

sigmoid function

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Synonyms

logistic function

Abstract

Definition

The sigmoid function carries the whole real line into the interval between 0 and 1, turning the unbounded score that a classifier assigns to a data point into a probability. It is the function $\sigma : \mathbb{R} \rightarrow (0, 1)$ given by

$$\sigma(z) := \frac{1}{1 + e^{-z}}.$$

It rises from $\sigma(z) \rightarrow 0$ as $z \rightarrow -\infty$ to $\sigma(z) \rightarrow 1$ as $z \rightarrow +\infty$, and passes through $\sigma(0) = 1/2$ (see Fig. 1). It is also called the logistic function, and it is the softmax function for two classes.

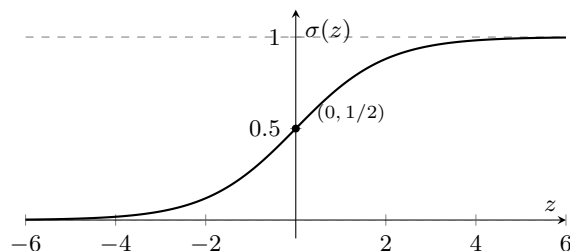


Figure 1: The sigmoid function $\sigma(z) = 1/(1 + e^{-z})$ carries the real line into the open interval $(0, 1)$, with $\sigma(0) = 1/2$

The sigmoid function is smooth and strictly increasing, and it satisfies the symmetry $\sigma(-z) = 1 - \sigma(z)$ (Bishop, 2006, Sect. 4.2). Its derivative is available in closed form, and in terms of the function itself,

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

Once $\sigma(z)$ has been evaluated, its derivative therefore costs one multiplication. That is what makes the sigmoid cheap to differentiate inside an artificial neural

network (ANN). The product $\sigma(z)(1 - \sigma(z))$ is at most $1/4$, and it approaches 0 for large $|z|$. A neuron whose score is large in magnitude therefore passes almost no gradient back. That is why the rectified linear unit (ReLU) replaced the sigmoid in the hidden layers of deep nets (Goodfellow et al., 2016, Sect. 6.3.2).

The sigmoid function turns a log-odds into a posterior probability. Let the binary label y be either 1 or -1 . Write a for the log-odds of the two classes at the feature vector $\mathbf{x}$,

$$a := \log \frac{\mathbb{P}(\mathbf{x} \mid y = 1) \mathbb{P}(y = 1)}{\mathbb{P}(\mathbf{x} \mid y = -1) \mathbb{P}(y = -1)}.$$

The posterior probability of the label 1 is then exactly $\mathbb{P}(y = 1 \mid \mathbf{x}) = \sigma(a)$, and the inverse recovers the log-odds, $a = \log(\sigma/(1 - \sigma))$ (Bishop, 2006, Sect. 4.2). A linear model with model parameters $\mathbf{w}$ scores the feature vector by $\mathbf{w}^\top \mathbf{x}$. Equating that score with a is what turns $\sigma(\mathbf{w}^\top \mathbf{x})$ into a posterior probability, so the reading is an assumption rather than a convention: that the log-odds is a linear function of the feature vector. The assumption holds exactly when the feature vectors of the two classes follow Gaussian random variables (Gaussian RVs) sharing one covariance matrix (Bishop, 2006, Sect. 4.2.1). Logistic regression instead fits $\mathbf{w}$ to a training set without assuming any form for those feature vectors (Bishop, 2006, Sect. 4.3).

Cross References

activation function, ReLU, softmax function, logistic regression, linear model, classifier, probability, smooth

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.
2. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.