

binary classification

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

two-class classification

Abstract

Binary classification is the machine learning (ML) problem of predicting which of two labels a data point carries, from its features. The label space holds two elements, written either $\{-1, +1\}$ or $\{0, 1\}$; the margin-based loss functions assume the first convention. A classifier partitions the feature space into two decision regions separated by a decision boundary, which is a hyperplane for a linear classifier. The natural loss function is the 0/1 loss, whose average over a dataset is 1 minus the accuracy, but it is not convex in the model parameters, so practical methods minimize a convex surrogate that upper bounds it: the logistic loss for logistic regression, the hinge loss for the support vector machine (SVM). Because the two kinds of mistake usually carry different costs, the confusion matrix is reported beside the accuracy.

Definition

A weather service that answers whether the maximum temperature of the coming day will reach 25 degrees Celsius returns one of two answers, yes or no, and nothing in between. The same shape of question is asked of an email (spam or not), of a credit application (default or not) and of a biopsy image (malignant or benign). Binary classification is the machine learning (ML) problem behind all of them: predict which of two labels a data point carries, from its features.

Binary classification is classification with a label space $\mathcal{Y}$ that holds two elements. The two elements are usually written $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$. The choice is a convention and not a modelling decision, but it must be fixed before a loss function is written down: the logistic loss and the hinge loss are defined through the margin $y \cdot h(\mathbf{x})$ and therefore assume $\{-1, +1\}$. A classifier is a hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$, and it partitions the feature space into two decision regions, one per label, separated by a decision boundary. For a linear classifier the decision boundary is a hyperplane $\mathbf{w}^\top \mathbf{x} + w_0 = 0$, and the label predicted for $\mathbf{x}$ is the side of that hyperplane on which $\mathbf{x}$ lies (Bishop, 2006, Sect. 4.1.1). A data point that lands on the side of the other label is misclassified and incurs a 0/1 loss of 1 (Fig. 1).

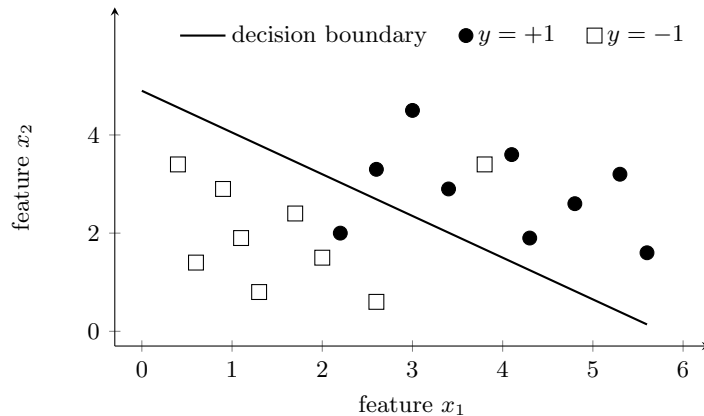


Figure 1: Binary classification with two features. The decision boundary of a linear classifier is a hyperplane, here a straight line, and it cuts the feature space into the two decision regions of the labels $y = +1$ (filled circles) and $y = -1$ (open squares). Two of the 18 data points lie on the far side of the line from their own label: they are misclassified, each contributes 1 to the 0/1 loss, and the average 0/1 loss is therefore $2/18 \approx 0.11$

The natural loss function for binary classification is the 0/1 loss, which charges 1 for a wrong label and 0 for a right one, so that its average over a dataset is the fraction of data points misclassified, i.e., 1 minus the accuracy. The 0/1 loss is not convex in the model parameters and empirical risk minimization (ERM) with it is computationally hard, so practical methods minimize a convex surrogate that upper bounds it (Shalev-Shwartz and Ben-David, 2014, Sect. 12.3): the logistic loss, which gives logistic regression, or the hinge loss, which gives the support vector machine (SVM).

A single number is a thin summary of a binary classifier. The two kinds of mistake, a wrong yes and a wrong no, generally carry different costs, and on imbalanced data the accuracy of the constant answer is already high (baseline). Both are reasons to report the confusion matrix, and the precision, recall and F_1 score computed from it, beside the accuracy.

For example, a spam filter reads the features of an email and returns $y = +1$ for spam and $y = -1$ for legitimate mail. A wrong $+1$ hides a legitimate email from the user while a wrong -1 only delivers one more piece of spam, so the two mistakes are not interchangeable and the classifier is tuned to keep the first rare.

Cross References

classification, classifier, decision boundary, 0/1 loss, logistic loss, logistic regression, accuracy, confusion matrix, imbalanced data, baseline

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.
2. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.