squared error 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

squared loss, quadratic loss

Abstract

The squared error loss of a hypothesis on a data point with a numeric label is the square of the difference between the label and the prediction that the hypothesis delivers for the features of the data point. Its average over a training set is the training error, also called the mean squared error (MSE), and minimizing that average over a linear model is least squares, whose solution the normal equations give in closed form. Up to an increasing affine transformation, the squared error loss is the negative log-likelihood function of the label under Gaussian noise around the prediction, so least squares is maximum likelihood estimation. The hypothesis with the smallest risk under the squared error loss is the conditional expectation of the label given the features, and the risk of any other hypothesis exceeds it by the expected squared difference from that conditional expectation. The squared error loss is convex and differentiable in the prediction with a slope proportional to the error, so a single outlier moves the learned hypothesis in proportion to its error, while the absolute error loss and the Huber loss bound that displacement.

Definition

A weather service predicts the maximum daytime temperature at Krems an der Donau from the morning minimum of the same day. Each day is a data point with the feature x, the morning minimum, and the label $y \in \mathbb{R}$, the maximum. A hypothesis h delivers the prediction $\widehat{y} = h(x)$, and the maximum measured in the afternoon says by how many degrees that prediction was off. For a data point with feature vector $\mathbf{x} \in \mathcal{X}$ and numeric label $y \in \mathbb{R}$, consider a hypothesis $h : \mathcal{X} \to \mathbb{R}$. The squared error loss it incurs is the square of the difference between the label and its prediction (Hastie et al., 2009, Sect. 2.4),

$$L\left((\mathbf{x}, y), h\right) := \big(y - h(\mathbf{x})\big)^2. \tag{1}$$

On 8 December 2024 the morning minimum at Krems was 3.0 degrees and the maximum 4.6 degrees. The linear hypothesis $\widehat{h}(x) = 9.62 + 1.01x$ learned

below predicts 12.66 degrees for that morning, so its squared error loss (1) on that day is $(4.6 - 12.66)^2 \approx 65.0$ squared degrees. The squared error loss is measured in the squared unit of the label, vanishes exactly when the prediction is correct, and is the same for a prediction too high as for one too low by the same amount.

As a function of the prediction $\widehat{y}$, the squared error loss (1) is a parabola with its minimum at $\widehat{y} = y$, as Fig. 1 shows. It is convex and differentiable, with derivative $-2(y - \widehat{y})$: its slope grows in proportion to the error. For a linear model $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, the squared error loss is a convex and differentiable function of the model parameters $\mathbf{w}$ as well, which is what makes gradient descent (GD) applicable. The absolute error loss $|y - \widehat{y}|$ is convex too but non-smooth at $\widehat{y} = y$, and its slope is ± 1 whatever the size of the error. The Huber loss with threshold $\delta > 0$ is quadratic for a small error and linear for a large one (Hastie et al., 2009, Eq. (10.23)). In terms of the error $e := y - h(\mathbf{x})$, it is

$$\begin{cases} e^2 & \text{for } |e| \le \delta, \\ 2\delta|e| - \delta^2 & \text{otherwise,} \end{cases} \tag{2}$$

so it agrees with the squared error loss for an error up to δ and its slope is capped at 2δ beyond (Huber, 1992).

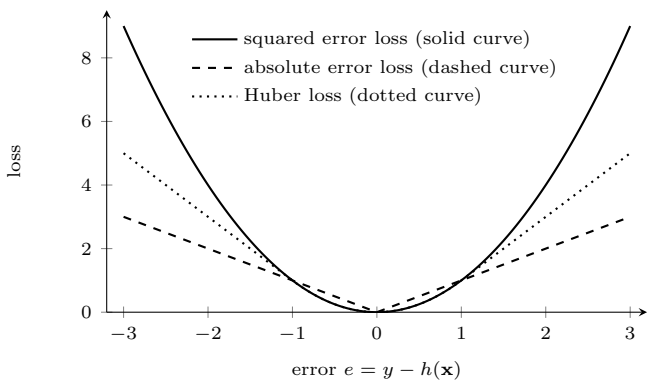


Figure 1: The squared error loss (1) (solid curve), the absolute error loss (dashed curve), and the Huber loss (2) with threshold $\delta = 1$ (dotted curve) as functions of the error $e = y - h(\mathbf{x})$. The three agree at an error of ± 1; beyond it, only the squared error loss keeps steepening

The average of the squared error loss over a training set $\mathcal{D}^{(\text{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}$,

$$\frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - h(\mathbf{x}^{(r)})\right)^2, \tag{3}$$

is the training error of h, also called its mean squared error (MSE), and empirical risk minimization (ERM) selects from a model the hypothesis with the smallest

training error. With a linear model, ERM under the squared error loss is linear regression by least squares (Legendre, 1805). The training error (3) then reads $(1/m)\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$, with the feature matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ whose rows are the feature vectors and the label vector $\mathbf{y} \in \mathbb{R}^m$. Its gradient vanishes exactly at the solutions of the normal equations $\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$, and when $\mathbf{X}$ has full column rank the unique solution is (Bishop, 2006, Sect. 3.1.1)

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{4}$$

The closed form (4) needs no iteration, so there is no update whose fixed point the solution would be. When $\mathbf{X}^\top \mathbf{X}$ is too large to invert, GD on the training error (3) is used instead, and its fixed points are again the solutions of the normal equations (see linear regression).

Take as training set every 18th day of 2024 at Krems, 21 days, with the constant feature 1 appended to the morning minimum so that $\widehat{\mathbf{w}}$ holds an intercept and a slope. The closed form (4) gives the line $\widehat{h}(x) = 9.62 + 1.01x$ of Fig. 2(a), with training error 12.75 squared degrees, or 3.57 degrees after taking the root. The best constant prediction under the squared error loss is the sample mean of the labels, 17.14 degrees, and its training error is their sample variance, 62.05: the average squared error of a constant c is that variance plus $(c - 17.14)^2$. Under the absolute error loss the best constant is the median of the labels instead, 15.80 degrees.

The squared error loss is the loss behind maximum likelihood estimation when the label is modeled as the prediction plus Gaussian noise. Let $y = h(\mathbf{x}) + \varepsilon$ with a Gaussian random variable (Gaussian RV) ε of mean zero and variance σ^2, so that y given $\mathbf{x}$ has the probability density function (pdf) $\mathcal{N}\left(h(\mathbf{x}), \sigma^2\right)$. The negative logarithm of that pdf at the observed label is (Bishop, 2006, Sect. 3.1.1, Eqs. (3.11)–(3.12))

$$-\ln \mathcal{N}\left(h(\mathbf{x}), \sigma^2\right)(y) = \frac{1}{2\sigma^2}\left(y - h(\mathbf{x})\right)^2 + \frac{1}{2}\ln\left(2\pi\sigma^2\right), \tag{5}$$

an increasing affine function of the squared error loss (1). For data points that are realizations of independent and identically distributed (i.i.d.) random variables (RVs), the negative log-likelihood function of the training set is the sum of the terms (5) over its data points, which is $m/(2\sigma^2)$ times the training error (3) plus a constant. The maximum likelihood estimate of $\mathbf{w}$ is therefore the least squares solution (4), whatever σ^2 is, and the maximum likelihood estimate of σ^2 is the training error itself, 12.75 for the 21 days. The logistic loss is the negative log-likelihood function for a binary label, and Laplacian instead of Gaussian noise gives the absolute error loss (Murphy, 2012, Sect. 7.4).

The risk of a hypothesis is its expected loss under a probabilistic model with a joint probability distribution $\mathbb{P}^{(\mathbf{x},y)}$ of the feature vector and the label. Under the squared error loss, the risk is minimized by the conditional expectation $h^*(\mathbf{x}) = \mathbb{E}\{y \mid \mathbf{x}\}$, the Bayes estimator for this loss (Bishop, 2006, Sect. 1.5.5, Eq. (1.89); Hastie et al., 2009, Sect. 2.4, Eq. (2.13)). The risk of any hypothesis

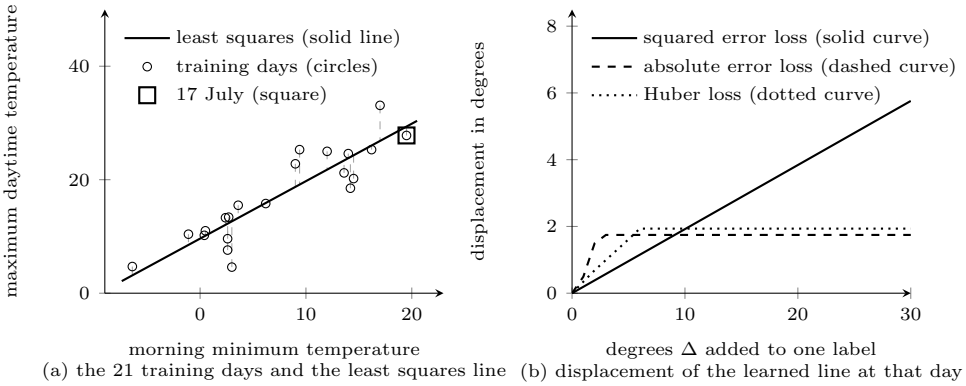
h splits into two parts (Bishop, 2006, Eq. (1.90)),

$$\mathbb{E}\big\{\big(y - h(\mathbf{x})\big)^2\big\} = \mathbb{E}\big\{\big(h^*(\mathbf{x}) - h(\mathbf{x})\big)^2\big\} + \mathbb{E}\big\{\big(y - h^*(\mathbf{x})\big)^2\big\}, \tag{6}$$

the expected squared difference between h and the conditional expectation, and the expected conditional variance of the label, which no hypothesis can reduce. For a hypothesis learned from a random training set, the first part of the decomposition (6) splits further into the squared bias and the variance of the learned prediction, the decomposition behind the bias–variance trade-off (Hastie et al., 2009, Sect. 7.3, Eq. (7.9)). Under the absolute error loss the Bayes estimator is the conditional median (Hastie et al., 2009, Eq. (2.18)). The two constant predictions above are the training set versions of these two facts, with the average over the 21 days in place of the expectation.

The growing slope of the parabola has a cost. A data point whose label is far from the prediction enters the gradient of the training error (3) in proportion to its error, so an outlier pulls the learned hypothesis toward itself without bound (Hastie et al., 2009, Sect. 10.6). For least squares the pull is explicit. Adding Δ to the label of the rth training day changes the solution (4) by $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{(r)}\Delta$. The prediction for that day changes by $\ell_r\Delta$, where $\ell_r = (\mathbf{x}^{(r)})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{(r)}$ is the leverage of the day.

Among the 21 days, the one with the warmest morning, 17 July 2024 with minimum 19.5 and maximum 27.8 degrees, has leverage 0.192. Each degree added to its label moves the least squares line at that morning by 0.192 degrees, which is 5.76 degrees for $\Delta = 30$, and the slope of the line rises from 1.01 to 1.37 (Fig. 2(b), solid curve). Under the absolute error loss, whose ERM solution is least absolute deviation regression, the line moves by at most 1.75 degrees, and not at all once Δ exceeds 3 (dashed curve). The slope of the absolute error loss is ± 1, so once the corrupted day is above the line, its term in the optimality condition no longer depends on the size of its error. The Huber loss with $\delta = 3$ moves its line by 1.93 degrees and not at all beyond $\Delta = 6$ (dotted curve), since its slope is capped. This bounded pull is what robustness to outliers means for a loss. The absolute error loss is non-smooth, and the Huber loss adds the threshold δ as a hyperparameter; the squared error loss has neither drawback, and no such bound.



(a) the 21 training days and the least squares line (b) displacement of the learned line at that day

Figure 2: (a) Every 18th day of 2024 at Krems, the 21 training days (circles), and the line learned by least squares (solid line): the squared error loss (1) of a day is the squared length of its dashed segment, and the training error (3) averages these over the 21 days. The square marks 17 July, whose label is corrupted in (b). (b) The displacement of the learned line at the morning minimum of that day when Δ degrees are added to its label: the least squares line (solid curve) moves by 0.192Δ without bound, while the line learned under the absolute error loss (dashed curve) and under the Huber loss with threshold $\delta = 3$ (dotted curve) stop moving once the corrupted day lies far enough above them. Data generated by `pythondemos/sqerrloss.py`

When the label is a vector, the squared error loss is the squared Euclidean norm of the difference between label and prediction. The clustering error of k-means, which compares a data point with the cluster centroid that stands in for it, and the reconstruction error of principal component analysis (PCA) and of an autoencoder are of this form, with the feature vector itself in the place of the label.

Cross References

loss, loss function, absolute error loss, Huber loss, logistic loss, 0/1 loss, training error, MSE, ERM, least squares, linear regression, normal equations, GD, maximum likelihood, Gaussian RV, risk, Bayes estimator, conditional expectation, bias–variance trade-off, outlier, robustness, least absolute deviation regression, Huber regression, clustering error

References

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

2. Legendre (1805). Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot, Paris.

3. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

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

5. Huber (1992). Robust estimation of a location parameter. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 492–518. Springer-Verlag.