Title: Double descent

URL: https://en.wikipedia.org/wiki/Double_descent

PageID: 3257325

Categories: Category:Machine learning, Category:Model selection, Category:Statistical classification, Category:Statistics stubs

-----

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

Double descent in statistics and machine learning is the phenomenon where a model with a small number of parameters and a model with an extremely large number of parameters both have a small training error , but a model whose number of parameters is about the same as the number of data points used to train the model will have a much greater test error than one with a much larger number of parameters. This phenomenon has been considered surprising, as it contradicts assumptions about overfitting in classical machine learning.

History

Early observations of what would later be called double descent in specific models date back to 1989.

The term "double descent" was coined by Belkin et. al. in 2019, when the phenomenon gained popularity as a broader concept exhibited by many models. The latter development was prompted by a perceived contradiction between the conventional wisdom that too many parameters in the model result in a significant overfitting error (an extrapolation of the bias–variance tradeoff), and the empirical observations in the 2010s that some modern machine learning techniques tend to perform better with larger models.

Theoretical models

Double descent occurs in linear regression with isotropic Gaussian covariates and isotropic Gaussian noise.

A model of double descent at the thermodynamic limit has been analyzed using the replica trick , and the result has been confirmed numerically.

A number of works have suggested that double descent can be explained using the concept of effective dimension : While a network may have a large number of parameters, in practice only a subset of those parameters are relevant for generalization performance, as measured by the local Hessian curvature. This explanation is formalized through PAC -Bayes compression-based generalization bounds, which show that less complex models are expected to generalize better under a Solomonoff prior .

Empirical examples

The scaling behavior of double descent has been found to follow a broken neural scaling law functional form.

See also

Grokking (machine learning)

References

Further reading

Mikhail Belkin; Daniel Hsu; Ji Xu (2020). "Two Models of Double Descent for Weak Features" . SIAM Journal on Mathematics of Data Science . 2 (4): 1167– 1180. arXiv : 1903.07571 . doi : 10.1137/20M1336072 .

Mount, John (3 April 2024). "The m = n Machine Learning Anomaly" .

Preetum Nakkiran; Gal Kaplun; Yamini Bansal; Tristan Yang; Boaz Barak; Ilya Sutskever (29 December 2021). "Deep double descent: where bigger models and more data hurt". Journal of Statistical Mechanics: Theory and Experiment . 2021 (12). IOP Publishing Ltd and SISSA Medialab srl: 124003. arXiv : 1912.02292 . Bibcode : 2021JSMTE2021l4003N . doi : 10.1088/1742-5468/ac3a74 . S2CID 207808916 .

Song Mei; Andrea Montanari (April 2022). "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve". Communications on Pure and Applied Mathematics . 75 (4): 667– 766. arXiv : 1908.05355 . doi : 10.1002/cpa.22008 . S2CID 199668852 .

Xiangyu Chang; Yingcong Li; Samet Oymak; Christos Thrampoulidis (2021). "Provable Benefits of Overparameterization in Model Compression: From Double Descent to Pruning Neural Networks". Proceedings of the AAAI Conference on Artificial Intelligence . 35 (8). arXiv : 2012.08749 .

External links

Brent Werness; Jared Wilber. "Double Descent: Part 1: A Visual Introduction" .

Brent Werness; Jared Wilber. "Double Descent: Part 2: A Mathematical Explanation" .

Understanding "Deep Double Descent" at evhub.

v

t

e

Outline

Index

Mean Arithmetic Arithmetic-Geometric Contraharmonic Cubic Generalized/power Geometric Harmonic Heronian Heinz Lehmer

Arithmetic

Arithmetic-Geometric

Contraharmonic

Cubic

Generalized/power

Geometric

Harmonic

Heronian

Heinz

Lehmer

Median

Mode

Average absolute deviation

Coefficient of variation

Interquartile range

Percentile

Range

Standard deviation

Variance

Central limit theorem

Moments Kurtosis L-moments Skewness

Kurtosis

L-moments

Skewness

Index of dispersion

Contingency table

Frequency distribution

Grouped data

Partial correlation

Pearson product-moment correlation

Rank correlation Kendall's τ Spearman's ρ

Kendall's τ

Spearman's ρ

Scatter plot

Bar chart

Biplot

Box plot

Control chart

Correlogram

Fan chart

Forest plot

Histogram

Pie chart

Q–Q plot

Radar chart

Run chart

Scatter plot

Stem-and-leaf display

Violin plot

Effect size

Missing data

Optimal design

Population

Replication

Sample size determination

Statistic

Statistical power

Sampling Cluster Stratified

Cluster

Stratified

Opinion poll

Questionnaire

Standard error

Blocking

Factorial experiment

Interaction

Random assignment

Randomized controlled trial

Randomized experiment

Scientific control

Adaptive clinical trial

Stochastic approximation

Up-and-down designs

Cohort study

Cross-sectional study

Natural experiment

Quasi-experiment

Population

Statistic

Probability distribution

Sampling distribution Order statistic

Order statistic

Empirical distribution Density estimation

Density estimation

Statistical model Model specification L p space

Model specification

L p space

Parameter location scale shape

location

scale

shape

Parametric family Likelihood (monotone) Location–scale family Exponential family

Likelihood (monotone)

Location–scale family

Exponential family

Completeness

Sufficiency

Statistical functional Bootstrap U V

Bootstrap

U

V

Optimal decision loss function

loss function

Efficiency

Statistical distance divergence

divergence

Asymptotics

Robustness

Estimating equations Maximum likelihood Method of moments M-estimator Minimum distance

Maximum likelihood

Method of moments

M-estimator

Minimum distance

Unbiased estimators Mean-unbiased minimum-variance Rao–Blackwellization Lehmann–Scheffé theorem Median unbiased

Mean-unbiased minimum-variance Rao–Blackwellization Lehmann–Scheffé theorem

Rao–Blackwellization

Lehmann–Scheffé theorem

Median unbiased

Plug-in

Confidence interval

Pivot

Likelihood interval

Prediction interval

Tolerance interval

Resampling Bootstrap Jackknife

Bootstrap

Jackknife

1- & 2-tails

Power Uniformly most powerful test

Uniformly most powerful test

Permutation test Randomization test

Randomization test

Multiple comparisons

Likelihood-ratio

Score/Lagrange multiplier

Wald

Z -test (normal)

Student's t -test

F -test

Chi-squared

G -test

Kolmogorov–Smirnov

Anderson–Darling

Lilliefors

Jarque–Bera

Normality (Shapiro–Wilk)

Likelihood-ratio test

Model selection Cross validation AIC BIC

Cross validation

AIC

BIC

Sign Sample median

Sample median

Signed rank (Wilcoxon) Hodges–Lehmann estimator

Hodges–Lehmann estimator

Rank sum (Mann–Whitney)

Nonparametric anova 1-way (Kruskal–Wallis) 2-way (Friedman) Ordered alternative (Jonckheere–Terpstra)

1-way (Kruskal–Wallis)

2-way (Friedman)

Ordered alternative (Jonckheere–Terpstra)

Van der Waerden test

Bayesian probability prior posterior

prior

posterior

Credible interval

Bayes factor

Bayesian estimator Maximum posterior estimator

Maximum posterior estimator

Correlation

Regression analysis

Pearson product-moment

Partial correlation

Confounding variable

Coefficient of determination

Errors and residuals

Regression validation

Mixed effects models

Simultaneous equations models

Multivariate adaptive regression splines (MARS)

Simple linear regression

Ordinary least squares

General linear model

Bayesian regression

Nonlinear regression

Nonparametric

Semiparametric

Isotonic

Robust

Homoscedasticity and Heteroscedasticity

Exponential families

Logistic (Bernoulli) / Binomial / Poisson regressions

Analysis of variance (ANOVA, anova)

Analysis of covariance

Multivariate ANOVA

Degrees of freedom

Cohen's kappa

Contingency table

Graphical model

Log-linear model

McNemar's test

Cochran–Mantel–Haenszel statistics

Regression

Manova

Principal components

Canonical correlation

Discriminant analysis

Cluster analysis

Classification

Structural equation model Factor analysis

Factor analysis

Multivariate distributions Elliptical distributions Normal

Elliptical distributions Normal

Normal

Decomposition

Trend

Stationarity

Seasonal adjustment

Exponential smoothing

Cointegration

Structural break

Granger causality

Dickey–Fuller

Johansen

Q-statistic (Ljung–Box)

Durbin–Watson

Breusch–Godfrey

Autocorrelation (ACF) partial (PACF)

partial (PACF)

Cross-correlation (XCF)

ARMA model

ARIMA model (Box–Jenkins)

Autoregressive conditional heteroskedasticity (ARCH)

Vector autoregression (VAR) (Autoregressive model (AR))

Spectral density estimation

Fourier analysis

Least-squares spectral analysis

Wavelet

Whittle likelihood

Kaplan–Meier estimator (product limit)

Proportional hazards models

Accelerated failure time (AFT) model

First hitting time

Nelson–Aalen estimator

Log-rank test

Bioinformatics

Clinical trials / studies

Epidemiology

Medical statistics

Chemometrics

Methods engineering

Probabilistic design

Process / quality control

Reliability

System identification

Actuarial science

Census

Crime statistics

Demography

Econometrics

Jurimetrics

National accounts

Official statistics

Population statistics

Psychometrics

Cartography

Environmental statistics

Geographic information system

Geostatistics

Kriging

Category

Mathematics portal

Commons

WikiProject

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

This statistics -related article is a stub . You can help Wikipedia by expanding it .

v

t

e