feature

Authors

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

Synonyms

covariate, explanatory variable, independent variable, input, predictor, regressor

Abstract

A feature of a data point is one of its attributes that can be measured or computed easily, without human supervision. The features of a data point are collected into a feature vector, which a hypothesis map reads to predict the label of the data point. The red-green-blue (RGB) pixel intensities of a digital image or the signal values of an audio recording are typical features. Which attributes of a data point serve as features, rather than as labels, is a design choice within a machine learning (ML) application.

Definition

A feature of a data point $\mathbf{z}$ is one of its attributes that is measurable, or can be computed easily, without the need for human supervision (Dodge, 2003; Everitt and Skrondal, 2010; Gujarati and Porter, 2009). The features of a data point are assembled into a feature vector $\mathbf{x} = \mathbf{\Phi}(\mathbf{z})$ by a feature transformation, a function $\mathbf{\Phi} : \mathcal{Z} \rightarrow \mathcal{X}$, viewing the data point itself as its raw features from which the computed features are obtained.

For example, if a data point is a digital image, then the red-green-blue (RGB) intensities of its pixels can serve as features. Another example is shown in Fig. 1, where the signal values of a finite-duration audio signal are used as its features.

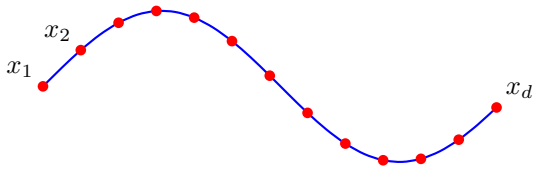

Figure 1: Audio signal (blue waveform) and its signal values (red dots), which can be used as features $x_1, \ldots, x_d$

New features can be constructed by transforming existing features. Such a transformation can be an arithmetic computation applied to the existing features. For example, if a data point has a numeric feature $x \in \mathbb{R}$, the quantities x^2, $\exp(-x)$, and $3x$ can also be used as features of the data point.

Instead of its raw signal values, the audio signal in Fig. 1 can be described by the magnitudes of its discrete-time Fourier transform (DTFT) (Oppenheim et al., 1999). These features capture the frequency content of the signal and are unchanged by time shifts, which is useful for a stationary signal, one whose statistical properties do not change over time. Another example is the activation of a neuron within an artificial neural network (ANN): it is a new feature derived from the input features by a sequence of basic computations encoded by the ANN.

Using activations as features underlies a second, narrower use of the word *feature* in mechanistic interpretability: a feature is a human-interpretable concept encoded inside a trained ANN, such as a curve detector or the presence of a wheel. The activations of a subset of d neurons (for example, a whole layer, or the neurons that an analyst selects for study) form a vector in a Euclidean space $\mathbb{R}^d$. Whereas the broader sense above treats a single neuron activation as one feature, here a feature is obtained as a function of this whole activation vector. One such function is the projection of the activation vector onto a one-dimensional subspace spanned by a unit vector (or direction) (Olah et al., 2020; Park et al., 2024). This unit vector can be obtained from a hyperplane separating data points that express the concept from those that do not (Kim et al., 2018) (see Fig. 2).

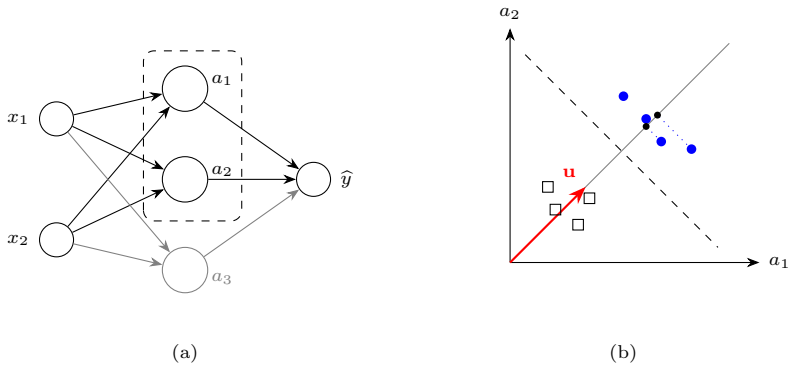

Figure 2: A feature as a function of an activation vector $\mathbf{a}$ obtained from a trained ANN. (a) An ANN with one hidden layer of three neurons; the activations a_1, a_2 of a selected subset (dashed box) form the activation vector $\mathbf{a} = \big(a_1, a_2\big)^\top$, while a_3 is not used. (b) Each data point maps to a point $(a_1, a_2)^\top$ in the activation space. Data points that express the concept (blue filled circles) and those that do not (open squares) are separated by a hyperplane (dashed line) with unit normal $\mathbf{u}$. The feature value is the projection of the activation vector onto $\mathbf{u}$ (blue dotted lines, shown for two points)

Another construction of an activation-based feature is based on regions instead of directions in the activation space (Crabbé and van der Schaar, 2022). In either case, constructing the function that maps the activation vector to the

feature requires a set of data points labeled as representing the concept or not.

Whether a given attribute is treated as a feature or a label is not inherent to the data point. It is more a design choice that depends on the machine learning (ML) application and the resources available for determining each attribute. An attribute should be used as a feature only if its value can be determined with sufficient accuracy.

The accuracy of features also matters for regulation. The EU AI Act requires the training data of high-risk machine learning systems (ML systems) to be relevant, representative, and as error-free as possible (European Parliament and Council of the European Union, 2024, Art. 10). When features are personal data, the general data protection regulation (GDPR) adds the data minimization principle and the accuracy principle (European Parliament and Council of the European Union, 2016, Art. 5(1)(c) and (d)).

Conversely, deliberately perturbing the features can be useful: adding small perturbations to the features of data points during training is a form of data augmentation that acts as regularization. For example, the ridge regression penalty equals the average loss under Gaussian feature perturbations.

Cross References

data point, label, feature vector, feature space, dataset, mechanistic interpretability, CAV, EU AI Act, data augmentation, regularization, ridge regression, Fourier transform

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

6. Olah et al. (2020). Zoom In: An Introduction to Circuits. Distill, 5(3).

7. Oppenheim et al. (1999). Discrete-Time Signal Processing. 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

8. Crabbé and van der Schaar (2022). Concept Activation Regions: A Generalized Framework for Concept-Based Explanations. In: Adv. Neural Inf. Process. Syst., pp. 2590–2607.

9. Kim et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Mach. Learn., pp. 2668–2677.

10. Park et al. (2024). The Linear Representation Hypothesis and the Geometry of Large Language Models. In: Proc. 41st Int. Conf. Mach. Learn., pp. 39643–39666.