

machine learning pipeline (ML pipeline)

Authors

Alexander Jung^{*}, Department of Computer Science, Aalto University, Espoo, Finland

^{*}Corresponding author: alex.jung@aalto.fi

Abstract

A machine learning pipeline (ML pipeline) is the composition of the functions by which a machine learning system (ML system) turns raw data into predictions. Its stages typically are data preprocessing, feature learning or hand-designed feature extraction, model training, and inference on new data points. Each stage is a function whose output is the input of the next, so the whole pipeline is one map from raw data to predictions and is applied identically during training and in deployment. Fitting stages such as feature scaling inside the pipeline, on the training set only, prevents data leakage from the validation set or test set.

Definition

The term ML pipeline refers to a composition (i.e., concatenation) of several functions within a machine learning system (ML system). The individual functions include data preprocessing, feature learning, model training, and inference. By combining them, an ML system turns raw data into predictions (Hapke and Nelson, 2020).

Cross References

function, ML system

References

1. Hapke and Nelson (2020). Building Machine Learning Pipelines: Automating Model Life Cycles with TensorFlow. O'Reilly Media, Sebastopol, CA, USA.