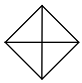


Why

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

Definition

We define a probability measure over the inputs. Let $(X, \mathcal{X}, \mu)$ be a probability space and $(Y, \mathcal{Y})$ a measurable space. Let $f : X \to Y$ measurable. We call the pair $((X, \mathcal{X}, \mu), f)$ a *probabilistic data-generation model.*

The idea is that to each instance $x \in X$, we associate a label $f(x) \in Y$. We call the law μ the *data-generating distribution* or *underlying distribution* and call f the *correct labeling function.* Many authors refer to a probabilistic data-generation model as the *statistical learning (theory) framework.*

Representative datasets

Let $(X, \mathcal{X}, \mu)$ and $f : X \to Y$ be a probabilistic data-generation model. We put a measure on the set of datasets X^n so that such "nonrepresentative" datasets have low measure. We use the product measure μ^n. We interpret this as a model of the inductor as acting on a training set of correctly labeled, independent and identically distributed inputs.

Let $0 < \delta < 1$. A set of datasets $\mathcal{S} \subset X^n$ is $1-\delta$-*representative* if $\mu^n(\mathcal{S}) \geq 1 - \delta$. We interpret this as follows: "the dataset we see is in $\mathcal{S}$ with high probability." Roughly speaking, if $\delta \ll 1$ then the set $\mathcal{S}$ is "fairly likely."

We think of $\mathcal{S}$ as the set of datasets for which our learning algorithm "succeeds," those datasets which are reasonable. In other words, the probability of a reasonable training set is $1-\delta$. We call δ the *confidence parameter*.

Predictor errors

The *error of a predictor* $h : X \to Y$ with respect to the probabilistic data-generation model is

$$\mu(\{x \in \mathcal{X} \mid h(x) \neq f(x)\}).$$

In other words, the error is the probability (w.r.t. the underlying distribution) that the predictor h mislabels a point. The error is measured with respect to the underlying distribution μ and correct labeling function f.

The *accuracy* of a predictor is

$$\mu(\{x \in X \mid h(x) = f(x)\}).$$

It is the measure of the set of points for which the predictor matches the correct labeling function: In other words, it is one minus the error of the predictor.

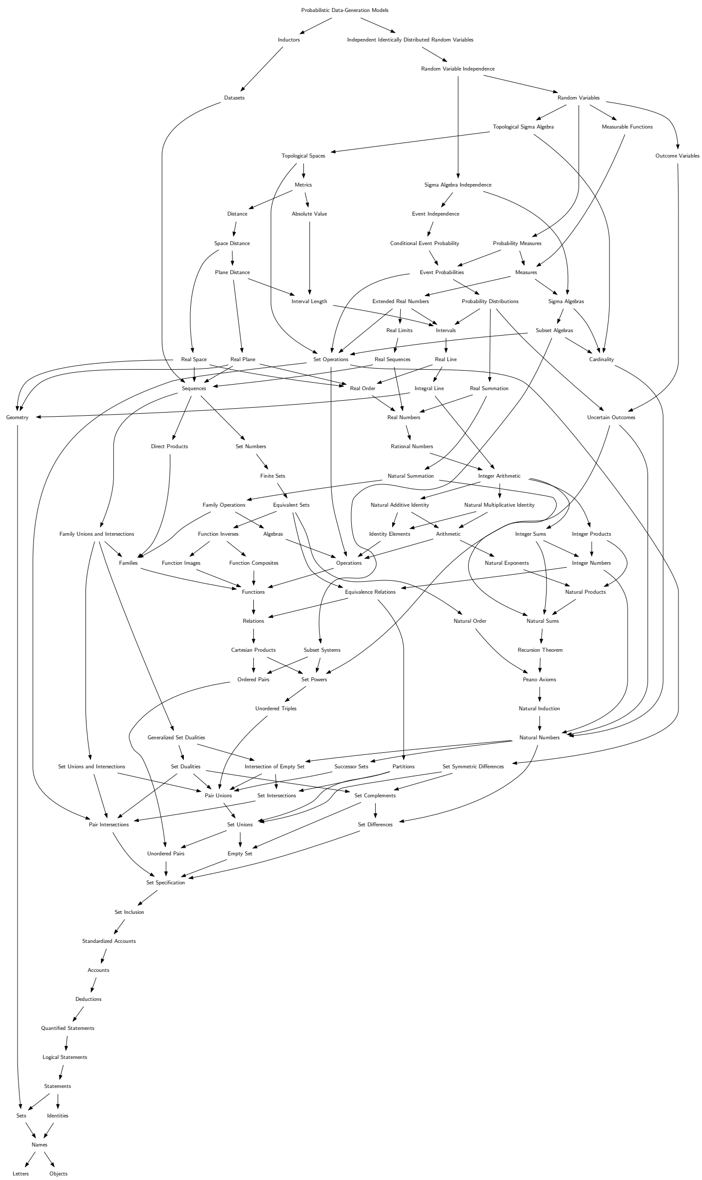
Inductor errors

Let $S = (x_1, y_1), \ldots, (x_n, y_n)$ a dataset of n paired records in $X \times Y$. Let $\mathcal{M}_{X \to Y}$ be the set of measurable functions from X to Y. The *dataset error of an inductor* $A : (X \times Y)^n \to \mathcal{M}_{X \to Y}$ on S is the error of the predictor $A(S)$.

Notice that if we have a probability measure on X^n, we can discuss the measure of a set of training sets. We may be interested in statements like: the measure of the set of training sets for which the error of the inductor is nonzero is small.

Other terminology

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


Probabilistic Data-Generation Models
Inductors
Independent Identically Distributed Random Variables
Random Variable Independence
Datasets
Random Variables
Topological Sigma Algebra
Measurable Functions
Topological Spaces
Outcome Variables
Metrics
Sigma Algebra Independence
Distance
Absolute Value
Event Independence
Space Distance
Conditional Event Probability
Probability Measures
Plane Distance
Event Probabilities
Measures
Interval Length
Extended Real Numbers
Probability Distributions
Sigma Algebras
Real Limits
Intervals
Subset Algebras
Real Space
Real Plane
Set Operations
Real Sequences
Real Line
Cardinality
Sequences
Real Order
Integral Line
Real Summation
Geometry
Real Numbers
Uncertain Outcomes
Direct Products
Set Numbers
Rational Numbers
Finite Sets
Natural Summation
Integer Arithmetic
Family Operations
Equivalent Sets
Natural Additive Identity
Natural Multiplicative Identity
Family Unions and Intersections
Function Inverses
Algebras
Identity Elements
Arithmetic
Integer Sums
Integer Products
Families
Function Images
Function Composites
Operations
Natural Exponents
Integer Numbers
Functions
Equivalence Relations
Natural Products
Relations
Natural Order
Natural Sums
Cartesian Products
Subset Systems
Recursion Theorem
Ordered Pairs
Set Powers
Peano Axioms
Unordered Triples
Natural Induction
Generalized Set Dualities
Natural Numbers
Set Unions and Intersections
Set Dualities
Intersection of Empty Set
Successor Sets
Partitions
Set Symmetric Differences
Pair Unions
Set Intersections
Set Complements
Pair Intersections
Set Unions
Set Differences
Unordered Pairs
Empty Set
Set Specification
Set Inclusion
Standardized Accounts
Accounts
Deductions
Quantified Statements
Logical Statements
Statements
Sets
Identities
Names
Letters
Objects