



Why

We compare inductors by comparing the predictors they produce. We compare predictors by judging predictions.

Definition

A *loss* function is a nonnegative real-valued function on pairs which is zero only on repeated pairs. It need not be symmetric.

We use loss functions to judge a prediction in comparison to the recorded postcept. The first argument of the loss function is the predicted postcept and the second is the recorded (observed, true, recorded) postcept.

The *loss of a predictor on a pair* is the result of the loss function on the pair. Similarly, the *loss of a predictor on a dataset* of record pairs is the sum of the losses on the pairs of the dataset. Likewise, the *average loss* of a predictor is the loss of the predictor on the dataset divided by the size of the dataset. The average loss is also known as the *empirical risk* of the predictor on the dataset.

Notation

Let $(a, b) \in A \times B$; $A, B \neq \emptyset$. For a loss function $\ell : B \times B \rightarrow \mathbf{R}$ and predictor $f : A \rightarrow B$, the loss of f on (a, b) is $\ell(f(a), b)$. Let $s = ((a^1, b^1), \dots, (a^n, b^n))$ be a record sequence. The loss of f on s is $\sum_{k=1}^n \ell(f(a^k), b^k)$. The average loss of f on s is $(1/n) \sum_{k=1}^n \ell(f(a^k), b^k)$.

Training and test loss

Recall that s is a dataset of records pairs in $A \times B$. We call a predictor $f : A \rightarrow B$ an *interpolator* of the dataset s if, for each pair (a^i, b^i) in the dataset, $f(a^i) = b^i$. An interpolator achieves zero loss on the dataset it interpolates.

The rub is that an interpolator may have nonzero loss a record pair which is not contained in the dataset used to construct it. For this reason, it is common to consider two datasets. First, the one used to construct the predictor (the *training dataset*) and second, one used to evaluate the predictor (the *test dataset*, *validation dataset* or *evaluation dataset*).

We judge a predictor by its average loss on the test dataset. We call this the *test loss*, in contrast with the *train loss* obtained on the dataset used to construct the predictor. A predictor whose average test loss is much larger than its average train loss is said to be *overfit* to the train dataset.¹

Roughly speaking, we judge an inductor by some aggregation metric of the loss of predictors it produces on datasets.

¹Many authorities will refer to this as the “problem” or “danger” of overfitting.

