Title: Vector database

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

PageID: 74020014

Categories: Category:Machine learning, Category:Types of databases

---

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

A vector database , vector store or vector search engine is a database that uses the vector space model to store vectors (fixed-length lists of numbers) along with other data items. Vector databases typically implement one or more approximate nearest neighbor algorithms, [1] [2] so that one can search the database with a query vector to retrieve the closest matching database records.

Vectors are mathematical representations of data in a high-dimensional space. In this space, each dimension corresponds to a feature of the data, with the number of dimensions ranging from a few hundred to tens of thousands, depending on the complexity of the data being represented. A vector's position in this space represents its characteristics. Words, phrases, or entire documents, as well as images, audio, and other types of data, can all be vectorized. [3]

These feature vectors may be computed from the raw data using machine learning methods such as feature extraction algorithms, word embeddings [4] or deep learning networks. The goal is that semantically similar data items receive feature vectors close to each other.

Vector databases can be used for similarity search , semantic search , multi-modal search , recommendations engines , large language models (LLMs), object detection , etc. [3]

Vector databases are also often used to implement retrieval-augmented generation (RAG), a method to improve domain-specific responses of large language models. The retrieval component of a RAG can be any search system, but is most often implemented as a vector database. Text documents describing the domain of interest are collected, and for each document or document section, a feature vector (known as an " embedding ") is computed, typically using a deep learning network, and stored in a vector database. Given a user prompt, the feature vector of the prompt is computed, and the database is queried to retrieve the most relevant documents. These are then automatically added into the context window of the large language model, and the large language model proceeds to create a response to the prompt given this context. [5]

Techniques

The most important techniques for similarity search on high-dimensional vectors include:

Hierarchical Navigable Small World (HNSW) graphs

Locality-sensitive Hashing (LSH) and Sketching

Product Quantization (PQ)

Inverted Files

and combinations of these techniques. [citation needed]

In recent benchmarks, HNSW-based implementations have been among the best performers. [6] [7] Conferences such as the International Conference on Similarity Search and Applications, SISAP and the Conference on Neural Information Processing Systems (NeurIPS) host competitions on vector search in large databases.

Implementations

See also

Curse of dimensionality – Difficulties arising when analyzing data with many aspects ("dimensions")

Machine learning – Study of algorithms that improve automatically through experience

Nearest neighbor search – Optimization problem in computer science

Recommender system – System to predict users' preferences

References

External links

Sawers, Paul (2024-04-20). "Why vector databases are having a moment as the AI hype cycle peaks" . TechCrunch . Retrieved 2024-04-23 .