validation error

Author and editors

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

Abstract

The validation error E_v of a hypothesis $\widehat{h}$ is its average loss over a validation set $\mathcal{D}^{(\mathrm{val})}$, computed with $\widehat{h}$ fixed. Because no data point of $\mathcal{D}^{(\mathrm{val})}$ entered the training, E_v estimates the risk of $\widehat{h}$, while the training error understates it. For data points that are realizations of independent and identically distributed (i.i.d.) random variables (RVs), E_v is itself the realization of an RV whose expectation is the risk and whose fluctuation shrinks as $\mathcal{D}^{(\mathrm{val})}$ grows. A training error far below the validation error signals overfitting, and comparing validation errors across candidate models is model selection.

Definition

A straight line fitted to forty days of weather recordings predicts each day's maximum daytime temperature from its morning minimum. Applied to twenty held-back days, the line misses each day's recorded maximum by some amount. The squared error loss squares each miss, and the average of the twenty squared misses is the validation error of the line. It is this number, not the average miss on the forty days the line was fitted to, that says whether the line is of any use on a day it has not seen.

The validation error E_v of a hypothesis $\widehat{h}$ is its average loss over a validation set $\mathcal{D}^{(\mathrm{val})}$,

$$E_v = \frac{1}{|\mathcal{D}^{(\mathrm{val})}|} \sum_{\mathbf{z} \in \mathcal{D}^{(\mathrm{val})}} L\left(\mathbf{z}, \widehat{h}\right),$$

computed with $\widehat{h}$ fixed. Because no data point of $\mathcal{D}^{(\mathrm{val})}$ entered the training, E_v estimates the risk of $\widehat{h}$ (Hastie et al., 2009, Sect. 7.2); the training error, computed over the training set whose average loss empirical risk minimization (ERM) minimized, understates it. For the line of the opening example the training error is 2.45, the validation error 2.87 and the risk 2.61.

The validation error is an average of $|\mathcal{D}^{(\mathrm{val})}|$ individual losses. For data points that are realizations of independent and identically distributed (i.i.d.) random variables (RVs), E_v is itself the realization of an RV whose expectation is the risk of $\widehat{h}$, since each summand has that expectation. Its variance is the variance of a single loss divided by $|\mathcal{D}^{(\mathrm{val})}|$, so the fluctuation around the risk shrinks as $\mathcal{D}^{(\mathrm{val})}$ grows. Fig. 1 measures the fluctuation for one fixed line: recomputed over 300 independent draws of $\mathcal{D}^{(\mathrm{val})}$ for each size, the validation error scatters around the risk, with a standard deviation that falls from 1.4 at $|\mathcal{D}^{(\mathrm{val})}| = 5$ to 0.3 at $|\mathcal{D}^{(\mathrm{val})}| = 80$.

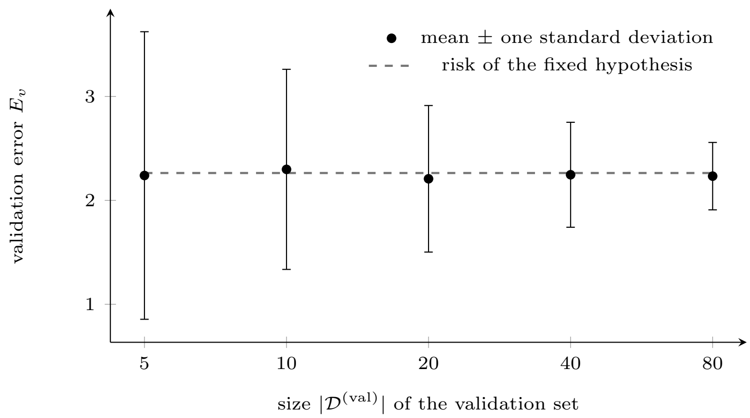

Figure 1: The validation error of one fixed hypothesis, a line fitted to 80 days of weather recordings, recomputed over 300 independent draws of the validation set for each size $|\mathcal{D}^{(\mathrm{val})}|$. Each bar spans one standard deviation either side of the mean. The mean sits on the risk (dashed) for every size, while the spread shrinks as $|\mathcal{D}^{(\mathrm{val})}|$ grows. Data generated by `pythondemos/valerr.py`

Read beside the training error, the validation error diagnoses a machine learning (ML) method: a training error far below the validation error signals overfitting (Hastie et al., 2009, Sect. 7.2). Comparing the validation errors of several candidate models is model selection (Bishop, 2006, Sect. 1.3); a validation set that is reused many times for this purpose is itself overfitted, which is why a test set is kept apart for the final assessment (Bishop, 2006, Sect. 1.3; Hastie et al., 2009, Sect. 7.2).

Cross References

validation set, validation, training set, test set, training error, loss, risk, hypothesis, ERM, overfitting, model selection

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

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