supervised learning

Authors

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

Abstract

Supervised learning is the machine learning (ML) setting in which every data point of the training set carries a label, the quantity to be predicted, alongside its features. The goal is to learn a hypothesis that maps the feature vector of a data point to its label and incurs a small loss on data points outside the training set. The standard formalization is empirical risk minimization (ERM), which chooses a hypothesis minimizing the average loss on the training set; without labels that average cannot be formed. The label space decides the learning task: a numeric label makes it regression, a discrete one classification. Labels do not by themselves deliver generalization, since a hypothesis space large enough to fit the training set exactly can incur an arbitrarily large loss elsewhere. Because labels are often expensive to obtain, unsupervised learning, self-supervised learning and semi-supervised learning (SSL) each reduce what must be annotated.

Definition

A weather station at Krems records, for every day of 2024, the minimum temperature in the morning and the maximum temperature during the day. To forecast the afternoon from the morning, each day is taken as a data point: its feature is the morning minimum, and the maximum of that day is the quantity to be predicted. For the days already recorded, that quantity is known, because it was measured. Supervised learning is the machine learning (ML) setting in which every data point of the training set comes with the quantity to be predicted, its label (Sutton and Barto, 2018, Sect. 1.1; Hastie et al., 2009, Ch. 2).

Formally, the training set consists of labeled data points $\mathcal{D}^{(\text{train})} = \big\{ \big(\mathbf{x}^{(r)}, y^{(r)}\big) \big\}_{r=1}^{m}$ with feature vectors $\mathbf{x}^{(r)} \in \mathcal{X}$ and labels $y^{(r)} \in \mathcal{Y}$. The goal is to learn a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that incurs a small loss on data points outside $\mathcal{D}^{(\text{train})}$, which is what generalization demands. That requirement is made precise by the independent and identically distributed assumption (i.i.d. assumption): the data points are realizations of independent and identically distributed (i.i.d.) random variables (RVs) with a common probability distribution, and the quantity to be small is the risk, the expected loss under it. The labels make this a learning task that can be formalized as empirical risk minimization (ERM): among the hypotheses of a hypothesis space $\mathcal{H}$, choose

one that minimizes the average loss on the training set,

$$\widehat{h} \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{m} \sum_{r=1}^{m} L\left(\left(\mathbf{x}^{(r)}, y^{(r)}\right), h\right).$$

Without labels this average cannot be formed, since the loss of a prediction is defined only against the label it is compared with. Whether the learned hypothesis is any good is measured on data points kept out of training, the validation set. Fig. 1 shows both for the days at Krems, with a hypothesis delivered by ERM over the linear model.

The label space decides which learning task it is. A numeric label space $\mathcal{Y} = \mathbb{R}$ makes the task regression, a discrete one makes it classification. The weather records deliver both from the same data points. With the maximum temperature of the day as label the task is regression. With the label *frost in the morning*, which is positive on 19% of the days, it is classification, and ERM with the 0/1 loss, which counts the data points whose label is predicted wrongly, delivers a classifier that is correct on 81% of the held-out days, against 71% for always answering with the more frequent of the two labels.

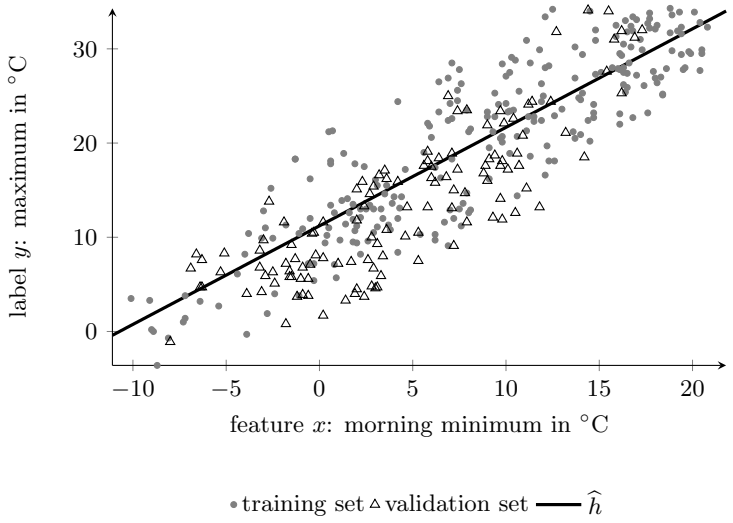

Figure 1: The 366 days of 2024 at Krems as labeled data points. The feature of a day is its morning minimum temperature, its label the maximum temperature of that day. The hypothesis $\widehat{h}$ is delivered by ERM over the linear model on the training set of the days from January to August; the days from September to December form the validation set and are not used for training. Data generated by `pythondemos/supervisedlearning.py`

The labels carry the information that makes the prediction possible, but they do not guarantee it. In Fig. 1 the learned hypothesis attains a training error of 17.8, measured as the mean squared error (MSE) in squared degrees

Celsius, against a sample variance of 80.7 for the labels. Its validation error on the held-out days is 24.3, against 108.1 for the hypothesis that answers with the sample mean of the training labels. Fitted instead to ten days over a hypothesis space of polynomials of degree twelve, ERM drives the training error to zero and the validation error to more than 10^7, against 23.3 for the linear model on the same ten days. A label for every data point of the training set is therefore not sufficient: the hypothesis space must also be small enough for the training error to say something about the loss outside $\mathcal{D}^{(\text{train})}$, which is what a validation set reveals (see overfitting) (Hastie et al., 2009, Sect. 7.2; Jung, 2022, Sect. 6.2).

The labels that make all of this possible are not free. Where one comes from varies: it can be measured, as the maximum temperature of a day is; it can be recorded, as the price at which a house sold is; or it must be annotated by a person, as the diagnosis shown by a medical image is. In the last case the labels are the scarce resource: annotating requires the expert whose judgment is to be reproduced, so a dataset of unlabeled data points is typically much cheaper to enlarge than a training set of labeled data points. Assembling ImageNet, a dataset of more than three million annotated images, took tens of thousands of people working through a crowdsourcing platform (Deng et al., 2009). Three settings answer this shortage differently. Unsupervised learning drops the labels entirely and looks for structure in the feature vectors alone. Self-supervised learning constructs the labels from the data points themselves, by withholding some of their features, and so returns an ordinary supervised problem whose training set no annotator has touched. Semi-supervised learning (SSL) uses a few labeled data points together with many unlabeled ones. Reinforcement learning (RL) has no label at all: it observes a reward for the action it took, and never the action that would have been correct (Sutton and Barto, 2018, Sect. 1.1).

Cross References

label, labeled data point, training set, validation set, classifier, i.i.d. assumption, risk, ERM, hypothesis, classification, regression, generalization, overfitting, unsupervised learning, self-supervised learning, SSL, RL

References

1. Deng et al. (2009). ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

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

3. Sutton and Barto (2018). Reinforcement Learning: An Introduction. 2nd ed., MIT Press, Cambridge, MA, USA.

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