training 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 training error E_t of a hypothesis $\widehat{h}$ is its average loss over the training set, the empirical risk of $\widehat{h}$ on the data points used for training. For a hypothesis delivered by empirical risk minimization (ERM) it is the smallest empirical risk that any hypothesis of the hypothesis space attains. Because the same data points selected $\widehat{h}$, the training error is optimistic: it lies below the risk on average, by a gap that grows with the size of the hypothesis space, so a hypothesis with zero training error may be worthless on new data points. Read beside the validation error, a training error far below it signals overfitting and a large training error signals underfitting.

Definition

A straight line is fitted to forty days of weather recordings to predict each day's maximum daytime temperature from its morning minimum. On each of these days the line misses the recorded maximum by some amount, between 0.0 and 4.0 degrees. The squared error loss squares each miss, and the mean of the forty squared misses, 2.45, is the training error of the line. No other line attains a smaller value on these days: the line was chosen to minimize exactly this number. The numbers are computed by `pythondemos/trainerr.py`.

The training error E_t of a hypothesis $\widehat{h}$ is its average loss over the training set $\mathcal{D}^{(\text{train})} = \{\mathbf{z}^{(r)}\}_{r=1}^{m}$,

$$E_t = \frac{1}{m} \sum_{r=1}^{m} L\left(\mathbf{z}^{(r)}, \widehat{h}\right),$$

i.e., the empirical risk of $\widehat{h}$ on the data points used for training. For a hypothesis delivered by empirical risk minimization (ERM), the training error is the smallest empirical risk that any hypothesis of the hypothesis space attains on $\mathcal{D}^{(\text{train})}$, since ERM minimizes precisely this average. With the 0/1 loss, the training error of a classifier is the fraction of training data points it misclassifies; with the logistic loss, it is the negative average log-likelihood of the training labels (Hastie et al., 2009, Sect. 7.2).

The training error is computed on the same data points that selected $\widehat{h}$, and this is what limits its value as a measure of quality. The goal of a machine

learning (ML) method is a small loss on data points it has not seen, formalized as a small risk for data points that are realizations of independent and identically distributed (i.i.d.) random variables (RVs). Because $\widehat{h}$ was picked to make the average on $\mathcal{D}^{(\text{train})}$ small, the training error is optimistic: on average over the draw of the training set it lies below the risk. The more the hypothesis space offers to choose from, the larger this gap (Hastie et al., 2009, Sect. 7.4). A hypothesis with zero training error may therefore be worthless on new data points. The validation error, computed on data points that did not enter the training, has no such optimism and estimates the risk; the difference between the two is the generalization gap.

Fig. 1 shows both errors for polynomials of degree 0 to 12 fitted by ERM to the forty days. The hypothesis spaces are nested, a polynomial of degree d being one of degree $d + 1$ with a zero coefficient, so the training error never increases with the degree: it falls from 29.7 for the constant to 2.45 for the line and 1.84 for degree 12. The risk, computed on 200000 fresh days, is smallest for the line, 2.61. It reaches 24.8 at degree 12, where the polynomial follows the noise of the forty recordings.

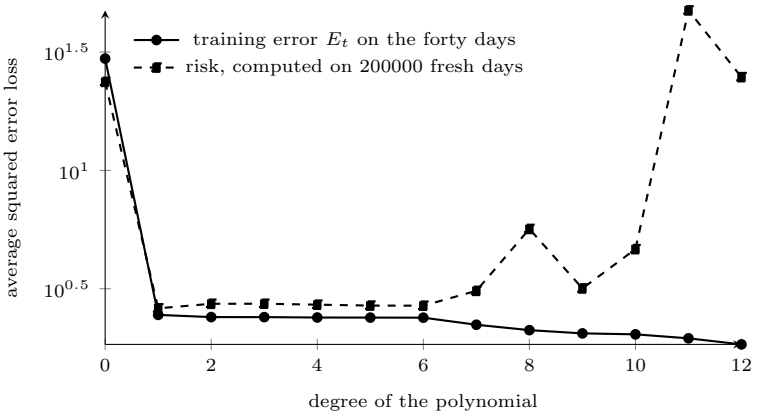

Figure 1: Training error and risk of the polynomial of each degree fitted by ERM to forty days of weather recordings. The training error never increases with the degree, since each hypothesis space contains the previous one, while the risk is smallest for the straight line and grows by an order of magnitude beyond degree 10. Data generated by `pythondemos/trainerr.py`

Read beside the validation error, the training error diagnoses an ML method. A training error far below the validation error signals overfitting (Hastie et al., 2009, Sect. 7.2). A training error that is itself large signals underfitting: a hypothesis space too small, or an optimizer that stopped short of the minimum. Fig. 1 also shows the trade that regularization makes: restricting the hypothesis space from degree 12 to the line raises the training error from 1.84 to 2.45 and lowers the risk from 24.8 to 2.61.

Cross References

training set, empirical risk, ERM, loss, validation error, risk, generalization gap, overfitting, underfitting, hypothesis

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.