



Why

A natural and (conceptually) simple inductor is to select a predictor which achieves zero error on the training dataset.

Definition

Let (X, \mathcal{X}, μ) and $f : X \rightarrow Y$ be a probabilistic supervised data model. For a dataset $(x_1, y_1), \dots, (x_n, y_n)$ in $X \times Y$, the *empirical error* of a predictor $h : X \rightarrow Y$ is

$$(1/n) |\{i \in \{1, 2, \dots, n\} \mid h(x_i) \neq y_i\}|,$$

An *empirical error minimizer* is a predictor whose empirical error is minimal. Since we assume that the dataset is consistent, the correct labeling function is always an empirical error minimizer, and achieves zero. If the dataset is incomplete, however, there may be other empirical error minimizers.

A family of inductors $(A^n)_n$ is an *empirical error minimizer* if, for each $n \in N$, and each dataset $D = ((x_i, y_i))_{i=1}^n$, $A^n(D)$ is an empirical risk minimizer of D .

Overfitting and incutive bias

Since a dataset may be incomplete, an empirical risk minimizer may be a “poor” predictor. For example, let $A \subset X \subset \mathbf{R}^2$ and $Y = \{0, 1\}$. Define $f : X \rightarrow Y$ by $f(x) = 1$ if $x \in A$ and $f(x) = 0$ otherwise. Suppose $A \in \mathcal{X}$ and $\mu(A) = 1/2$.

For any training set $(x_1, y_1), \dots, (x_n, y_n)$ in $X \times Y$, the hypothesis $h : X \rightarrow Y$ defined by $h(x) = y_i$ if $x = x_i$ and $h(x) = 0$ otherwise achieves zero empirical error but has “actual” error $1/2$. The predictor is said to be *overfit* or to exhibit *overfitting*. It is said to fit the training dataset “too well.”

By adjusting the hypothesis class one may “guard against” overfitting. In this context, the constrained hypothesis class is called an *inductive bias*.

Other terminology

Other terminology for the empirical error includes *empirical risk*. For these reasons, the learning paradigm of selecting a predictor h to minimize the empirical risk is called *empirical risk minimization* or *ERM*.

