



## Why

We want to discuss an inductor’s performance on consistent (but possibly incomplete) datasets.

We take two steps. First, put a measure on the set of training sets and only consider high-measure subsets. Second, consider predictors performing well in some tolerance.

## Definition

Let  $(X, \mathcal{X}, \mu)$  be a probability space and  $(Y, \mathcal{Y})$  a measurable space. Let  $f : X \rightarrow Y$  measurable. We call the pair  $((X, \mathcal{X}, \mu), f)$  a *(supervised) probabilistic data model*.

We interpret  $\mu$  as the *data-generating distribution* or *underlying distribution* and  $f$  as the *correct labeling function*. Many authors refer to a supervised probabilistic data model as the *statistical learning (theory) framework*.

## Probable datasets

We put a measure on the set of datasets by using the product measure  $(X^n, \mathcal{X}^n, \mu^n)$ . We interpret this as a model for a training set of independent and identically distributed inputs.

For  $\delta \in (0, 1)$ ,  $\mathcal{S} \subset X^n$  is  $1 - \delta$ -*representative* if  $\mu^n(\mathcal{S}) \geq 1 - \delta$ . If  $\mathcal{S}$  is  $1 - \delta$ -representative for small  $\delta$ , we think of  $\mathcal{S}$  as a set of “probable” or “reasonable” datasets. We call  $\delta$  the *confidence parameter*.

## Predictor error

The *error* of (measurable)  $h : X \rightarrow Y$  (under  $\mu$  and  $f$ ) is

$$\text{error}_{\mu, f}(h) = \mu(\{x \in X \mid h(x) \neq f(x)\}).$$

We interpret this as the probability that the predictor mislabels a point. The *accuracy* of  $h$  is  $1 - \text{error}_{\mu, f}(h)$ .

Since  $(f, g) \mapsto \mu[f(x) \neq g(x)]$  is a metric on  $L^2(X, \mathcal{X}, \mu, Y)$  we can talk about the error as the “distance” from the correct label classifier. Thus we will say that  $\varepsilon \in (0, 1)$ , (measurable)  $h : \mathcal{X} \rightarrow \mathcal{Y}$   $\varepsilon$ -*approximates* the correct labeling function  $f$  if  $\text{error}(h) \leq \varepsilon$ . Roughly speaking, if  $\varepsilon \ll 1$ , the the error of the hypothesis is “fairly small.” We call  $\varepsilon$  the *accuracy parameter*, since the accuracy of such a predictor is  $1 - \varepsilon$ .

### Other terminology

A *hypothesis class* is a subset of the measurable functions from  $X \rightarrow Y$ . Other names for the error of a classifier include the *generalization error*, the *risk* or the *true error* or *loss*.

