

generalization

Authors

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

*Corresponding author: alex.jung@aalto.fi

Abstract

Generalization is the ability of a machine learning (ML) method to make accurate predictions on data points that have not been used during training. Many ML methods implement training via empirical risk minimization (ERM), i.e., learning a hypothesis $\hat{h}$ that minimizes the empirical risk on the training set $\mathcal{D}^{(\text{train})}$. The learned $\hat{h}$ is then used to compute predictions on new data points outside $\mathcal{D}^{(\text{train})}$. A probabilistic model for data generation quantifies the prediction errors beyond $\mathcal{D}^{(\text{train})}$ via the risk of $\hat{h}$, i.e., the expected loss incurred on a new randomly chosen data point. The difference between the risk and the empirical risk is the generalization gap. A complementary deterministic view characterizes generalization via the robustness of $\hat{h}$ to small perturbations of the features of a data point.

Definition

A weather forecaster trained on past data is useful only if its predictions are accurate on future unseen data points. Generalization refers to the ability of a learned hypothesis $\hat{h}$ to make similarly accurate predictions on data points that have not been used during training. It is the central goal of machine learning (ML): learning patterns in the features and labels of data points that extend beyond the training set.

The canonical learning technique is empirical risk minimization (ERM): the machine learning system (ML system) obtains $\hat{h} \in \mathcal{H}$ by minimizing the empirical risk on the training set $\mathcal{D}^{(\text{train})}$. A low empirical risk on $\mathcal{D}^{(\text{train})}$ does not guarantee accurate predictions on data points outside $\mathcal{D}^{(\text{train})}$. Online learning and Bayesian inference face the same generalization challenge.

Studying generalization mathematically requires formalizing the notion of *unseen* data points. A widely used approach is to assume a probabilistic model for the generation of data points, such as the independent and identically distributed assumption (i.i.d. assumption). Here, data points are interpreted as independent random variables (RVs) with an identical, fixed but unknown, probability distribution $\mathbb{P}$. Fig. 1 sketches this setup: the shaded ellipse depicts the region of typical data points under $\mathbb{P}$, and the two large filled circles are the data points of the training set that ERM fits with the learned hypothesis $\hat{h}$. The risk of $\hat{h}$ is the expected loss:

$$\bar{L}(\hat{h}) = \mathbb{E} \left\{ L(\mathbf{z}, \hat{h}) \right\}.$$

The difference $\bar{L}(\hat{h}) - \widehat{L}(\hat{h}|\mathcal{D}^{(\text{train})})$ is the generalization gap.

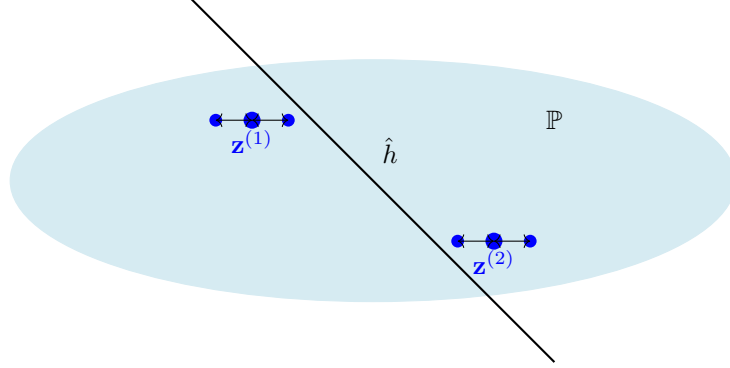


Figure 1: Two data points $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$ used as a training set to learn a hypothesis $\hat{h}$ via ERM. The shaded ellipse marks the region of typical data points under an assumed probability distribution $\mathbb{P}$

For any fixed hypothesis $h \in \mathcal{H}$, $\bar{L}(h)$ is a deterministic number that depends only on h and $\mathbb{P}$, whereas $\widehat{L}(h|\mathcal{D}^{(\text{train})})$ is an RV that depends on $\mathcal{D}^{(\text{train})}$, which, in turn, is a realization of an independent and identically distributed (i.i.d.) sequence of RVs $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)} \sim \mathbb{P}$. Bounding the generalization gap therefore means controlling the probability of the (undesirable) event

$$\mathcal{E}^{(h)} := \left\{ \mathcal{D}^{(\text{train})} : |\widehat{L}(h|\mathcal{D}^{(\text{train})}) - \bar{L}(h)| > \varepsilon \right\},$$

where $\varepsilon > 0$ is a user-specified tolerance for the generalization gap. Each $h \in \mathcal{H}$ defines a different event $\mathcal{E}^{(h)}$ (see Fig. 2), and generalization requires low probability of $\mathcal{E}^{(h)}$ uniformly over $\mathcal{H}$ (Bishop, 2006, Sect. 1.5; Shalev-Shwartz and Ben-David, 2014, Ch. 4).

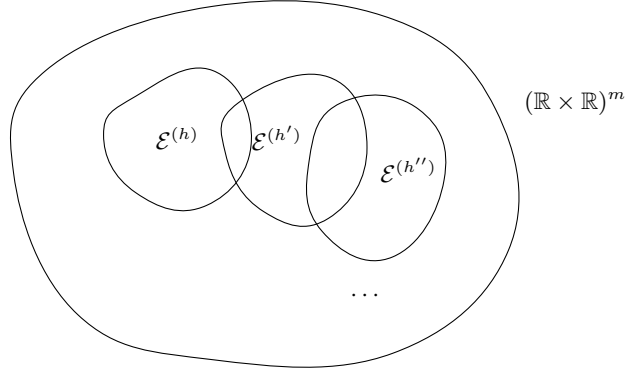


Figure 2: Consider an ML application with feature space $\mathcal{X} = \mathbb{R}$ and label space $\mathcal{Y} = \mathbb{R}$. A training set of size m is a point in the sample space $(\mathbb{R} \times \mathbb{R})^m$ of all possible training sets of size m (the outer region). Each hypothesis $h \in \mathcal{H}$ defines an event $\mathcal{E}^{(h)} \subseteq (\mathbb{R} \times \mathbb{R})^m$ (an inner region) collecting those training sets on which h fails to generalize. Ensuring that the ML method is likely to generalize requires the union $\bigcup_{h \in \mathcal{H}} \mathcal{E}^{(h)}$ to have low probability, not just each individual $\mathcal{E}^{(h)}$

Probability theory is not the only route to generalization. A complementary deterministic approach uses robustness: a good hypothesis $\hat{h}$ should not change its prediction $\hat{h}(\mathbf{x})$ much if the features $\mathbf{x}$ of a data point $\mathbf{z}$ are slightly perturbed. For example, an object detection system trained on smartphone photos should still detect the object even if a single pixel is changed (Su et al., 2019), and it should deliver the same result if the object in the image is rotated (Mallat, 2016). The arrows in Fig. 1 indicate small perturbations of each data point in the training set of an ML method.

Cross References

ERM, empirical risk, risk, generalization gap, i.i.d. assumption, training set, validation error, test set, overfitting, validation, regularization, hypothesis, hypothesis space, online learning, Bayesian inference

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.
2. Mallat (2016). Understanding deep convolutional networks. Philos. Trans. Roy. Soc. A, 374(2065), pp. Art. no. 20150203.
3. Su et al. (2019). One Pixel Attack for Fooling Deep Neural Networks. IEEE Trans. Evol. Comput., 23(5), pp. 828–841.

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