mean squared error (MSE)

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

Abstract

The mean squared error (MSE) of a hypothesis is the average of its squared error loss, the squared difference between the label of a data point and the prediction for it. The average is either an expectation under a probabilistic model for the data, the independent and identically distributed assumption (i.i.d. assumption), which makes the MSE the risk under the squared error loss, or the average over the data points of a dataset, which makes it the empirical risk, the training error or the validation error. The two readings coincide for the empirical distribution of the dataset, which assigns the same probability to each of its data points. When the i.i.d. assumption holds, the average over a dataset estimates the expectation and improves with the sample size. The root MSE, its square root, is measured in the unit of the label.

Definition

Consider a hypothesis h that predicts the maximum daytime temperature of a day at Krems an der Donau (Austria) from its morning minimum. A weather service that uses it reports one number for how far the predictions are off on average, and the MSE is that number: the average of the squared error loss, the squared difference between the label of a data point and the prediction for it. The word *average* has two readings here, and the MSE is defined in both.

The first reading needs a probabilistic model for how the data come about. The independent and identically distributed assumption (i.i.d. assumption) is the simplest one: it treats the daily weather measurements as realizations of independent and identically distributed (i.i.d.) random variables (RVs), so that each day, a data point with feature vector $\mathbf{x}$ and label y, is an independent draw from one probability distribution $\mathbb{P}$ that does not change from day to day. Under this model, the MSE of h is the expectation of its squared error loss,

$$\bar{L}\big(h\big) := \mathbb{E}\big\{\big(y - h(\mathbf{x})\big)^2\big\}, \tag{1}$$

which is the risk of h under the squared error loss (Hastie et al., 2009, Sect. 2.4). The expectation (1) is taken over the probability distribution $\mathbb{P}$ of the data point, which the i.i.d. assumption postulates but does not specify: in a machine learning (ML) application, $\mathbb{P}$ is unknown. What can be computed is the MSE

on a dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$ of sample size m, the average of the squared error loss over its data points,

$$\widehat{L}\big(h|\mathcal{D}\big) := \frac{1}{m}\sum_{r=1}^{m}\big(y^{(r)} - h(\mathbf{x}^{(r)})\big)^2, \tag{2}$$

which is the empirical risk of h on $\mathcal{D}$ under the squared error loss. It is the training error when $\mathcal{D}$ is the training set and the validation error when $\mathcal{D}$ is the validation set. For a linear model learned by least squares on the 365 days of 2023 at Krems, the MSE (2) is 21.1 on those days and 19.2 on the 366 days of 2024, in squared degrees. Its square root, the root MSE of 4.4 degrees, has the unit of the label.

The two definitions are one definition applied to two probability distributions. The empirical distribution $\mathbb{P}^{(\mathcal{D})}$ of the dataset $\mathcal{D}$ assigns the probability $1/m$ to each of its data points. The expectation of the squared error loss under $\mathbb{P}^{(\mathcal{D})}$ is therefore a sum of m terms with weight $1/m$ each,

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}^{(\mathcal{D})}}\big\{\big(y - h(\mathbf{x})\big)^2\big\} = \widehat{L}\big(h|\mathcal{D}\big). \tag{3}$$

By the identity (3), the average (2) is the expectation (1) with the empirical distribution of $\mathcal{D}$ in place of $\mathbb{P}$. Every statement about the MSE as an expectation then holds for the average with $\mathbb{P}^{(\mathcal{D})}$ in that role, such as its split into the variance of the label around its conditional mean and the squared deviation of h from that mean (Bishop, 2006, Sect. 1.5.5). Fig. 1(a) draws the empirical distribution of the squared errors of the 366 days of 2024. Its mean is the MSE 19.2, while its median is 9.1 and two thirds of the days have a loss below the MSE: a few far-off days weigh heavily in a squared average. Conversely, when the data points of $\mathcal{D}$ are realizations of i.i.d. RVs with probability distribution $\mathbb{P}$, the average (2) is an estimate of the expectation (1) that improves with the sample size m by the law of large numbers. Fig. 1(b) shows this for the days of 2024 in a random order: over the first 30 days, the average lies between 12.6 and 26.7 in nine of ten orders, over the first 180 days between 17.0 and 21.3. How far the training error can lie from the risk is the subject of generalization.

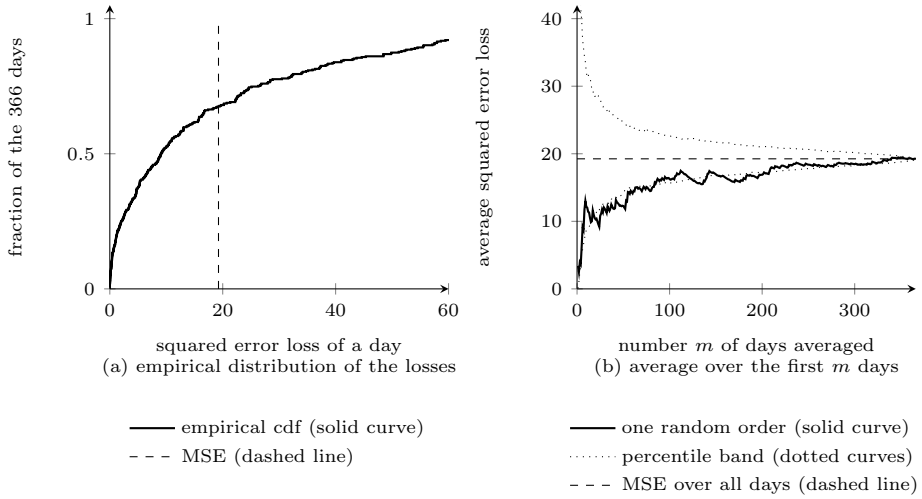


Figure 1: The squared error loss of a linear model, learned by least squares on the days of 2023 at Krems, on the 366 days of 2024. (a) The empirical cdf of the 366 losses (solid curve), cut at 60 squared degrees with 29 days beyond the cut, and their mean, the MSE (2) (dashed line). (b) The average of the losses over the first m days of 2024 in one random order (solid curve), the band between the 5th and the 95th percentile over 500 orders, which holds this average for nine orders in ten (dotted curves), and the MSE over all 366 days (dashed line). Data generated by `pythondemos/mse.py`

Squaring the deviation is what makes a far-off day weigh heavily, and that is the MSE's main limitation. A single mismeasured label enters the average in proportion to the square of its deviation, so it can dominate the number reported for a whole year. The mean absolute error (MAE), the average of the absolute error loss instead of the squared error loss, weighs each day in proportion to its deviation alone. The difference shows in what each average asks a constant prediction to be: the MSE is smallest at the mean of the labels, the MAE at their median (Hastie et al., 2009, Sect. 10.6). On the 366 days of 2024 at Krems, these are 17.91 and 17.60 degrees. A decimal point in the wrong place moves one of those 366 labels by 5, 10, 20, or 50 degrees. The mean then moves by 0.014, 0.027, 0.055, and 0.137 degrees, the error divided by the sample size. The median moves by at most 0.1 degrees, the step to the neighboring label, however large the error becomes. The MSE is therefore the wrong summary for a record with outliers or a long-tailed error distribution, and the MAE or a least absolute deviation regression is the usual answer (Hastie et al., 2009, Sect. 10.6). What the MSE buys in return is the mean it points at: it is differentiable everywhere, which is what lets least squares have a closed form.

Cross References

squared error loss, risk, empirical risk, training error, validation error, empirical distribution, expectation, probabilistic model, i.i.d. assumption, law of large numbers, generalization, MAE, absolute error loss, outlier, linear regression, loss

References

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

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