0/1 loss

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

misclassification loss, zero-one loss

Abstract

The 0/1 loss of a classifier on a data point is 1 if the predicted label differs from the true label and 0 if the two agree, so it is the indicator function of a misclassification. Its average over a dataset is the fraction of misclassified data points, which is one minus the accuracy, and its expectation is the probability of a misclassification, whose smallest value over all classifiers is the Bayes risk. For binary classification it depends on the data point only through the margin, and it uses the real-valued hypothesis only through its sign. As a function of the model parameters it is piecewise constant with a zero gradient wherever the gradient exists, and finding the linear classifier with the fewest misclassifications on a training set is NP-hard, so it is not used as the objective function of training. Empirical risk minimization (ERM) uses a convex surrogate instead, the hinge loss or the logistic loss divided by $\ln 2$, each an upper bound on the 0/1 loss whose minimizers over all functions have the sign of the classifier with the Bayes risk. The 0/1 loss remains the loss by which a classifier is evaluated on a test set.

Definition

A frost warning for Krems an der Donau is issued in the evening and is either right or wrong the next morning: frost came, or it did not. Counting a wrong warning as 1 and a right one as 0 is the 0/1 loss. In general, consider a data point with features $\mathbf{x} \in \mathcal{X}$ and a label y from a finite label space $\mathcal{Y}$, and a classifier $h : \mathcal{X} \to \mathcal{Y}$ that predicts the label. The 0/1 loss incurred by h on the data point $(\mathbf{x}, y)$ is (Goodfellow et al., 2016, Sect. 5.1.2; Shalev-Shwartz and Ben-David, 2014, Sect. 3.2.2)

$$L^{(0/1)}\left((\mathbf{x}, y), h\right) := \begin{cases} 0 & \text{if } h(\mathbf{x}) = y, \\ 1 & \text{if } h(\mathbf{x}) \neq y. \end{cases} \tag{1}$$

It is the indicator function of a misclassification, and it takes the same two values whether the label space holds two elements or many.

The average of the 0/1 loss (1) over a dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$ is the fraction of misclassified data points,

$$\frac{1}{m} \sum_{r=1}^{m} L^{(0/1)} \left((\mathbf{x}^{(r)}, y^{(r)}), h \right) = 1 - \operatorname{acc}(h \mid \mathcal{D}), \tag{2}$$

which is one minus the accuracy of h on $\mathcal{D}$. On a training set, the average (2) is the training error of h under the 0/1 loss, and on a test set it is the error rate reported for a classification method. For data points that are realizations of independent and identically distributed (i.i.d.) random variables (RVs) with a common probability distribution, the expectation of the 0/1 loss is the probability that h misclassifies a data point, which is the risk of h under the 0/1 loss (Shalev-Shwartz and Ben-David, 2014, Sect. 3.2.2). The risk is smallest for the classifier that assigns each $\mathbf{x}$ the label value with the largest conditional probability given $\mathbf{x}$, and that smallest risk is the Bayes risk (Bishop, 2006, Sect. 1.5.1).

For binary classification with $\mathcal{Y} = \{-1, +1\}$, a classifier is usually built from a real-valued hypothesis $h : \mathcal{X} \to \mathbb{R}$ that predicts $\widehat{y} = \operatorname{sign}(h(\mathbf{x}))$. The 0/1 loss (1) of this classifier depends on the data point only through the margin $y \cdot h(\mathbf{x})$,

$$L^{(0/1)}\left((\mathbf{x}, y), h\right) = \begin{cases} 0 & \text{if } y \cdot h(\mathbf{x}) > 0, \\ 1 & \text{if } y \cdot h(\mathbf{x}) \leq 0. \end{cases} \tag{3}$$

A positive margin is a correct prediction, and a data point on the decision boundary $h(\mathbf{x}) = 0$ counts as misclassified. The margin form (3) uses h only through its sign, so multiplying h by any positive number leaves the 0/1 loss of every data point unchanged.

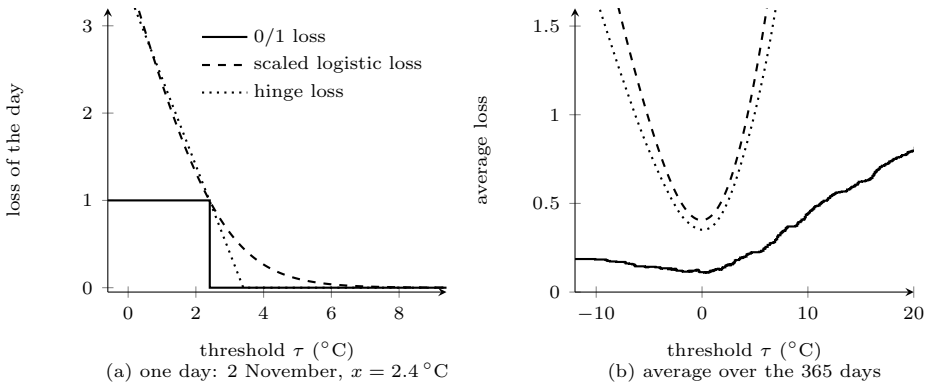
The 0/1 loss is the natural loss for classification, and empirical risk minimization (ERM) with it would take the training error (2) as the objective function to minimize over a hypothesis space $\mathcal{H}$. Two obstacles stand in the way. First, as a function of the model parameters of h, the average 0/1 loss is piecewise constant: it changes only when the decision boundary crosses a data point, and between two such crossings its gradient is zero. Gradient descent (GD) and its variants therefore receive no direction of improvement from it, and the average is neither convex nor smooth (Shalev-Shwartz and Ben-David, 2014, Sect. 12.3). Second, the problem is hard even for methods that use no gradient. If some linear classifier classifies every data point of the training set correctly, a linear program finds one (Shalev-Shwartz and Ben-David, 2014, Sect. 9.1). Without that guarantee, finding the linear classifier with the fewest misclassifications on a given training set is NP-hard, and so is finding one whose number of correct predictions is within a fixed factor of the largest achievable (Ben-David et al., 2003; Shalev-Shwartz and Ben-David, 2014, Sect. 8.2.2).

Fig. 1 shows both obstacles in the frost warning of the opening. Each of the $m = 365$ days of 2024 at Krems is a data point whose single feature x is the lowest temperature of the day and whose label is $y = +1$ if the lowest temperature of the following day is below 0 °C (frost) and $y = -1$ otherwise. Frost follows

68 of the 365 days. The warning compares the day's lowest temperature with a threshold τ: the hypothesis $h(x) = \tau - x$ predicts frost when it is positive, and the single model parameter τ selects a classifier from the hypothesis space of all thresholds. The 0/1 loss of one day is a step in τ, from 1 to 0 at the day's temperature, and the average over the year is a staircase with 195 steps, one at each distinct temperature. Its smallest value, $40/365 = 0.110$, is attained by every threshold strictly between 0.2 and 0.3 and between 0.4 and 0.6 °C: the set of minimizers is not a single point, and the slope of the staircase is zero at every threshold that is not a jump. A warning that never predicts frost has average 0/1 loss $68/365 = 0.186$, and one that always predicts frost has 0.814.

Practical classification methods replace the 0/1 loss in ERM by a surrogate loss function that is convex in $h(\mathbf{x})$ and an upper bound on the 0/1 loss (Shalev-Shwartz and Ben-David, 2014, Sect. 12.3). The hinge loss $\max\{0, 1 - y \cdot h(\mathbf{x})\}$, which gives the support vector machine (SVM), lies on or above the 0/1 loss for every margin (Shalev-Shwartz and Ben-David, 2014, Sect. 12.3). The logistic loss $\ln(1 + \exp(-y \cdot h(\mathbf{x})))$, which gives logistic regression, equals $\ln 2 \approx 0.69$ at margin 0 and is an upper bound only after division by $\ln 2$. Scaling a surrogate so that it passes through 1 at margin 0 is the usual convention when loss functions are compared (Bishop, 2006, Fig. 7.5; Hastie et al., 2009, Sect. 10.6). Because the surrogate is an upper bound, the risk of any hypothesis under the 0/1 loss is at most its risk under the surrogate, so a hypothesis with a small surrogate risk misclassifies rarely (Shalev-Shwartz and Ben-David, 2014, Sect. 12.3). Both surrogates are also calibrated for classification: for a convex function of the margin that is differentiable at 0 with a negative derivative there, every sequence of hypotheses whose surrogate risk tends to its smallest possible value has a 0/1 risk that tends to the Bayes risk (Bartlett et al., Thms. 1 and 6). Over all functions of $\mathbf{x}$, the hinge loss is minimized by the sign of $\mathbb{P}(y = +1 \mid \mathbf{x}) - 1/2$ and the logistic loss by the logarithm of the odds of $y = +1$, and the sign of either minimizer is the classifier with the Bayes risk (Hastie et al., 2009, Table 12.1). Unlike the 0/1 loss, a surrogate changes when h is multiplied by a positive number, so it depends on the scale of h and not only on its sign.

For the frost warning, the average logistic loss divided by $\ln 2$ and the average hinge loss are convex functions of the threshold that lie above the staircase everywhere (Fig. 1). Each has a single minimizer, $\tau = -0.04$ °C for the logistic loss and 0.06 °C for the hinge loss, and at both the average 0/1 loss is $41/365 = 0.112$, one day more than the smallest value of the staircase. At the minimizer of the logistic loss, the average logistic loss divided by $\ln 2$ is 0.405 for h, 0.333 for $h/2$, and 0.670 for $2h$, while the average 0/1 loss is the same for all three.



(a) one day: 2 November, $x = 2.4\,°\mathrm{C}$ (b) average over the 365 days

Figure 1: The frost warning $h(x) = \tau - x$ at Krems as a function of its threshold τ. (a) The 0/1 loss (solid curve), the logistic loss divided by $\ln 2$ (dashed curve), and the hinge loss (dotted curve) of one day, 2 November 2024, whose lowest temperature was $x = 2.4\,°\mathrm{C}$ and which was followed by frost. The 0/1 loss jumps from 1 to 0 at $\tau = 2.4$, where both surrogates equal 1, and the surrogates lie on or above the 0/1 loss at every threshold. (b) The averages of the same three losses over the 365 days of 2024. The average 0/1 loss is a staircase with 195 steps and its smallest value 0.110 just above $0\,°\mathrm{C}$; the two averaged surrogates are convex curves with one minimizer each, both within $0.1\,°\mathrm{C}$ of $0\,°\mathrm{C}$. Data generated by `pythondemos/zerooneloss.py`

A small average 0/1 loss on the training set is not the goal of training but a means to a small risk, and the two can differ. Fig. 2 shows a training set on which a linear classifier h' misclassifies one data point while a classifier h'' whose decision boundary bends around that data point misclassifies none. A small shift of a feature vector near this decision boundary flips the prediction of h'', so its zero training error does not indicate good generalization; this is overfitting. Restricting the hypothesis space to linear classifiers is a form of regularization, the route that shrinks the hypothesis space.

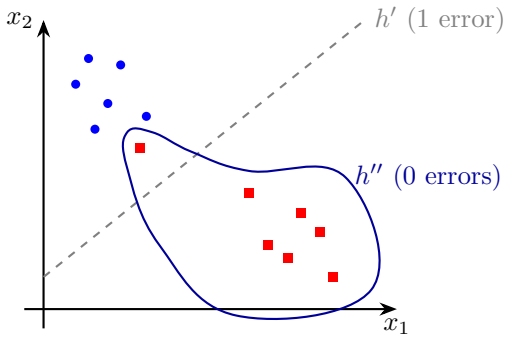


Figure 2: Two classifiers on one training set of data points with two features, drawn as blue circles and red squares by their label value. The linear classifier h' (dashed line) misclassifies one data point, the red square among the blue circles, so its average 0/1 loss on the training set is 1/13. The classifier h'' (solid curve) bends its decision boundary around that data point and reaches an average 0/1 loss of zero, but a small shift of a feature vector near this decision boundary flips its prediction

Cross References

loss, loss function, classification, binary classification, classifier, accuracy, training error, risk, Bayes risk, margin, hinge loss, logistic loss, ERM, SVM, logistic regression, confusion matrix, imbalanced data

References

1. Bartlett et al.. Convexity, Classification, and Risk Bounds. Journal of the American Statistical Association, 101(473), pp. 138–156.

2. Ben-David et al. (2003). On the Difficulty of Approximately Maximizing Agreements. Journal of Computer and System Sciences, 66(3), pp. 496–514.

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

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

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

6. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.