Title: List of datasets for machine-learning research

URL: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

PageID: 49082762

Categories: Category:Datasets in machine learning

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

These datasets are used in machine learning (ML) research and have been cited in peer-reviewed academic journals . Datasets are an integral part of the field of machine learning. Major advances in this field can result from advances in learning algorithms (such as deep learning), computer hardware , and, less-intuitively, the availability of high-quality training datasets. [1] High-quality labeled training datasets for supervised and semi-supervised machine learning algorithms are usually difficult and expensive to produce because of the large amount of time needed to label the data. Although they do not need to be labeled, high-quality datasets for unsupervised learning can also be difficult and costly to produce. [2] [3] [4]

Many organizations, including governments, publish and share their datasets . The datasets are classified, based on the licenses, as Open data and Non-Open data .

The datasets from various governmental-bodies are presented in List of open government data sites . The datasets are ported on open data portals . They are made available for searching, depositing and accessing through interfaces like Open API . The datasets are made available as various sorted types and subtypes.

List of sorting used for datasets

The data portal is classified based on its type of license. The open source license based data portals are known as open data portals which are used by many government organizations and academic institutions .

List of open data portals

https://github.com/sebneu/ckan_instances/blob/master/instances.csv

https://dataverse.org/metrics

List of portals suitable for multiple types of applications

The data portal sometimes lists a wide variety of subtypes of datasets pertaining to many machine learning applications .

List of portals suitable for a specific subtype of applications

The data portals which are suitable for a specific subtype of machine learning application are listed in the subsequent sections.

Image data

Text data

These datasets consist primarily of text for tasks such as natural language processing , sentiment analysis , translation, and cluster analysis .

Reviews

News articles

Messages

Twitter and tweets

Dialogues

Legal

Other text

Categorization

citation analysis

Sound data

These datasets consist of sounds and sound features used for tasks such as speech recognition and speech synthesis .

Speech

Music

Other sounds

(WAV)

Signal data

Datasets containing electric signal information requiring some sort of signal processing for further analysis.

Electrical

Motion-tracking

Other signals

Chemical data

Datasets from physical systems.

Chemical Reactions with transition states (TS)

OpenReACT-CHON-EFH

OpenReACT-CHON-EFH (O pen Re action Dataset of A tomic C onfigura T ions comprising C , H , O and N with E nergies, F orces and H essians) is a 2025 open-access benchmark for machine-learning interatomic potentials.

RTP set – 35,087 stationary-point geometries (reactant, transition state and product) drawn from 11,961 elementary reactions, each labeled with density-functional energies, atomic forces and full Hessian matrices at the ωB97X-D/6-31G(d) level.

IRC set – 34,248 structures along 600 minimum-energy reaction paths, used to test extrapolation beyond trained stationary points.

NMS set – 62,527 off-equilibrium geometries generated by normal-mode sampling to probe model robustness under thermal perturbations.

The collection underpins the study Does Hessian Data Improve the Performance of Machine Learning Potentials? and was used to train and benchmark the machine-learning interatomic potentials reported therein. [202]

The dataset itself is distributed under a CC licence via Figshare. [203]

Physical data

Datasets from physical systems.

High-energy physics

Systems

Astronomy

Earth science

Other physical

Biological data

Datasets from biological systems.

Human

Dataset [263]

Animal

Fungi

Plant

Microbe

Drug discovery

Anomaly data

[333] [334]

[335]

(Last Updated - 2020)

Question answering data

This section includes datasets that deals with structured data.

Further details are provided in the project's GitHub repository and respective Hugging Face dataset card .

Dialog or instruction prompted data

This section includes datasets that contains multi-turn text with at least two actors, a "user" and an "agent". The user makes requests for the agent, which performs the request.

3: conversation id, utterances, vertical, scenario, instructions.

Cybersecurity

Mechanisms of Attack Domains of Attack

Software Development Hardware Design [permanent dead link] Research Concepts

2009 , 2010 2011 , 2012 , 2013 , 2014 , 2015 , 2016 , 2017 , 2018 , 2019 , 2020 , 2021 , 2022 .

Data files can also be downloaded here .

Data is also available here .

Alternate list of reports .

Climate and sustainability

Code data

Workshops

Multivariate data

Financial

Weather

Census

Transit

Internet

Games

Other multivariate

Curated repositories of datasets

As datasets come in myriad formats and can sometimes be difficult to use, there has been considerable work put into curating and standardizing the format of datasets to make them easier to use for machine learning research.

OpenML: [502] Web platform with Python, R, Java, and other APIs for downloading hundreds of machine learning datasets, evaluating algorithms on datasets, and benchmarking algorithm performance against dozens of other algorithms.

PMLB: [503] A large, curated repository of benchmark datasets for evaluating supervised machine learning algorithms. Provides classification and regression datasets in a standardized format that are accessible through a Python API.

Metatext NLP: https://metatext.io/datasets web repository maintained by community, containing nearly 1000 benchmark datasets, and counting. Provides many tasks from classification to QA, and various languages from English, Portuguese to Arabic.

Appen : Off The Shelf and Open Source Datasets hosted and maintained by the company. These biological, image, physical, question answering, signal, sound, text, and video resources number over 250 and can be applied to over 25 different use cases. [504] [505]

See also

Comparison of deep learning software

List of manual image annotation tools

List of biological databases

References

v

t

e

Differentiable programming

Information geometry

Statistical manifold

Automatic differentiation

Neuromorphic computing

Pattern recognition

Ricci calculus

Computational learning theory

Inductive bias

IPU

TPU

VPU

Memristor

SpiNNaker

TensorFlow

PyTorch

Keras

scikit-learn

Theano

JAX

Flux.jl

MindSpore

Portals Computer programming Technology

Computer programming

Technology