Are hyperparameters vibes?

Nico Formánek
High Performance Computing Center (HLRS)

April 24th, 2025

Ben Recht on hyperparameters

https://www.argmin.net/p/in-defense-of-typing-monkeys

First, **hyperparameters are just vibes.**

Ben Recht on hyperparameters

https://www.argmin.net/p/in-defense-of-typing-monkeys

> *First,* **hyperparameters are just vibes.** *What are the hyperparameters in a neural network?*

Ben Recht on hyperparameters

https://www.argmin.net/p/in-defense-of-typing-monkeys

First, **hyperparameters are just vibes.** *What are the hyperparameters in a neural network?*

1. *The number of units in each layer*
2. *The number of layers*
3. *The architecture*
4. *The regularization*
5. *The weight initialization*
6. *The ADAM parameters*
7. *The input encoding*
8. *The length of the hamburger train*
9. *The random seed*
10. *The bugs in the code*

Ben Recht on hyperparameters

https://www.argmin.net/p/in-defense-of-typing-monkeys

First, **hyperparameters are just vibes.** *What are the hyperparameters in a neural network?*

1. *The number of units in each layer*
2. *The number of layers*
3. *The architecture*
4. *The regularization*
5. *The weight initialization*
6. *The ADAM parameters*
7. *The input encoding*
8. *The length of the hamburger train*
9. *The random seed*
10. *The bugs in the code*

As far as I can tell, the only parts of the neural network that are not parameters are the weights after initialization.

Ben Recht on hyperparameters

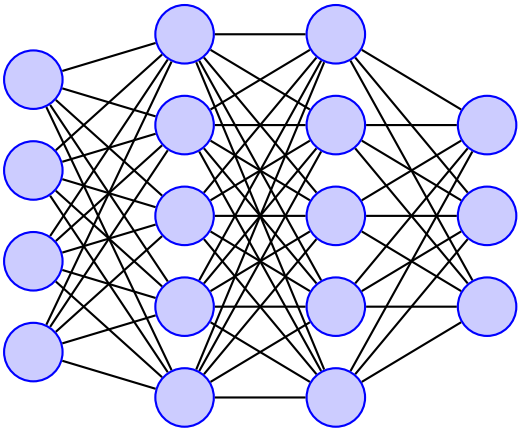
https://www.argmin.net/p/in-defense-of-typing-monkeys

First, **hyperparameters are just vibes.** *What are the hyperparameters in a neural network?*

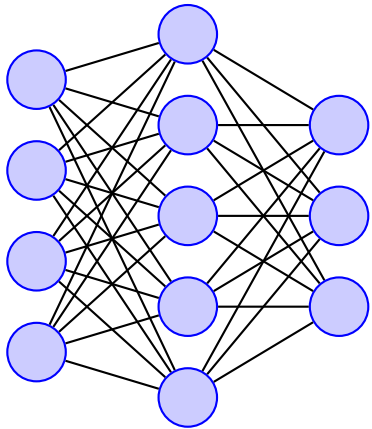
1. *The number of units in each layer*
2. *The number of layers*
3. *The architecture*
4. *The regularization*
5. *The weight initialization*
6. *The ADAM parameters*
7. *The input encoding*
8. *The length of the hamburger train*
9. *The random seed*
10. *The bugs in the code*

As far as I can tell, the only parts of the neural network that are not parameters are the weights after initialization.

The number of layers



$\phi = 2$



$\phi = 1$

Working definition

Definition (Hyperparameter)

A hyperparameter is a an implicit or explicit choice one has to make before one can run an algorithm.

Working definition

Definition (Hyperparameter)

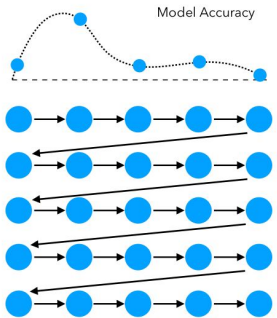
A hyperparameter is a an implicit or explicit choice one has to make before one can run an algorithm.

Can one make an implicit choice? Or is it made for us?

How are hyperparameters set in practice?

- Often baked into the algorithm (i.e. implicit) and thus *ignored*.
- Sometimes set according to *best practices*.
- Sometimes *searched* or *optimized* for.

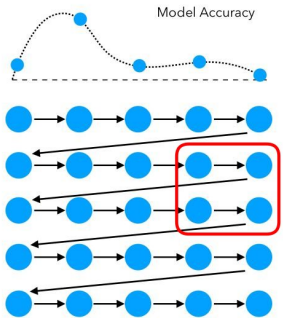
Hyperparameter search



Grid search - a type of brute force.

- Search spaces are big (typically exponential in the dimension of hyperparameters).

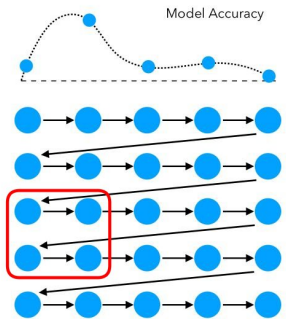
Hyperparameter search



Grid search - a type of brute force.

- Search spaces are big (typically exponential in the dimension of hyperparameters).
- Idea: restrict search space.

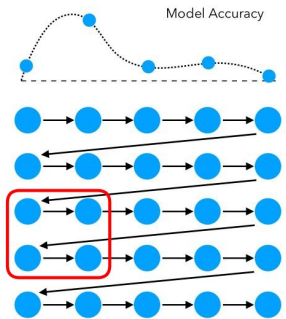
Hyperparameter search



Grid search - a type of brute force.

- Search spaces are big (typically exponential in the dimension of hyperparameters).
- Idea: restrict search space.
- How? Controlled by (hyper-)hyperparameters.

Hyperparameter search



Grid search - a type of brute force.

- Search spaces are big (typically exponential in the dimension of hyperparameters).
- Idea: restrict search space.
- How? Controlled by (hyper-)hyperparameters.
- Where to stop?

Philosophical interlude - Münchhausen trilemma

All reasoning must come to an end. According to Hans Albert (Philosopher, follower of Popper) it must end thus:

1. Infinite regress.
2. Circularity.
3. Dogmatic.

(At least according to orthodox epistemology. Obviously Albert thought he had something better.)

Justified choices

- We all want to be rational.
- Maybe there is (at least theoretically) a best way of choosing.
- Wish: independence from ways of choosing and evaluating.
- Evade the Münchhausen trilemma.

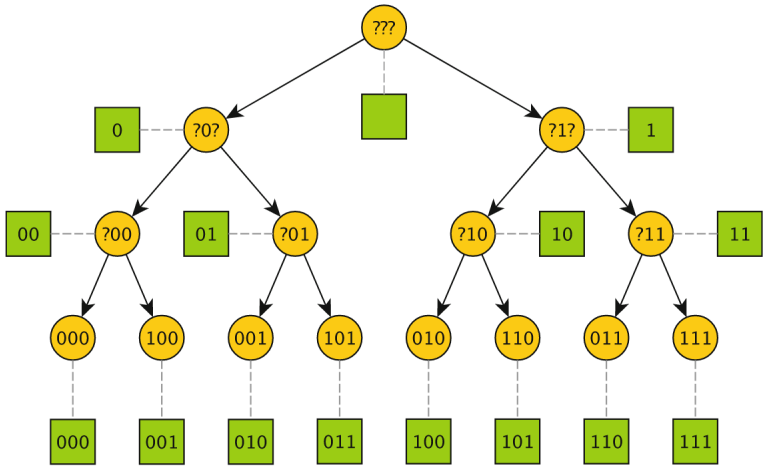
The maximally agnostic setting

- A (deterministic, non-repeating) search algorithm A: in every step select some ϕ and use f to evaluate it.
- Sometimes called black box search (because we only need $f(\phi)$).
- Grid search is an example of black box search.

The maximally agnostic setting

- A (deterministic, non-repeating) search algorithm A: in every step select some ϕ and use f to evaluate it.
- Sometimes called black box search (because we only need $f(\phi)$).
- Grid search is an example of black box search.
- One hyperparameter $\phi \in \{1, 2, 3\}$.
- A function to evaluate it $f : \{1, 2, 3\} \to \{0, 1\}$
- Some way to measure the performance of A on f.

A picture says more than a thousand words



Behaviour of one search A on all possible $f : \{1, 2, 3\} \to \{0, 1\}$

Measuring performance in a maximally agnostic way

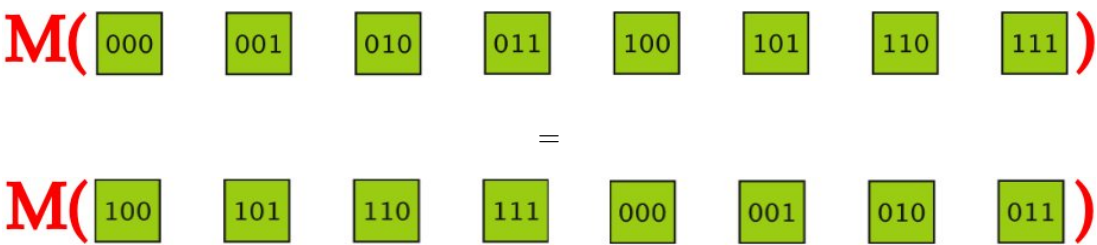
Maximally agnostic performance measure:

- Depends only implicitly on f.
- Measures aggregate performance of A on all f.
- Order of trace should not matter (Principle of insufficient reason).

Measuring performance in a maximally agnostic way

Maximally agnostic performance measure:

- Depends only implicitly on f.
- Measures aggregate performance of A on all f.
- Order of trace should not matter (Principle of insufficient reason).



NFL - a short history

In this setting one can show that, generally,

Theorem (No free lunch)

$\forall A, B, M : M(\{Tr(A, f)\}) = M(\{Tr(B, f)\})$

- First results proved in the 90s (Wolpert & Macready, Schaffer)
- Learning, optimization/search.
- Various generalizations later.
- Practical implications discussed a lot.

Some reactions to NFL

> *About half of the people in the audience to which my work was directed told me that my result was completely obvious and common knowledge–which is perfectly fair. Of course, the other half argued just as strongly that the result wasn't true.*

(Cullen Schaffer)

- Unexpected, practically relevant (Duda & Hart)
- Unexpected, practically irrelevant (Giraud-Carrier & Provost)
- Expected, practically irrelevant (Hutter)
- Expected, practically relevant (-)

Coping with NFL

Definition (Giraud-Carrier & Provost, weak)

The weak assumption of Machine Learning is that the process that presents us with learning problems, call it Ω, induces a non-uniform probability distribution, p_Ω, over the f_i's.

- The weak assumption has been justified by the *anthropic principle* (Christianini, McDermott).
- By Albert's standards this would count as circular reasoning.
- For an informed choice of A knowledge of p_Ω is necessary.
- Do we have that knowledge?

Coping with NFL

Definition (Giraud-Carrier & Provost, strong)

The strong assumption of Machine Learning is that p_Ω is **explicitly or implicitly known**, at least to a useful approximation.

Coping with NFL

Definition (Giraud-Carrier & Provost, strong)

The strong assumption of Machine Learning is that p_Ω is **explicitly or implicitly known**, at least to a useful approximation.

- This stands in stark contrast to the agnostic setting which is normally claimed in ML (von Luxburg & Schölkopf: *No assumptions on P.*)
- Is implicit knowledge sufficient to make an explicit choice?
- And how does it get justified?

McDermott on why our implicit choices are justified

The specialisation to this large, ill-defined subset has happened through intuition, trial and error, theoretical understanding of the properties of the problems of interest, and gradual matching of their properties in algorithms. Algorithms are designed according to researchers' intuitions, and formal and informal knowledge. Meanwhile, algorithms evolve since those which seem to work well are kept and varied, and those which do not are thrown away (algorithms not published or re-implemented, papers not cited, or results not replicated).

Metareasons for Metaheuristics

- The process creating (implicit) choices for hyperparameters is not completely under our control.
- It seems impossible to argue that it leads to optimal choices. (Regress, circle or dogma?)
- Hyperparameters are vibes!

Shameless advertisment

We will host a **Conference on Uncertainty** this Summer at HLRS.

CFP is still open! `https://philo.hlrs.de/cfp-sas-25-uncertainty/`

The End