overfitting

Authors

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

Abstract

Overfitting is the failure mode of a machine learning (ML) method that fits its training set too closely: the learned hypothesis incurs a small empirical risk on the training set but a large risk, i.e., a large expected loss on data points outside the training set. Equivalently, the learned hypothesis has a large generalization gap. Overfitting typically arises when the hypothesis space is too large relative to the number of data points in the training set. Regularization counteracts overfitting by pruning the hypothesis space, augmenting the training set, or adding a penalty term to the empirical risk. Validation detects overfitting by comparing the training error with the validation error obtained on a validation set.

Definition

An image classifier that memorizes every photograph in its training set but fails to correctly classify previously unseen images is overfitting. In general, consider a machine learning (ML) method that uses empirical risk minimization (ERM) to learn a hypothesis $\hat{h} \in \mathcal{H}$ with minimum empirical risk $\widehat{L}\big(\hat{h}|\mathcal{D}^{(\text{train})}\big)$ on a given training set $\mathcal{D}^{(\text{train})}$. The method overfits $\mathcal{D}^{(\text{train})}$ if the empirical risk $\widehat{L}\big(\hat{h}|\mathcal{D}^{(\text{train})}\big)$ is small while the risk $\bar{L}\big(\hat{h}\big)$, i.e., the expected loss incurred on data points outside $\mathcal{D}^{(\text{train})}$, is large. In other words, an overfitting ML method learns a hypothesis with a large generalization gap (Hastie et al., 2009; Jung, 2022).

Overfitting typically occurs when the hypothesis space $\mathcal{H}$ is too large relative to the number of data points in the training set. Regularization counteracts overfitting in three elementary forms: pruning the model, adding a penalty term to the empirical risk, or augmenting the training set via data augmentation. Validation detects overfitting: an ML method that overfits delivers a small training error but a large validation error on a validation set.

Fig. 1 illustrates overfitting for polynomial regression. Polynomials of degree $0, \ldots, 9$ are fitted, using ERM with the squared error loss, to a training set of $m = 10$ data points. The training error decreases with increasing degree. The polynomial of degree 9 has 10 model parameters and interpolates the $m = 10$ data points. The validation error, in contrast, is smallest for degree 3 and grows by several orders of magnitude for larger degrees: the high-degree polynomials overfit the training set.

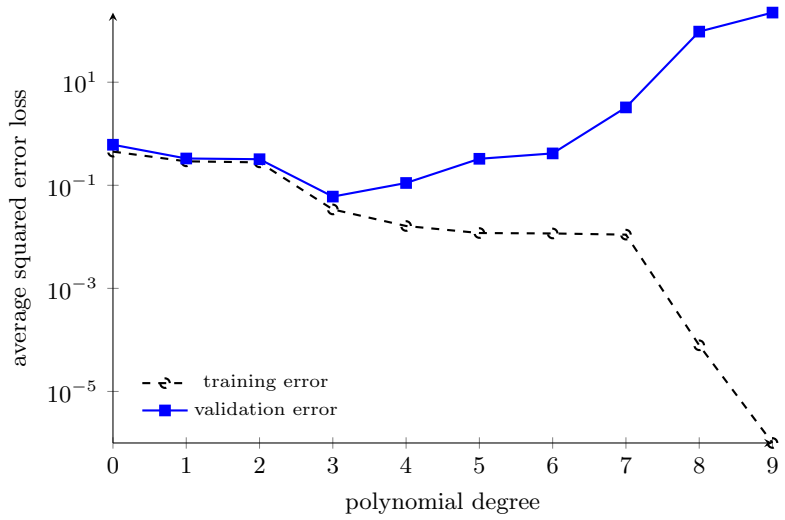

Figure 1: Training error and validation error of polynomial regression over the polynomial degree. The training set consists of $m = 10$ data points with feature x drawn uniformly from $[0, 1]$ and label $y = \sin(2\pi x) + \varepsilon$ with additive Gaussian noise ε. The validation set consists of 100 data points drawn from the same probability distribution. Data generated by `pythondemos/overfitting.py`

Cross References

ERM, generalization gap, generalization, regularization, underfitting, validation

References

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

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.