Title: Embedding (machine learning)

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

PageID: 79351893

Categories: Category:Artificial neural networks, Category:Machine learning, Category:Natural language processing

-----

Embedding in machine learning refers to a representation learning technique that maps complex, high-dimensional data into a lower-dimensional vector space of numerical vectors.

## Technique

It also denotes the resulting representation, where meaningful patterns or relationships are preserved. As a technique, it learns these vectors from data like words, images, or user interactions, differing from manually designed methods such as one-hot encoding . This process reduces complexity and captures key features without needing prior knowledge of the domain.

## Similarity

In natural language processing , words or concepts may be represented as feature vectors , where similar concepts are mapped to nearby vectors. The resulting embeddings vary by type, including word embeddings for text (e.g., Word2Vec ), image embeddings for visual data, and knowledge graph embeddings for knowledge graphs , each tailored to tasks like NLP, computer vision , or recommendation systems . This dual role enhances model efficiency and accuracy by automating feature extraction and revealing latent similarities across diverse applications.

To measure the distance between two embeddings, a similarity measure can be used to find the overall similarity of the concepts represented by the embeddings. If the vectors are normalized to have a magnitude of 1, then the similarity measures are proportional to cos ■ ( θ a b ) {\displaystyle \cos \left(\theta _{ab}\right)} .

The cosine similarity disregards the magnitude of the vector when determining similarity, so it is less biased towards training data that appears very frequently. The dot product includes the magnitude inherently, so it will tend to value more popular data. Generally, for high-dimensional vector spaces, vectors tend to converge in distance, so Euclidean distance becomes less reliable for large embedding vectors.

## See also

Latent space

Feature extraction

Dimensionality reduction

Word embedding

Neural network

Reinforcement learning

## References