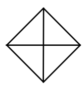


Why

We analyze the performance of inductors in a framework in which we assume the dataset comes as random samples some probability space.

Definition

Let $\mathcal{X}$ be the domain set, or set of inputs and let $\mathcal{Y}$ be the label set, or set of outs. Let $\mathcal{D} = (\mathcal{X}, \mathcal{A}, \mathbf{P})$ be a probability space. Let $f : \mathcal{X} \to \mathcal{Y}$.

Let a $(x_1, y_1), \dots, (x_n, y_n)$ of size n generated is by sampling x_i from $\mathcal{D}$ and then labeling it with $y_i = f(x_i)$.[1] For this reason we call $\mathcal{D}$ the *data-generating distribution* or *underlying distribution* and we call f the *correct labeling function.*

Measures of success

Let $A \in \mathcal{A}$, then the *error of a classifier* (or of a prediction rule) $h : \mathcal{X} \to \mathcal{Y}$ is

$$\mathbf{P}(\{x \in \mathcal{X}\} h(X) \neq f(x)).$$

In other words, the probability (w.r.t. the underling distribution) that the classifier h mislabels a point. Many authors associate an event $A \in \mathcal{A}$ with a function $\pi : \mathcal{X} \to \{0, 1\}$

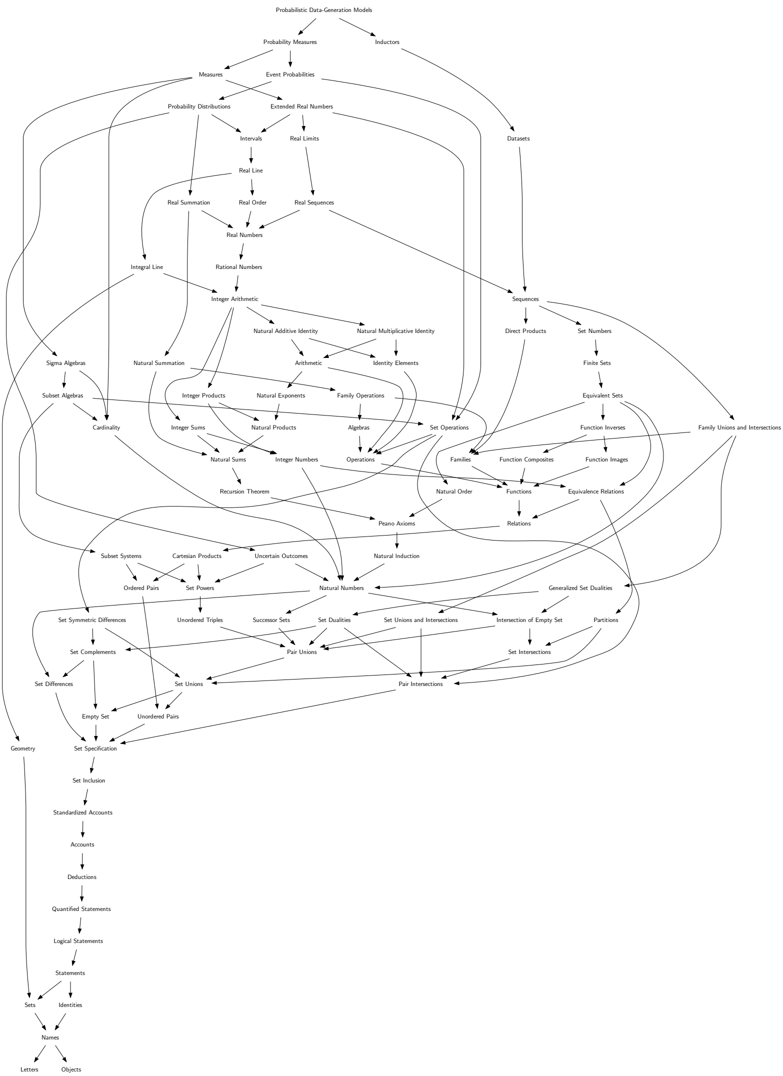
[1] Future editions will be more precise by what we mean by sampling. In other words, future editions will likely treat of these “samples” as random variables.

so that $A = \{x \in \mathcal{X} \mid \pi(x) = 1\}$ and it is common to write $\mathbf{P}[\pi(x)]$ for $\mathbf{P}(A)$.

The error is measured with respect to the distribution $\mathcal{D}$ and correct labeling function f. Other names for the error of a classifier include the *generalization error*, the *risk* or the *true error* or *loss* of h.

Statistical learning theorem

Many authors refer to a data-generating distribution along with an input set, output set, correct labeling function, and set of predictors as the *statistical learning theory framework*.


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