Title: Labeled data

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

PageID: 54033657

Categories: Category: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

Labeled data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and augments each piece of it with informative tags called judgments . For example, a data label might indicate whether a photo contains a horse or a cow, which words were uttered in an audio recording, what type of action is being performed in a video, what the topic of a news article is, what the overall sentiment of a tweet is, or whether a dot in an X-ray is a tumor.

Labels can be obtained by having humans make judgments about a given piece of unlabeled data. [1] Labeled data is significantly more expensive to obtain than the raw unlabeled data.

The quality of labeled data directly influences the performance of supervised machine learning models in operation, as these models learn from the provided labels. [2]

Crowdsourced labeled data

In 2006, Fei-Fei Li , the co-director of the Stanford Human-Centered AI Institute, initiated research to improve the artificial intelligence models and algorithms for image recognition by significantly enlarging the training data . The researchers downloaded millions of images from the World Wide Web and a team of undergraduates started to apply labels for objects to each image. In 2007, Li outsourced the data labeling work on Amazon Mechanical Turk , an online marketplace for digital piece work . The 3.2 million images that were labeled by more than 49,000 workers formed the basis for ImageNet , one of the largest hand-labeled database for outline of object recognition . [3]

Automated data labelling

After obtaining a labeled dataset, machine learning models can be applied to the data so that new unlabeled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabeled data. [4]

Challenges

Data-driven bias

Algorithmic decision-making is subject to programmer-driven bias as well as data-driven bias. Training data that relies on bias labeled data will result in prejudices and omissions in a predictive model , despite the machine learning algorithm being legitimate. The labeled data used to train a specific machine learning algorithm needs to be a statistically representative sample to not bias the results. [5] For example, in facial recognition systems underrepresented groups are subsequently often misclassified if the labeled data available to train has not been representative of the population,. In 2018, a study by Joy Buolamwini and Timnit Gebru demonstrated that two facial analysis datasets that have been used to train facial recognition algorithms, IJB-A and Adience, are composed of 79.6% and 86.2% lighter skinned humans respectively. [6]

Human error and inconsistency

Human annotators are prone to errors and biases when labeling data. This can lead to inconsistent labels and affect the quality of the data set. The inconsistency can affect the machine learning model's ability to generalize well. [7]

Domain expertise

Certain fields, such as legal document analysis or medical imaging , require annotators with specialized domain knowledge. Without the expertise, the annotations or labeled data may be inaccurate, negatively impacting the machine learning model's performance in a real-world scenario. [8]

References