

independent and identically distributed assumption (i.i.d. assumption)

Authors

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

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

Abstract

The independent and identically distributed assumption (i.i.d. assumption) is the probabilistic model that represents the data points of a dataset as realizations of independent and identically distributed (i.i.d.) random variables (RVs): each data point is drawn from the same probability distribution, independently of the others. For a fixed hypothesis, the average loss on a training set then has expectation equal to the risk, the expected loss on a new data point, which is what justifies empirical risk minimization (ERM) and the use of a validation set to estimate the loss outside the training set. The assumption fails under distribution shift, when the data points met after training follow a different probability distribution, and for data points recorded over time whose consecutive values are dependent.

Definition

The independent and identically distributed (i.i.d.) assumption is a widely used probabilistic model for the generation of data points. In particular, data points are represented as i.i.d. random variables (RVs).

Cross References

i.i.d., probabilistic model, data point, RV

References

- 1.