baseline

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

A baseline is a reference level against which the performance of a trained model is compared. This comparison allows verifying whether the achieved average loss is satisfactory, i.e., whether the learned hypothesis is already close to optimal. Baselines can be obtained from human performance, from an existing machine learning (ML) method, or from an accurate probabilistic model for the data generation. Given such a probabilistic model, the smallest achievable risk is the Bayes risk, incurred by the Bayes estimator of the label given the features. In practice, however, the Bayes estimator is infeasible for at least two reasons. First, the probability distribution of the data generation process might be impossible to determine. Second, computing the Bayes estimator might be computationally too expensive.

Definition

Consider some machine learning (ML) method that produces a learned hypothesis (or trained model) $\widehat{h} \in \mathcal{H}$. Its quality is typically evaluated by computing the average loss on a test set. But how can it be verified that the measured performance is satisfactory? In other words, is the learned hypothesis already so close to optimal that there is little point in investing more resources (for data collection or computation) to improve it? To answer such questions, a baseline (or reference level) for the smallest loss achievable by any method is needed.

Sometimes a baseline can be read off a scatterplot such as Fig. 1, which shows weather records of the station Krems an der Donau: each of the 366 days of 2024 is a data point whose feature x is the minimum air temperature of the day and whose label y is its maximum air temperature. On the 15 days with x between 11.5 and 12.5 degrees Celsius, the label ranged from 13.2 to 33.4 degrees Celsius. Two of these days even share the same feature value $x = 11.8$, with labels 13.2 and 29.9. A hypothesis is a map, so it assigns both days the same prediction and incurs, on these two days, an average squared error loss of at least $(16.7/2)^2 \approx 69.7$. Averaging the variance of the label within all such one-degree bands yields 19.34 (in squared degrees Celsius) as an estimate for the smallest achievable average squared error loss. A hypothesis learned by linear regression incurs an average squared error loss of 19.17 on the 366 days, within one percent of that estimate: this hypothesis is already close to optimal.

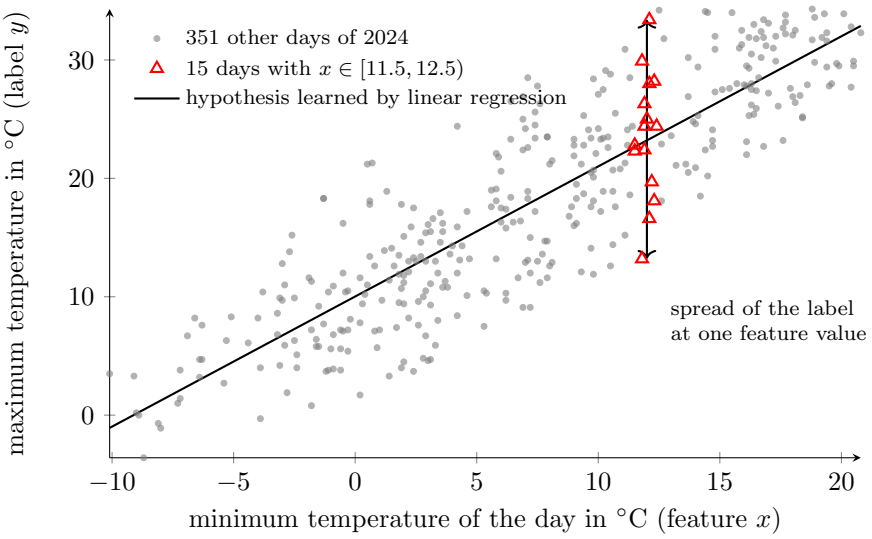

Figure 1: Daily temperatures at the weather station Krems an der Donau during 2024. Each data point is one day, with the minimum temperature as its feature x and the maximum temperature as its label y. Days with (nearly) the same feature value carry widely different label values; this spread limits the average loss achievable by any ML method that predicts the maximum temperature based on the minimum temperature. Data generated by `pythondemos/baseline.py`

A baseline might be obtained from human performance, e.g., the misclassification rate of dermatologists who diagnose cancer from visual inspection of skin (Salinas et al., 2024). For large language models (LLMs), baselines are often obtained from a benchmark, i.e., a fixed collection of test problems together with an evaluation protocol. As an example, the benchmark GSM8K measures the ability of an LLM to solve grade-school mathematics word problems (Cobbe et al., 2021).

Another source for a baseline is an existing, but for some reason unsuitable, ML method: it might be computationally too expensive for the intended ML application, yet its test set error can still serve as a baseline. A more principled source of a baseline is a probabilistic model. In many cases, given a probabilistic model $p(\mathbf{x}, y)$, it is possible to precisely determine the minimum achievable risk among any hypotheses (not even required to belong to the hypothesis space $\mathcal{H}$) (Lehmann and Casella, 1998).

This minimum achievable risk (referred to as the Bayes risk) is the risk of the Bayes estimator of the label y of a data point, given its features $\mathbf{x}$. For the squared error loss, the Bayes risk equals the variance of the label around its posterior mean, averaged over the features; the band-wise estimate in Fig. 1 approximates exactly this quantity.

Note that, for a given choice of loss function, the Bayes estimator (if it ex-

ists) is fully determined by the probability distribution $\mathbb{P}$ (Lehmann and Casella, 1998, Chap. 4). However, computing the Bayes estimator and Bayes risk presents two main challenges. First, the probability distribution $\mathbb{P}$ is unknown and must be estimated from observed data. Second, even if $\mathbb{P}$ were known, computing the Bayes risk exactly may be computationally infeasible (Cooper, 1990).

A widely used probabilistic model is the multivariate normal distribution $(\mathbf{x}, y) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ for data points characterized by numeric features and labels. Here, for the squared error loss, the Bayes estimator is given by the posterior mean $\mu_{y|\mathbf{x}}$ of the label y, given the features $\mathbf{x}$ (Gray, 2009; Lehmann and Casella, 1998). The corresponding Bayes risk is given by the posterior variance $\sigma^2_{y|\mathbf{x}}$ (see Fig. 2). The posterior variance can be read off the inverse of the covariance matrix: $\sigma^2_{y|\mathbf{x}}$ is the reciprocal of the diagonal entry of $\mathbf{C}^{-1}$ that corresponds to the label y (Bishop, 2006, Sect. 2.3.1). In practice, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{C}$ can be estimated from a dataset by the sample mean and the sample covariance matrix (Anderson, 2003).

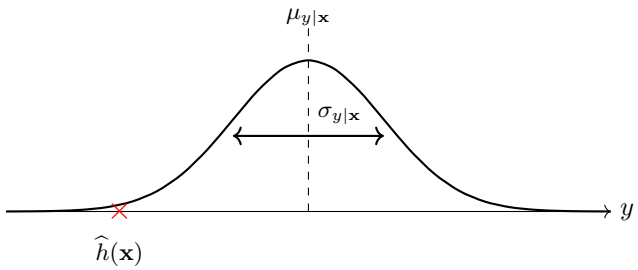

Figure 2: For data points that are realizations of independent and identically distributed (i.i.d.) random variables (RVs) with a common multivariate normal distribution, the minimum risk (under squared error loss) is obtained by the Bayes estimator $\mu_{y|\mathbf{x}}$ of the label y based on its features $\mathbf{x}$. The corresponding minimum risk is given by the posterior variance $\sigma^2_{y|\mathbf{x}}$. This quantity can serve as a baseline for the average loss of a trained model $\widehat{h}$

Cross References

Bayes risk, Bayes estimator, test set, probabilistic model, risk, accuracy, benchmark

References

1. Anderson (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., Wiley, Hoboken, NJ.

2. Bishop (2006). Pattern Recognition and Machine Learning. Springer

Science+Business Media, New York, NY, USA.

3. Cobbe et al. (2021). Training Verifiers to Solve Math Word Problems. arXiv preprint arXiv:2110.14168. URL: https://arxiv.org/abs/2110.14168.

4. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.

5. Lehmann and Casella (1998). Theory of Point Estimation. 2nd ed., Springer-Verlag, New York, NY, USA.

6. Salinas et al. (2024). A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. npj Digit. Med., 7(1), pp. Art. no. 125.

7. Cooper (1990). The computational complexity of probabilistic inference using Bayesian belief networks. Artif. Intell., 42(2–3), pp. 393–405.