data point

Authors

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

Abstract

A data point is the elementary information-carrying unit on which machine learning (ML) methods operate. The properties of a data point fall into two categories: features, which are easily measurable or computable, and labels, which are higher-level facts that typically can only be determined using human expertise. Whether a given property is treated as a feature or a label is a design choice that depends on the ML application.

Definition

A data point is an object that conveys information (Cover and Thomas, 2006); such objects are the elementary units on which machine learning (ML) methods operate. Examples include students, radio signals, trees, images, random variables (RVs), real numbers, or proteins. Data points of the same type are described by two categories of properties.

The first category includes features that are measurable or computable properties of a data point. They can be automatically extracted or computed using sensors, computers, or other data collection systems. The second category includes labels that are higher-level facts, or quantities of interest, that typically require human expertise or domain knowledge to determine rather than being directly measurable. Whether a given property serves as a feature or a label is ultimately a design choice that depends on the resources available in a given ML application.

Fig. 1 shows an image as an example of a data point. Several properties of the image can serve as features: the color intensities $x_1, \ldots, x_d$ of all image pixels, the timestamp x_{d+1} of the image capture, and the spatial location x_{d+2} of the image capture. Higher-level facts can serve as labels: the number of cows y_1, the number of wolves y_2, and the condition of the pasture y_3 (e.g., healthy or overgrazed).

Figure 1: Illustration of a data point consisting of an image. Different properties of the image can serve as features, such as the average color intensities of the image pixels. Higher-level facts about the image, such as the presence of different object categories, can serve as labels

For a data point that represents a patient, a feature could be the body weight, while the label could be the presence of cancer, the ground-truth fact of interest. Determining this label requires a medical expert (or even a committee of experts) to examine the patient. A property treated as a label in one setting (e.g., a cancer diagnosis) may serve as a feature in another, where reliable automation (e.g., image analysis) allows the property to be computed without human intervention. ML aims to predict the label of a data point from its features.

Both kinds of properties are error-prone. For many ML applications, it is rarely possible to access the true labels. For example, a medical expert's diagnosis can be wrong, rendering it a noisy proxy for the patient's true condition. Errors of this kind, called label noise, arise whenever a label is only a proxy for the true quantity of interest, whether that proxy is produced by human judgment or by an automated system.

Features provide also a source of error within an ML method. A feature is obtained by measuring or computing a property of a data point. Physical and chemical sensing devices have a finite measurement uncertainty (Kimothi, 2002), and computing a feature on a finite computer introduces rounding errors (Golub and Van Loan, 2013, Sect. 2.7). These imperfections make the recorded value deviate from the true feature value. This discrepancy is feature noise.

ML methods should be robust to both feature and label noise. Their predictions on new data points should not degrade sharply when some features or labels in the training set are corrupted. The EU AI Act reflects this concern: training data must, to the best extent possible, be "free of errors" (European Parliament and Council of the European Union, 2024, Art. 10(3)), and high-risk artificial intelligence systems (high-risk AI systems) must attain an appropriate level of robustness (European Parliament and Council of the European Union, 2024, Art. 15(1)).

Robustness to feature and label noise can be improved by data augmentation, which deliberately perturbs the features or the label of a training data point

(see Fig. 2). Training on such perturbed data points acts as regularization and reduces the influence of corrupted feature or label values on the learned hypothesis.

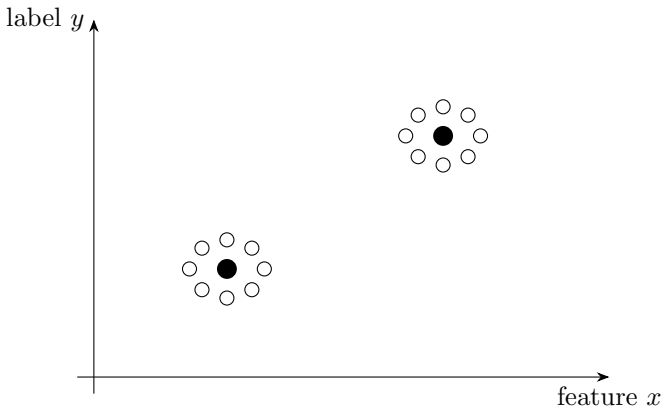

Figure 2: Data augmentation surrounds a data point (filled circle) with a cloud of perturbed copies (open circles). It is a form of regularization that makes an ML method less sensitive to feature and label noise

Cross References

data, feature, label, dataset, robustness, EU AI Act, data augmentation

References

1. 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.

2. Golub and Van Loan (2013). Matrix Computations. 4th ed., The Johns Hopkins Univ. Press, Baltimore, MD, USA.

3. Cover and Thomas (2006). Elements of Information Theory. 2nd ed., Wiley, Hoboken, NJ, USA.

4. Kimothi (2002). The Uncertainty of Measurements: Physical and Chemical Metrology Impact and Analysis. American Society for Quality (ASQ), La Vergne.