

model selection

Author and editors

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

Editors: Konstantina Oliumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Model selection chooses among several trained hypotheses by comparing the validation error each of them incurs on a validation set. Each candidate is the output of empirical risk minimization (ERM) run with a different hypothesis space or a different hyperparameter setting, such as the learning rate of a gradient-based method or the depth of a decision tree. The training error cannot serve as the criterion: it falls as the hypothesis space grows, while the risk eventually rises again. k -fold cross-validation (k -fold CV) supplies the validation error when the training set is too small to spare a validation set. A test set held out from every candidate and from the comparison itself is what remains for estimating the risk of the selected hypothesis.

Definition

Two forecasts of a day's maximum daytime temperature y from its morning minimum x have been fitted to the same six days of weather recordings: a straight line and a degree-five polynomial (Fig. 1). Each of them is a hypothesis that empirical risk minimization (ERM) on the training set $\mathcal{D}^{(\text{train})}$ of those six days delivers, the line from a linear model and the polynomial from the larger hypothesis space of polynomial regression. The polynomial passes through every day of $\mathcal{D}^{(\text{train})}$, so its training error, the average squared error loss over $\mathcal{D}^{(\text{train})}$, is zero; the line misses each of them and its training error is 0.47. The training error therefore ranks the polynomial first, and following that ranking would pick the candidate that predicts worst on a day it has not seen. Model selection ranks instead by the validation error on a validation set $\mathcal{D}^{(\text{val})}$ that ERM never saw, which reverses the order, and it leaves a test set $\mathcal{D}^{(\text{test})}$ untouched for estimating the risk of the winner. The twenty days held back from the fit are split fourteen to six for these two roles, so the validation errors below are averages over fourteen days, not over all twenty.

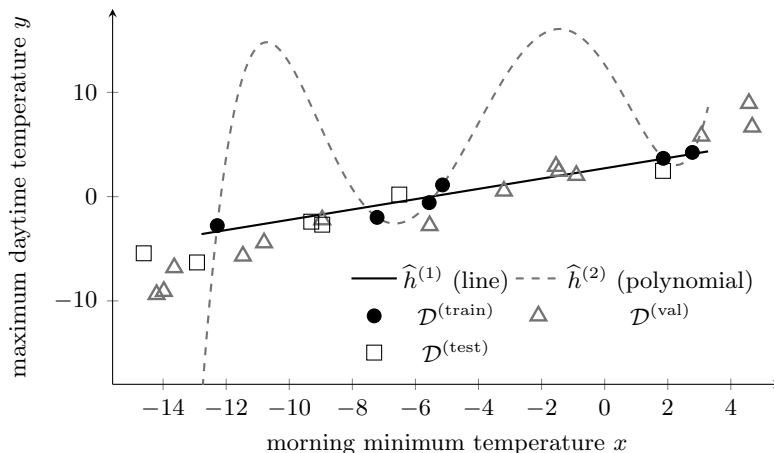


Figure 1: Model selection between two candidates fitted to the same six days (filled circles). The degree-five polynomial $\hat{h}^{(2)}$ passes through every day of the training set, so its training error is zero against 0.47 for the straight line $\hat{h}^{(1)}$. Averaged over the fourteen validation days (open triangles), the squared error loss is 2065 for the polynomial and 7.06 for the line, so the validation error selects the line. The six days of the test set (open squares) are held out from both ERM and the comparison, so the average squared error loss on them, 1.83, estimates the risk of the selected hypothesis. Data generated by `pythondemos/validation.py`

Model selection is the step of a machine learning (ML) workflow that chooses among several trained hypotheses by comparing the validation error each of them incurs (Shalev-Shwartz and Ben-David, 2014, Sect. 11.2.2). Each candidate $\hat{h}^{(c)}$ is the output of ERM run with a different hypothesis space $\mathcal{H}^{(c)}$, a different hyperparameter setting, or both. Typical hyperparameters include the learning rate of a gradient-based method, the strength α of a regularization penalty term, and the depth of a decision tree. The training error cannot serve as the criterion: it falls as the hypothesis space grows, while the risk eventually rises again (Hastie et al., 2009, Sect. 7.2). Model selection therefore follows model training, e.g., via ERM, and model validation, e.g., via k -fold cross-validation (k -fold CV): it re-runs training once per candidate and reads the validation error that validation returns.

The basic workflow has three stages (Fig. 2).

- 1) For each candidate c , run ERM on the training set with hypothesis space $\mathcal{H}^{(c)}$ and the corresponding hyperparameter setting. This yields a trained hypothesis $\hat{h}^{(c)} \in \mathcal{H}^{(c)}$.
- 2) Compute the validation error $E_v(\hat{h}^{(c)})$ of each $\hat{h}^{(c)}$ on a validation set held out from ERM. k -fold CV is the standard estimator of the validation error

when the training set is too small to spare a validation set (Hastie et al., 2009, Sect. 7.10).

- 3) Select the candidate with the smallest validation error. Its index is $c^* = \arg \min_c E_v(\hat{h}^{(c)})$. The selected hypothesis is $\hat{h}^{(c^*)}$.

For example, in predicting customer churn, the candidates are obtained by running ERM once with hypothesis space $\mathcal{H}^{(1)} =$ linear models and once with $\mathcal{H}^{(2)} =$ decision trees; the candidate with smaller validation error is retained, and its loss is reported on the test set.

The test set $\mathcal{D}^{(\text{test})}$ is held out from all three stages and used only afterwards. The independent and identically distributed assumption (i.i.d. assumption) makes its role precise. Suppose the data points of $\mathcal{D}^{(\text{train})}$, $\mathcal{D}^{(\text{val})}$, and $\mathcal{D}^{(\text{test})}$ are realizations of independent and identically distributed (i.i.d.) random variables (RVs) with a common probability distribution $\mathbb{P}^{(\mathbf{z})}$. The average loss on $\mathcal{D}^{(\text{test})}$ is then an unbiased estimator of the risk of $\hat{h}^{(c^*)}$, i.e., of its expected loss on a data point that no stage has seen. Selecting on $\mathcal{D}^{(\text{test})}$ destroys this property: a test set used repeatedly to keep the candidate with the smallest error underestimates the risk of the candidate finally chosen (Hastie et al., 2009, Sect. 7.2).

A/B testing is a further, live-deployment stage of model selection: two short-listed hypotheses $h^{(A)}$ and $h^{(B)}$ are compared on data points drawn from the operating machine learning system (ML system), which can differ from the probability distribution behind $\mathcal{D}^{(\text{train})}$, $\mathcal{D}^{(\text{val})}$, and $\mathcal{D}^{(\text{test})}$. A/B testing is therefore a special case of model selection restricted to two pre-trained candidates and live data, not a synonym for model selection.

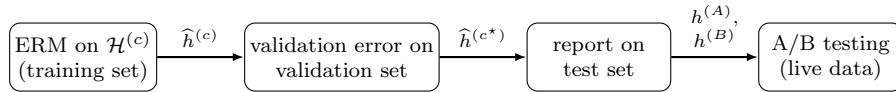


Figure 2: The ML workflow with model selection. For each candidate hypothesis space and hyperparameter setting, $\mathcal{H}^{(c)}$, ERM on the training set produces a candidate trained hypothesis $\hat{h}^{(c)}$. The candidate with the smallest validation error on the validation set is selected as $\hat{h}^{(c^*)}$. Its loss is then reported on the test set, which has remained unused until this point. A/B testing compares two short-listed hypotheses $h^{(A)}$ and $h^{(B)}$ on live data from the operating ML system

Cross References

validation, validation error, k -fold CV, validation set, test set, hyperparameter, hypothesis space, overfitting, A/B testing

References

1. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, 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.