# autoencoder

## Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

## Abstract

An autoencoder is a machine learning (ML) method that learns an encoder map $h : \mathbf{x} \mapsto \mathbf{z}$ and a decoder map $h^* : \mathbf{z} \mapsto \widehat{\mathbf{x}}$ jointly, so that the reconstruction $h^*(h(\mathbf{x}))$ is close to the original feature vector $\mathbf{x}$. The code $\mathbf{z}$ typically has fewer entries than $\mathbf{x}$, which makes an autoencoder a dimensionality reduction and feature learning method. Training uses empirical risk minimization (ERM) with a loss that measures the reconstruction error, and no labels are needed, so it is a form of self-supervised learning. With linear models for encoder and decoder, the learned code space coincides with that of principal component analysis (PCA); artificial neural networks (ANNs) as encoder and decoder allow nonlinear codes.

## Definition

An autoencoder is a machine learning (ML) method that simultaneously learns an encoder map $h \in \mathcal{H}$ and a decoder map $h^* \in \mathcal{H}^*$. Different autoencoders use different models $\mathcal{H}, \mathcal{H}^*$, e.g., artificial neural networks (ANNs) with different architectures. The special case of an autoencoder using (vector-valued) linear models for $\mathcal{H}, \mathcal{H}^*$ results in principal component analysis (PCA).

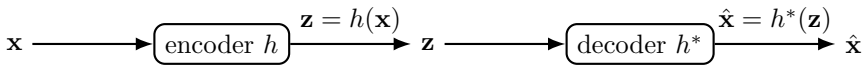


Figure 1: Autoencoder with an encoder $h$ mapping $\mathbf{x} \mapsto \mathbf{z}$ and a decoder $h^*$ mapping $\mathbf{z} \mapsto \hat{\mathbf{x}}$.

The training of the encoder and decoder can be implemented via empirical risk minimization (ERM) using a loss that measures the deviation of the reconstructed feature vector $h^*\big(h\big(\mathbf{x}\big)\big)$ from the original feature vector $\mathbf{x}$.

## Cross References

encoder, feature learning, dimensionality reduction

## References

1.