label

Authors

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

Synonyms

target, response, ground truth, quantity of interest

Abstract

The ultimate goal of machine learning (ML) is the accurate prediction of the label of a data point from its features. A label is an attribute of a data point that represents a quantity of interest. Determining the label of a data point is error-prone and often involves human annotation. The choice of label determines the learning task: whether the task is classification or regression depends on the label space together with the loss function. In self-supervised learning, the label is constructed from the features themselves, so that no manual annotation is required.

Definition

A label of a data point $\mathbf{z}$ is one of its attributes that represents a higher-level fact, or quantity of interest, and that typically requires human expertise or domain knowledge to determine (Dodge, 2003; Everitt and Skrondal, 2010; Gujarati and Porter, 2009). Unlike a feature, which can be computed or measured easily, a label is usually harder to obtain.

For a data point that represents a patient, a feature could be the body weight, while the label could be the presence of a disease. Whether a given attribute serves as a feature or a label is often a design choice that depends on the resources available in a given machine learning (ML) application. If an oncologist is readily available, a cancer prognosis can serve as a feature. On the other hand, if no oncologist is available, the cancer prognosis is a label that needs to be predicted from patient features.

One way to make the notion of a label precise is by introducing a labeling function $\bar{h} : \mathcal{Z} \to \mathcal{Y}$. The domain of this function is the data point space $\mathcal{Z}$, the set of all possible data points that can occur in a given ML application. Its values lie in the label space $\mathcal{Y}$, the set of all possible values the label of a data point can take on. The function assigns each data point $\mathbf{z}$ its label $y = \bar{h}(\mathbf{z})$ (Shalev-Shwartz and Ben-David, 2014, Ch. 2).

Note that the labeling function $\bar{h}$ acts on the data point $\mathbf{z}$ itself, that is, on the whole patient. A trained model (or learned hypothesis) $\hat{h}$ instead acts on the feature vector $\mathbf{x}$, a fixed list of recorded features such as the patient's body weight and age. The feature vector usually captures only part of a data point.

For example, in a health-care application, the data point space $\mathcal{Z}$ contains patients, whereas the feature space $\mathcal{X}$ contains only their recorded feature vectors. Two patients with the same feature vector may therefore carry different labels, since they can differ in information that was never recorded as a feature. No hypothesis $\hat{h}$ that reads only $\mathbf{x}$ can then deliver perfect predictions.

In practice, the true label $\bar{h}(\mathbf{z})$ of a data point is rarely known exactly. Even with full access to the data point, its label must be determined by human judgment or measurement, which is imperfect, so the available label is only a noisy version of $\bar{h}(\mathbf{z})$ (Hastie et al., 2009, Sect. 2.6).

For instance, an object category assigned by a human annotator or a diagnosis obtained from a clinician may be wrong, even though the annotator or clinician examines the data point directly. The discrepancy between such an observed label and the true label is called label noise and can degrade training and the resulting generalization. This source of error differs from the incompleteness of the recorded features discussed above: here the data point is accessible, but its label is determined imperfectly.

Label noise also matters for regulation. The EU AI Act requires the minimization of label noise and its impact on high-risk machine learning systems (ML systems) (European Parliament and Council of the European Union, 2024, Arts. 10(3) and 15(1)). When labels are personal data, the general data protection regulation (GDPR) requires them to be accurate (European Parliament and Council of the European Union, 2016, Art. 5(1)(d)).

One way to reduce the effect of label noise is label smoothing, which slightly perturbs the labels of data points used for model training (Goodfellow et al., 2016, Sect. 7.5.1). It is a form of data augmentation that acts as regularization.

The choice of label (or labeling function $\bar{h}$) determines a specific learning task. For the image of a cow herd in Fig. 1, the label could be a binary indicator of whether the image contains cows (binary classification), the number of cows or the average green level (regression), or a masked center pixel predicted from the surrounding pixels (self-supervised learning).

$y \in \{0, 1\}$: "contains cows?" $\rightarrow$ *binary classification*

$y \in \{1, 2, 3, \ldots\}$: "number of cows" $\rightarrow$ *regression*

$y \in \mathbb{R}$: "average green level" $\rightarrow$ *regression*

$y =$ center pixel of patch $\rightarrow$ *self-supervised learning*

Figure 1: The same data point yields different learning tasks depending on the choice of label

Whether a learning task is a classification problem or a regression problem

depends not only on the label space $\mathcal{Y}$ but also on the loss function. For example, the label space $\mathcal{Y} = \{0, 1, 2, 3, \ldots\}$ can be used for a classification method that uses the 0/1 loss as the loss function. The same label space can also be used for a regression method that uses the squared error loss as the loss function.

In self-supervised learning, the label is constructed from the features themselves, so no manual annotation is required. A prominent example is large language model (LLM) training: consider the sentence "All human beings are born free and equal," tokenized as the sequence ("All", "human", "beings", "are", "born, "free", "and", "equal"). The label for the subsequence ("All", "human", "beings") is the next token "are," the label for ("All", "human", "beings", "are") is "born," and so on. By sliding through the sequence, a single sentence produces many labeled data points from plain text alone.

Cross References

data point, feature, label space, learning task, classification, regression, self-supervised learning, LLM, EU AI Act, robustness, data augmentation, regularization

References

1. Dodge (2003). The Oxford Dictionary of Statistical Terms. Oxford Univ. Press, New York, NY, USA.

2. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

3. Everitt and Skrondal (2010). The Cambridge Dictionary of Statistics. 4th ed., Cambridge Univ. Press, Cambridge, U.K.

4. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

5. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

6. Gujarati and Porter (2009). Basic Econometrics. 5th ed., McGraw-Hill/Irwin, New York, NY, USA.

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

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