

logistic loss

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

log loss, binary cross-entropy loss

Abstract

The logistic loss is a convex and smooth surrogate for the 0/1 loss in binary classification, defined for a label $y \in \{-1, +1\}$ and a real-valued hypothesis h as a decreasing function of the margin $y \cdot h(\mathbf{x})$. Like the hinge loss, it uses a data point only through that margin. Rescaled by its value at margin zero it is an upper bound on the 0/1 loss, with equality there, so a small logistic loss forces a small 0/1 loss. Unlike the hinge loss it is differentiable everywhere, which suits it to gradient-based optimization methods such as gradient descent (GD). Running empirical risk minimization (ERM) with a linear hypothesis and the logistic loss gives logistic regression.

Definition

A spam filter either flags an email or does not, so its mistakes are counted rather than measured: the 0/1 loss charges 1 for every wrong answer. That count is flat almost everywhere as a function of the model parameters, so it gives gradient descent (GD) no direction to move in. The logistic loss replaces it by a convex and smooth surrogate that lies above it.

Consider a binary classification problem with label $y \in \{-1, +1\}$ and a real-valued hypothesis $h : \mathcal{X} \rightarrow \mathbb{R}$. On a data point $(\mathbf{x}, y)$ the logistic loss is (Bishop, 2006)

$$L((\mathbf{x}, y), h) := \log(1 + \exp(-y \cdot h(\mathbf{x}))). \quad (1)$$

Like the hinge loss, the logistic loss depends on the margin $y \cdot h(\mathbf{x})$ (see 0/1 loss), and after division by $\log 2$ it is a convex upper bound on the 0/1 loss, with equality at margin 0 (see Fig. 1).

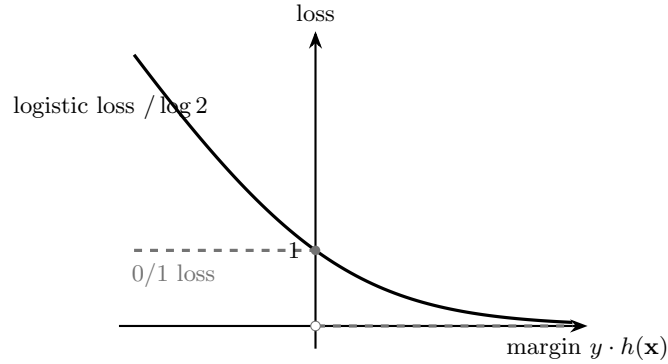


Figure 1: The logistic loss divided by $\log 2$ (solid) and the 0/1 loss (dashed) as functions of the margin $y \cdot h(\mathbf{x})$. The rescaled logistic loss lies on or above the 0/1 loss for every margin, with equality at margin 0; the rescaling is needed because $\log(1 + \exp(0)) = \log 2 < 1$

Unlike the hinge loss, the logistic loss is everywhere differentiable. It is therefore well suited for gradient-based optimization methods such as GD. Empirical risk minimization (ERM) with a linear hypothesis and the logistic loss yields logistic regression. For example, in medical diagnosis, logistic regression trained with the logistic loss predicts whether a patient has a disease ($y = +1$) or not ($y = -1$) based on clinical features.

Cross References

0/1 loss, hinge loss, classifier, logistic regression, GD, convex, binary classification

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.