Title: Leakage (machine learning)

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

PageID: 62817500

Categories: Category:Machine learning, Category:Statistical classification

-----

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

In statistics and machine learning , leakage (also known as data leakage or target leakage) is the use of information in the model training process which would not be expected to be available at prediction time, causing the predictive scores (metrics) to overestimate the model's utility when run in a production environment. [1]

Leakage is often subtle and indirect, making it hard to detect and eliminate. Leakage can cause a statistician or modeler to select a suboptimal model, which could be outperformed by a leakage-free model. [1]

Leakage modes

Leakage can occur in many steps in the machine learning process. The leakage causes can be sub-classified into two possible sources of leakage for a model: features and training examples. [1]

Feature leakage

Feature or column-wise leakage is caused by the inclusion of columns which are one of the following: a duplicate label, a proxy for the label, or the label itself. These features, known as anachronisms , will not be available when the model is used for predictions, and result in leakage if included when the model is trained. [2]

For example, including a "MonthlySalary" column when predicting "YearlySalary"; or "MinutesLate" when predicting "IsLate".

Training example leakage

Row-wise leakage is caused by improper sharing of information between rows of data. Types of row-wise leakage include:

Premature featurization ; leaking from premature featurization before Cross-validation /Train/Test split (must fit MinMax / ngrams /etc on only the train split, then transform the test set)

Duplicate rows between train/validation/test (for example, oversampling a dataset to pad its size before splitting; or, different rotations/augmentations of a single image; bootstrap sampling before splitting; or duplicating rows to up sample the minority class)

Non-independent and identically distributed random (non-IID) data Time leakage (for example, splitting a time-series dataset randomly instead of newer data in test set using a train/test split or rolling-origin cross-validation) Group leakage—not including a grouping split column (for example, Andrew Ng's group had 100k x-rays of 30k patients, meaning ~3 images per patient. The paper used random splitting instead of ensuring that all images of a patient were in the same split. Hence the model partially memorized the patients instead of learning to recognize pneumonia in chest x-rays. [3] [4])

Time leakage (for example, splitting a time-series dataset randomly instead of newer data in test set using a train/test split or rolling-origin cross-validation)

Group leakage—not including a grouping split column (for example, Andrew Ng's group had 100k x-rays of 30k patients, meaning ~3 images per patient. The paper used random splitting instead of ensuring that all images of a patient were in the same split. Hence the model partially memorized the patients instead of learning to recognize pneumonia in chest x-rays. [3] [4])

A 2023 review found data leakage to be "a widespread failure mode in machine-learning (ML)-based science", having affected at least 294 academic publications across 17 disciplines, and causing a potential reproducibility crisis . [5]

Detection

Data leakage in machine learning can be detected through various methods, focusing on performance analysis, feature examination, data auditing, and model behavior analysis. Performance-wise, unusually high accuracy or significant discrepancies between training and test results often indicate leakage. [6] Inconsistent cross-validation outcomes may also signal issues.

Feature examination involves scrutinizing feature importance rankings and ensuring temporal integrity in time series data. A thorough audit of the data pipeline is crucial, reviewing pre-processing steps, feature engineering, and data splitting processes. [7] Detecting duplicate entries across dataset splits is also important.

For language models, the Min-K% method can detect the presence of data in a pretraining dataset. It presents a sentence suspected to be present in the pretraining dataset, and computes the log-likelihood of each token, then compute the average of the lowest K of these. If this exceeds a threshold, then the sentence is likely present. [8] [9] This method is improved by comparing against a baseline of the mean and variance. [10]

Analyzing model behavior can reveal leakage. Models relying heavily on counter-intuitive features or showing unexpected prediction patterns warrant investigation. Performance degradation over time when tested on new data may suggest earlier inflated metrics due to leakage.

Advanced techniques include backward feature elimination, where suspicious features are temporarily removed to observe performance changes. Using a separate hold-out dataset for final validation before deployment is advisable. [7]

See also

AutoML

Concept drift (where the structure of the system being studied evolves over time, invalidating the model)

Overfitting

Resampling (statistics)

Supervised learning

Training, validation, and test sets

References