validation set

Authors

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

Abstract

A validation set consists of data points which have not been used for model training. For a learned hypothesis, the average loss on the validation set, the validation error, indicates how well the hypothesis predicts the labels of data points outside the training set. The validation error is used for model selection: the hypotheses learned by different machine learning (ML) methods are compared, and the one with the smallest validation error is selected. The selected hypothesis depends on the validation set, so assessing it requires a test set whose data points entered neither training nor selection. The size of the validation set determines how reliable the validation error is, and Hoeffding's inequality prescribes the size needed for the risk to lie, with high probability, within a given uncertainty band around the measured validation error. When data points are scarce, k-fold cross-validation (k-fold CV) averages the validation errors obtained from using different folds of the dataset as validation set.

Definition

Consider three days of weather recordings: for each day, the morning minimum temperature is the feature and the maximum daytime temperature is the label (Jung, 2022, Ch. 1). Fig. 1 shows a training set $\mathcal{D}^{(\text{train})}$ of three such days along with two hypotheses learned from it: the straight line $\widehat{h}^{(1)}$ delivered by empirical risk minimization (ERM) on the linear model, and a degree-two polynomial $\widehat{h}^{(2)}$ that passes through all three data points and therefore attains a training error of zero. Three further days, marked by open triangles, were held back from model training: they form a validation set, on which the loss of each learned hypothesis can be evaluated.

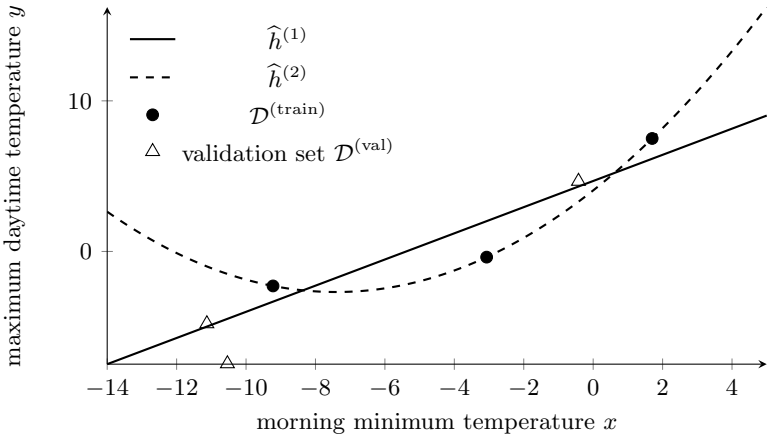

Figure 1: Data points representing daily weather conditions in Finland: each data point is a day, with the morning minimum temperature as its feature and the maximum daytime temperature as its label. The three filled circles form the training set $\mathcal{D}^{(\text{train})}$; both curves are hypotheses learned from $\mathcal{D}^{(\text{train})}$. The three open triangles were held back from model training and form a validation set $\mathcal{D}^{(\text{val})}$: the average loss on them estimates how well each hypothesis predicts the labels of data points that were not used to learn it. Data generated by `pythondemos/valset.py`

A validation set $\mathcal{D}^{(\text{val})}$ consists of data points which have not been used for model training. For a hypothesis $\widehat{h}$ learned from the training set, the resulting average loss on the validation set,

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

is the validation error: it indicates how well $\widehat{h}$ predicts the labels of data points outside the training set (Jung, 2022, Sect. 6.2). In Fig. 1, the polynomial $\widehat{h}^{(2)}$ attains a training error of zero but a validation error of 17.3, while the line $\widehat{h}^{(1)}$ has a training error of 2.9 and a validation error of 3.0.

The validation error is used for model selection: the hypotheses learned by different machine learning (ML) methods are compared, and the one with the smallest validation error is selected (Jung, 2022, Sect. 6.3). Fig. 2 shows the selection between the two learned hypotheses of the weather example: the line is selected, since $3.0 < 17.3$. The selected hypothesis depends on $\mathcal{D}^{(\text{val})}$: its validation error is the smallest of the compared validation errors, so the guarantee for a single fixed hypothesis no longer applies, and a valid bound must hold uniformly over all compared candidates, growing with their number (Shalev-Shwartz and Ben-David, 2014, Sect. 11.2). Assessing the selected hypothesis therefore requires a third dataset whose data points entered neither training

nor selection: the test set (Hastie et al., 2009, Sect. 7.2). In Fig. 2, the three days of the test set, marked by open squares, yield a test error of 1.3 for the selected line.

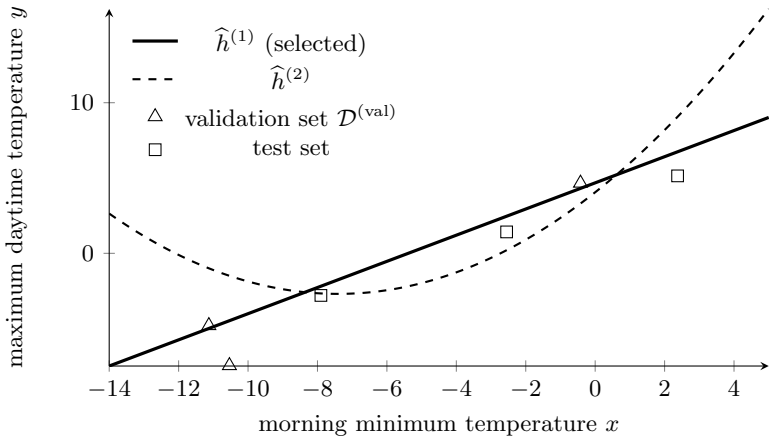

Figure 2: Model selection with a validation set. The two curves are the hypotheses of Fig. 1; the validation error over the open triangles selects the line (3.0 against 17.3 for the polynomial), drawn thick. The selected hypothesis depends on $\mathcal{D}^{(\mathrm{val})}$, so a third set of days that entered neither training nor selection, the test set (open squares), assesses it: the selected line incurs a test error of 1.3. Data generated by `pythondemos/valset.py`

The size $|\mathcal{D}^{(\mathrm{val})}|$ determines how well the validation error reflects the overall performance of a learned hypothesis: the validation error is an average of $|\mathcal{D}^{(\mathrm{val})}|$ terms, and a validation set that is too small results in an unreliable validation error (Jung, 2022, Sect. 6.2.1). A concentration inequality quantifies the reliability. For a loss with values in $[0, 1]$ and a fixed hypothesis $\widehat{h}$, Hoeffding's inequality (Shalev-Shwartz and Ben-David, 2014, Thm. 11.1) guarantees that, with probability at least $1 - \delta$, the risk of $\widehat{h}$ lies within an uncertainty band of half-width Δ around the measured validation error, provided that

$$|\mathcal{D}^{(\mathrm{val})}| \geq \frac{\ln(2/\delta)}{2\Delta^2}.$$

Fig. 3 shows this lower bound as a function of the half-width Δ for three values of δ. Narrowing the band is costly: halving Δ quadruples the required size. To have the validation error within $\Delta = 0.1$ of the risk with $\delta = 0.05$, it suffices to hold back $|\mathcal{D}^{(\mathrm{val})}| \geq 185$ data points, regardless of the hypothesis space and of the size of $\mathcal{D}^{(\mathrm{train})}$.

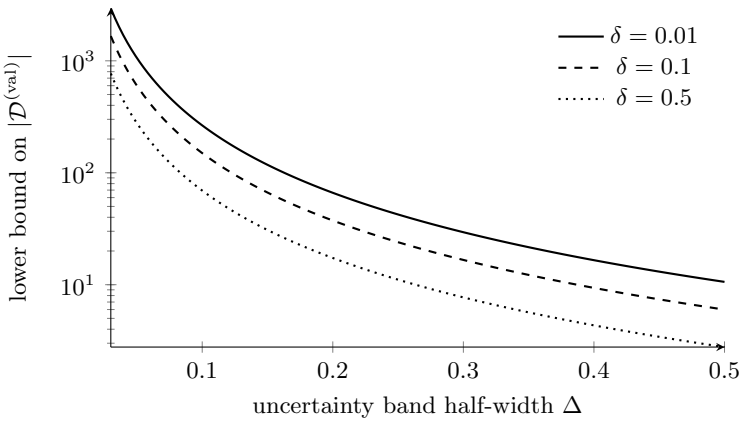

Figure 3: The lower bound $\ln(2/\delta)/(2\Delta^2)$ on the size $|\mathcal{D}^{(\text{val})}|$ as a function of the half-width Δ of the uncertainty band around the measured validation error, for three values of the confidence parameter δ. The vertical axis is logarithmic: halving Δ quadruples the required size, while tightening δ from 0.5 to 0.01 raises it by a factor below four

When data points are scarce, holding back a large validation set leaves few data points for the training set. As a remedy, k-fold cross-validation (k-fold CV) divides the dataset into k folds and uses each fold in turn as validation set: every fold yields a noisy validation error, an average over few data points, and averaging the k per-fold validation errors produces a more reliable estimate of the expected loss (Jung, 2022, Sect. 6.2.2; Stone, 1974). In a complete workflow, the available dataset is thus split three ways: a training set to learn hypotheses, a validation set to select among them, and a test set, used only once, to assess the selected hypothesis.

Cross References

validation, training set, test set, data point, risk, hypothesis, loss, validation error, training error, k-fold CV, LOO-CV, model selection

References

1. Jung (2022). Machine Learning: The Basics. Springer Nature, Singapore, Singapore.

2. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

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

New York, NY, USA.

4. Stone (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), pp. 111–147.