training set

Authors

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

*Corresponding author: alex.jung@aalto.fi

Abstract

The ultimate goal of machine learning (ML) is to learn a hypothesis that accurately predicts the label of any data point based on its features. A training set is a dataset used to compare the performance of candidate hypotheses: it consists of data points for which the loss incurred by each candidate can be evaluated. Model training methods then pick the hypothesis that performs best on the training set. For example, empirical risk minimization (ERM) learns the hypothesis minimizing the average loss on the training set, the empirical risk; the minimum value itself is the training error. Since the learned hypothesis is picked to perform well on the training set, a small training error can be misleading and result in overfitting. The validation error on a held-back validation set probes whether the learned hypothesis also predicts well outside the training set.

Definition

The ultimate goal of a machine learning (ML) method is to learn a hypothesis (or to train a model) that incurs a small prediction error for any data point. The prediction error is measured by some loss function, and the goal is formalized probabilistically as a small risk: the expected loss under the probability distribution from which the data points are drawn (independent and identically distributed assumption (i.i.d. assumption)). The method must therefore compare the candidate hypotheses from its hypothesis space on data points for which the loss can be evaluated. For example, in regression or classification, the loss can only be computed for data points with a known label. These data points form a dataset referred to as the training set, $\mathcal{D}^{(\text{train})} = \{\mathbf{z}^{(r)}\}_{r=1}^{m}$.

For example, empirical risk minimization (ERM) minimizes the average loss on the training set, the empirical risk, to learn a hypothesis

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{r=1}^{m} L\left(\mathbf{z}^{(r)}, h\right).$$

The empirical risk of the learned hypothesis is the training error. More generally, the training set is the input of the map $\mathcal{A}$ that defines a training method: $\widehat{h} = \mathcal{A}(\mathcal{D}^{(\text{train})})$, and training is to compute this map.

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 of these three days along

with two hypotheses learned from it: the straight line $\widehat{h}^{(1)}$ delivered by 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.

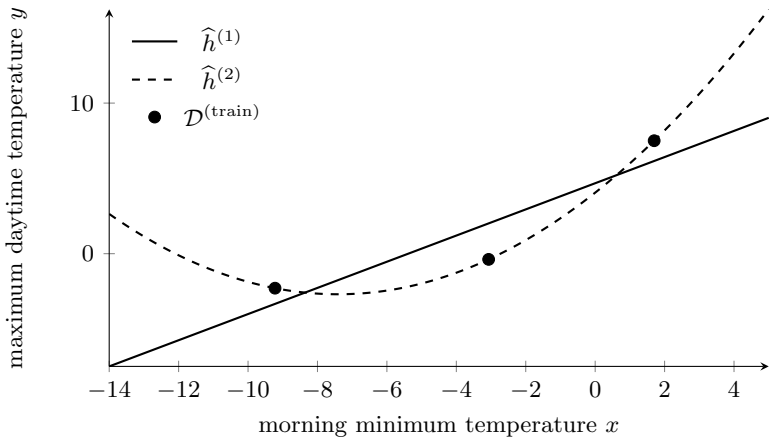

Figure 1: Data points representing daily weather conditions in Finland: each data point is a day, with the morning minimum temperature as its feature and the maximum daytime temperature as its label. The three filled circles form the training set $\mathcal{D}^{(\text{train})}$. Both curves are hypotheses learned from $\mathcal{D}^{(\text{train})}$: the solid straight line $\widehat{h}^{(1)}$ minimizes the average squared error on $\mathcal{D}^{(\text{train})}$ over the linear model, while the dashed degree-two polynomial $\widehat{h}^{(2)}$ passes through all three data points and attains a training error of zero. Both incur a small average loss on $\mathcal{D}^{(\text{train})}$, but the two curves differ sharply outside the training set. Data generated by `pythondemos/trainset.py`

Which data points are contained in $\mathcal{D}^{(\text{train})}$ decides what is learned. The choice of the training set is therefore crucial for trustworthy artificial intelligence (trustworthy AI): a training set that underrepresents a group of data points yields a learned hypothesis that incurs a larger loss on that group, a source of unfairness (see fairness) (Buolamwini and Gebru, 2018; Barocas et al., 2019; Mehrabi et al., 2021). Since $\widehat{h}$ is chosen to make this particular average small, the training error understates the loss that the same hypothesis incurs on data points outside $\mathcal{D}^{(\text{train})}$ (Jung, 2022, Sect. 6.2). The training error is therefore a poor estimate of the loss incurred outside the training set. Fig. 1 shows how misleading it can be: the polynomial $\widehat{h}^{(2)}$ has a smaller training error than the straight line, yet the two curves differ sharply away from the training data points, which is where the loss on new data points is incurred (see overfitting).

A second subset of data points, the validation set $\mathcal{D}^{(\text{val})}$, is therefore held back from model training (Fig. 2): the average loss on $\mathcal{D}^{(\text{val})}$, the validation error, probes the learned hypothesis on data points it was not trained on. Comparing the two errors diagnoses the ML method: if both are small, the learned

hypothesis also predicts well outside $\mathcal{D}^{(\text{train})}$; a small training error paired with a validation error far above it means the hypothesis space is too large compared to the number of data points in $\mathcal{D}^{(\text{train})}$ (Jung, 2022, Sect. 6.6; Hastie et al., 2009).

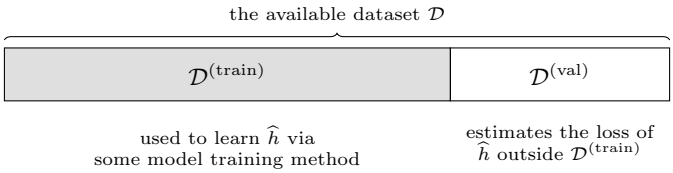

Figure 2: The available dataset split into two disjoint subsets. Only the shaded subset, the training set $\mathcal{D}^{(\text{train})}$, enters the training method $\mathcal{A}$ that delivers a learned hypothesis $\widehat{h}$; the validation set is what remains to estimate how that hypothesis behaves on data points it was not trained on

Cross References

training, dataset, validation set, test set, data point, ERM, hypothesis, loss, training error, validation error, hypothesis space, overfitting

References

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

2. Mehrabi et al. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), pp. 115:1–115:35.

3. Barocas et al. (2019). Fairness and Machine Learning. fairmlbook.org.

4. Buolamwini and Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91.

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