# test set


## Authors

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



## Abstract

A test set is a dataset of data points used neither for training a model nor for choosing between candidate models based on a validation set. For a hypothesis learned and selected without reference to the test set, the average loss on it estimates the risk under the independent and identically distributed assumption (i.i.d. assumption). This distinguishes it from the validation set, whose average loss flows back into the choice of the hypothesis and is therefore optimistic for the chosen candidate. The estimate stays valid only as long as no design choice depends on the test set; repeated evaluation with selection turns it into a validation set, and overlap with the training set constitutes test set contamination.


## Definition

A straight line and a degree-two polynomial have both been fitted to historic days of weather recordings, and the line was selected because it incurred the smaller average loss on days held back from the fit (model selection via a validation set). Quoting that same average as the line's expected error on future days would understate the error: the line was selected for scoring well on exactly those held-back days. The days reserved to answer the final question — how large a loss to expect on a new day — form the test set.

A test set $\mathcal{D}^{(\text{test})}$ is a dataset of data points used neither for training a model, e.g., via empirical risk minimization (ERM) on the training set $\mathcal{D}^{(\text{train})}$, nor for choosing between candidate models based on a validation set $\mathcal{D}^{(\text{val})}$ (Fig. 1). For a hypothesis $\widehat{h}$ that was learned and selected without reference to $\mathcal{D}^{(\text{test})}$, the test error is the average loss

$$\frac{1}{|\mathcal{D}^{(\text{test})}|} \sum_{\mathbf{z} \in \mathcal{D}^{(\text{test})}} L\left(\mathbf{z}, \widehat{h}\right).$$

It estimates the risk of $\widehat{h}$ under the independent and identically distributed assumption (i.i.d. assumption) (Hastie et al., 2009). The distinction from the validation set is the direction of information flow: the validation error flows back into the choice of $\widehat{h}$, so it is an optimistic estimate for the chosen candidate; the test error is computed once, after all choices are frozen, and flows back into nothing.

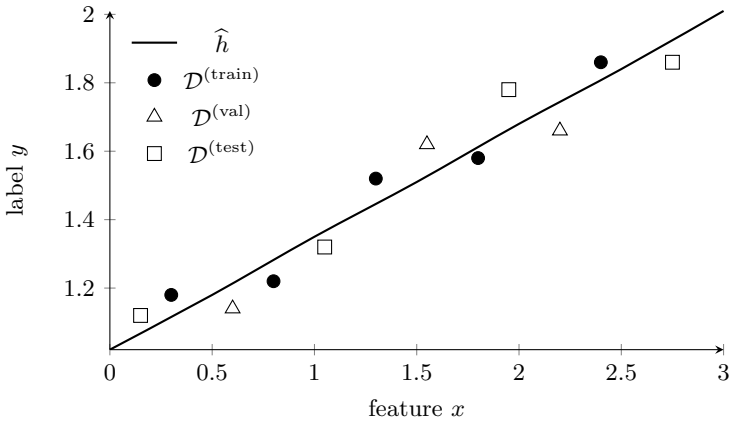


Figure 1: Three roles for data points. The filled circles form the training set on which $\widehat{h}$ is learned, the open triangles the validation set used to select among candidate models, and the open squares the test set: it enters neither training nor model selection, so the average loss on it estimates the risk of $\widehat{h}$

The estimate stays valid only as long as the test set is used this way. Evaluating many candidate models on $\mathcal{D}^{(\text{test})}$ and keeping the best turns the test set into a validation set, and its error into a validation error; data points of $\mathcal{D}^{(\text{test})}$ that enter the training set constitute test set contamination. In an ML competition, the labels of the test set are therefore withheld from the participants: no design choice of theirs can then depend on the test data, and the reported error keeps its meaning as a risk estimate.

## Cross References

training set, validation set, model selection, risk, generalization, test set contamination, data leakage, ERM

## References

1. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.