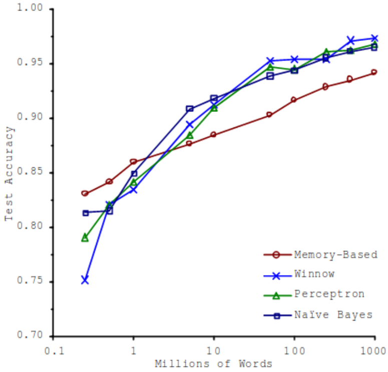
$$\Sigma_{f \in B} \mathbb{E}(\Phi | f, m, \alpha) + \Sigma_{f \in Y^X - B} \mathbb{E}(\Phi | f, m, \alpha) = constant \quad (3)$$

The constant on the right-handside depends on Φ, but is independent of both α and B.

Results

Final

Expressed differently, Eq. (3) says that $\Sigma_f \mathbb{E}(\Phi|f, m, \alpha)$ is independent of α. This is the No Free Lunch Theorem.



Criticism

No Incentive to Develop Algorithms

The foremost criticism portrayed towards this work, researchers argued that since no algorithm is expected to perform better than any other (over all set of problems), then it doesn't make sense to invest time in developing ML algorithms.

However, this notion is misleading, as this work only suggests that you need some prior knowledge of the dataset in your algorithm in order for it to work well.

Criticism

No Notion As 'All Problems'

Researchers argued that the space of all problems contains datasets that are completely random and chaotic, and nothing like the 'real world' that exhibits patterns and structure.

This is a valid piece of argument, but still doesn't allow a researcher to claim that his/her algorithm works better on all 'real world problems' as that set is ever growing.

Criticism

Existing Counterexamples

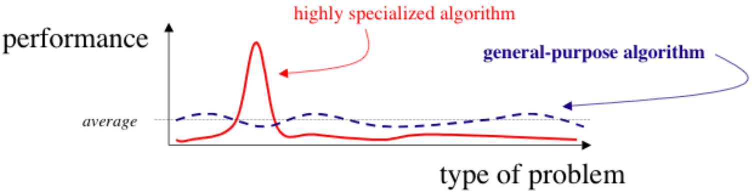
There has been an interesting counterexample, where a researcher proved that for a particular problem space, an algorithm would outperform another algorithm on 90% of the problem instances.

However analysis showed that by virtue of being exponentially worse on the remaining 10% of the problems. Its average performance abided by the NFL theorem and the two algorithms were equal when looking at mean performance, yet the hill-climber was better in 90% of the problems.

Impact

Foundational Discovery

NFL theorem is one of the most foundational pieces of work that comes under regular use while designing algorithms and still an area of active research with the two papers being cited more than 8000 times.



Spawned Many Areas of Work

The field of data analysis, which employs thousands of engineers owes its origin to the idea that investing effort in better data may have more impact than purely algorithmic work.

Also, this was a major point of consideration while the field of transfer learning was developed, where the experience gained from one algorithm is passed onto another.

Impact

Proven Empirically

Microsoft Researchers M. Banko and E. Brill in a 2001 paper (*Scaling to Very Very Large Corpora for Natural Language Disambiguation*) showed that, for the given problem, very different algorithms performed virtually the same. However, adding more examples (words) to the training set monotonically increased the accuracy of the model.

Conclusion

CS399 Seminar: No Free Lunch Theorem for Machine Learning

- Ultimately, randomly selecting a good algorithm without making any structural assumptions on the problem is similar to the proverbial "needle in the haystack".
- It will serve us well to consider the problem and the data, and act accordingly, engineering a solution with a good fit to the task at hand.
- As Peter Norvig (Director of Google Research) said "We don't have better algorithms. We just have more data."