Title: Hyperparameter (machine learning)

URL: https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)

PageID: 32402755

Categories: Category:Machine learning, Category:Model selection

-----

In machine learning , a hyperparameter is a parameter that can be set in order to define any configurable part of a model 's learning process. Hyperparameters can be classified as either model hyperparameters (such as the topology and size of a neural network) or algorithm hyperparameters (such as the learning rate and the batch size of an optimizer). These are named hyper parameters in contrast to parameters , which are characteristics that the model learns from the data.

Hyperparameters are not required by every model or algorithm. Some simple algorithms such as ordinary least squares regression require none. However, the LASSO algorithm, for example, adds a regularization hyperparameter to ordinary least squares which must be set before training. [1] Even models and algorithms without a strict requirement to define hyperparameters may not produce meaningful results if these are not carefully chosen. However, optimal values for hyperparameters are not always easy to predict. Some hyperparameters may have no meaningful effect, or one important variable may be conditional upon the value of another. Often a separate process of hyperparameter tuning is needed to find a suitable combination for the data and task.

As well as improving model performance, hyperparameters can be used by researchers to introduce robustness and reproducibility into their work, especially if it uses models that incorporate random number generation .

Considerations

The time required to train and test a model can depend upon the choice of its hyperparameters. [2] A hyperparameter is usually of continuous or integer type, leading to mixed-type optimization problems. [2] The existence of some hyperparameters is conditional upon the value of others, e.g. the size of each hidden layer in a neural network can be conditional upon the number of layers. [2]

Difficulty-learnable parameters

The objective function is typically non-differentiable with respect to hyperparameters. [clarification needed] As a result, in most instances, hyperparameters cannot be learned using gradient-based optimization methods (such as gradient descent), which are commonly employed to learn model parameters. These hyperparameters are those parameters describing a model representation that cannot be learned by common optimization methods, but nonetheless affect the loss function. An example would be the tolerance hyperparameter for errors in support vector machines .

Untrainable parameters

Sometimes, hyperparameters cannot be learned from the training data because they aggressively increase the capacity of a model and can push the loss function to an undesired minimum (overfitting to the data), as opposed to correctly mapping the richness of the structure in the data. For example, if we treat the degree of a polynomial equation fitting a regression model as a trainable parameter , the degree would increase until the model perfectly fit the data, yielding low training error, but poor generalization performance.

Tunability

Most performance variation can be attributed to just a few hyperparameters. [3] [2] [4] The tunability of an algorithm, hyperparameter, or interacting hyperparameters is a measure of how much performance can be gained by tuning it. [5] For an LSTM , while the learning rate followed by the network size are its most crucial hyperparameters, [6] batching and momentum have no significant effect on its performance. [7]

Although some research has advocated the use of mini-batch sizes in the thousands, other work has found the best performance with mini-batch sizes between 2 and 32. [8]

Robustness

An inherent stochasticity in learning directly implies that the empirical hyperparameter performance is not necessarily its true performance. [2] Methods that are not robust to simple changes in hyperparameters, random seeds , or even different implementations of the same algorithm cannot be integrated into mission critical control systems without significant simplification and robustification. [9]

Reinforcement learning algorithms, in particular, require measuring their performance over a large number of random seeds, and also measuring their sensitivity to choices of hyperparameters. [9] Their evaluation with a small number of random seeds does not capture performance adequately due to high variance. [9] Some reinforcement learning methods, e.g. DDPG (Deep Deterministic Policy Gradient), are more sensitive to hyperparameter choices than others. [9]

Optimization

Hyperparameter optimization finds a tuple of hyperparameters that yields an optimal model which minimizes a predefined loss function on given test data. [2] The objective function takes a tuple of hyperparameters and returns the associated loss. [2] Typically these methods are not gradient based, and instead apply concepts from derivative-free optimization or black box optimization.

Reproducibility

Apart from tuning hyperparameters, machine learning involves storing and organizing the parameters and results, and making sure they are reproducible. [10] In the absence of a robust infrastructure for this purpose, research code often evolves quickly and compromises essential aspects like bookkeeping and reproducibility . [11] Online collaboration platforms for machine learning go further by allowing scientists to automatically share, organize and discuss experiments, data, and algorithms. [12] Reproducibility can be particularly difficult for deep learning models. [13] For example, research has shown that deep learning models depend very heavily even on the random seed selection of the random number generator . [14]

See also

Hyper-heuristic

Replication crisis

References

v

t

e

History timeline

timeline

Companies

Projects

Parameter Hyperparameter

Hyperparameter

Loss functions

Regression Bias–variance tradeoff Double descent Overfitting

Bias–variance tradeoff

Double descent

Overfitting

Clustering

Gradient descent SGD Quasi-Newton method Conjugate gradient method

SGD

Quasi-Newton method

Conjugate gradient method

Backpropagation

Attention

Convolution

Normalization Batchnorm

Batchnorm

Activation Softmax Sigmoid Rectifier

Softmax

Sigmoid

Rectifier

Gating

Weight initialization

Regularization

Datasets Augmentation

Augmentation

Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann

Claude Shannon

Shun'ichi Amari

Kunihiko Fukushima

Takeo Kanade

Marvin Minsky

John McCarthy

Nathaniel Rochester

Allen Newell

Cliff Shaw

Herbert A. Simon

Oliver Selfridge

Frank Rosenblatt

Bernard Widrow

Joseph Weizenbaum

Seymour Papert

Seppo Linnainmaa

Paul Werbos

Geoffrey Hinton

John Hopfield

Jürgen Schmidhuber

Yann LeCun

Yoshua Bengio

Lotfi A. Zadeh

Stephen Grossberg

Alex Graves

James Goodnight

Andrew Ng

Fei-Fei Li

Alex Krizhevsky

Ilya Sutskever

Oriol Vinyals

Quoc V. Le

Ian Goodfellow

Demis Hassabis

David Silver

Andrej Karpathy

Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category