data

Authors

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

Abstract

This entry distinguishes three main uses of the term *data*. In a general sense, data refer to an abstract raw material that is used (or consumed) by machine learning (ML) methods. As a data point, it is an elementary information-carrying unit. As a dataset, it is an indexed collection of data points. A related regulatory term, personal data, is defined in the general data protection regulation (GDPR) and the EU AI Act.

Definition

A weather station measures the temperature at its location and stores the value 23.4 °C together with a timestamp (see Fig. 1(a)). Such recorded values are data: representations of information, recorded in a form suitable for storage, communication, and processing by machine learning systems (ML systems) (International Organization for Standardization and International Electrotechnical Commission, 2015). Every machine learning (ML) method fits a model to data, and the usefulness of the hypothesis the method learns is limited by the quality and quantity of the available data.

The word *data* is used in three technical senses, summarized in Fig. 1, and additionally in a legal sense. In its broadest use, data refer to recorded information. This sense applies when the distinction between unit and collection is irrelevant. Examples are raw sensor measurements such as the temperature reading in Fig. 1(a) and compound terms such as data parallelism, networked data, and data augmentation.

As a formal unit, a data point is the fundamental information-carrying unit (Fig. 1(b)). Its information is contained in two types of properties: features (easy to measure or compute) and labels (difficult to measure or compute, or in need of human annotation) (Shalev-Shwartz and Ben-David, 2014, Ch. 2).

As a formal collection, a dataset $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ is the indexed collection of data points on which model training and validation are performed (Fig. 1(c)) (Goodfellow et al., 2016, Sect. 5.1).

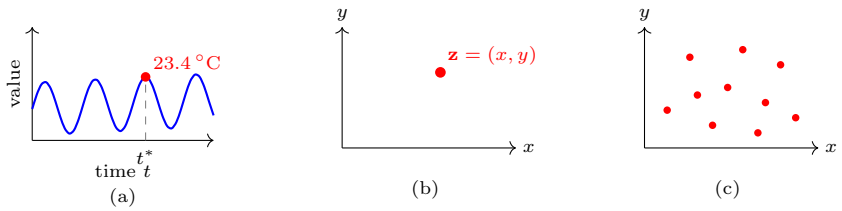

Figure 1: Three technical meanings of the word *data* in ML applications. (a) Data in the general sense: a temperature reading sampled at time instant t^*. (b) A data point $\mathbf{z} = (x, y)$ with a single feature x and label y. (c) A dataset that consists of a finite number of data points

The word *data* also carries a regulatory meaning. The general data protection regulation (GDPR) (European Parliament and Council of the European Union, 2016) and the EU AI Act (European Parliament and Council of the European Union, 2024) use personal data and biometric data as defined legal terms. Here, data denote any recorded information about an identifiable natural person. Several regulatory obligations on ML systems, including the data minimization principle, the right to erasure, and the right to explanation, refer to data in this legal sense (European Parliament and Council of the European Union, 2016, Art. 4(1); European Parliament and Council of the European Union, 2024, Art. 3).

Cross References

data point, dataset, model, hypothesis, GDPR, EU AI Act, personal data

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. 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.

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

4. International Organization for Standardization and International Electrotechnical Commission (2015). Information technology — Vocabulary. ISO/IEC 2382:2015. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v2:en.

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