Why

We study inductors by defining a probabilistic model of the data generation.

Definition

Let $\mathcal{X}$ be a set of inputs and $\mathcal{Y}$ be a set of outputs. Let $f : \mathcal{X} \to \mathcal{Y}$. Let $x_i : \Omega \to \mathcal{X}$ be independent and identically distributed random variables defined on some probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and define $y_i : \Omega \to \mathcal{Y}$ by $y_i \equiv f(x_i)$ for $i = 1, \ldots, n$.

We call $((\Omega, \mathcal{A}, \mathcal{P}), \{x_i : \Omega \to \mathcal{X}\}_{i=1}^n, f : \mathcal{X} \to \mathcal{Y})$ a *probabilistic data-generation model* It is natural to call the law μ of x_i (which is the same for each i) the *data-generating distribution* or *underlying distribution* and to call f the *correct labeling function.*

Many authors refer to a probabilistic data-generation model as a *statistical learning (theory) framework.* It is also common to define $S : \Omega \to (\mathcal{X} \times \mathcal{Y})^n$ by $S = ((x_1, y_1), \ldots, (x_n, y_n))$ and refer to this as the *training dataset* of the probabilistic model. Note well that it is a random variable.

Juding predictors under the model

The upshot of this framework is that we can theoretically analyze the performance of a candidate predictor, without recourse to some test dataset.

Let $h : \mathcal{X} \to \mathcal{Y}$ be a hypothesis predictor. The *error* of this

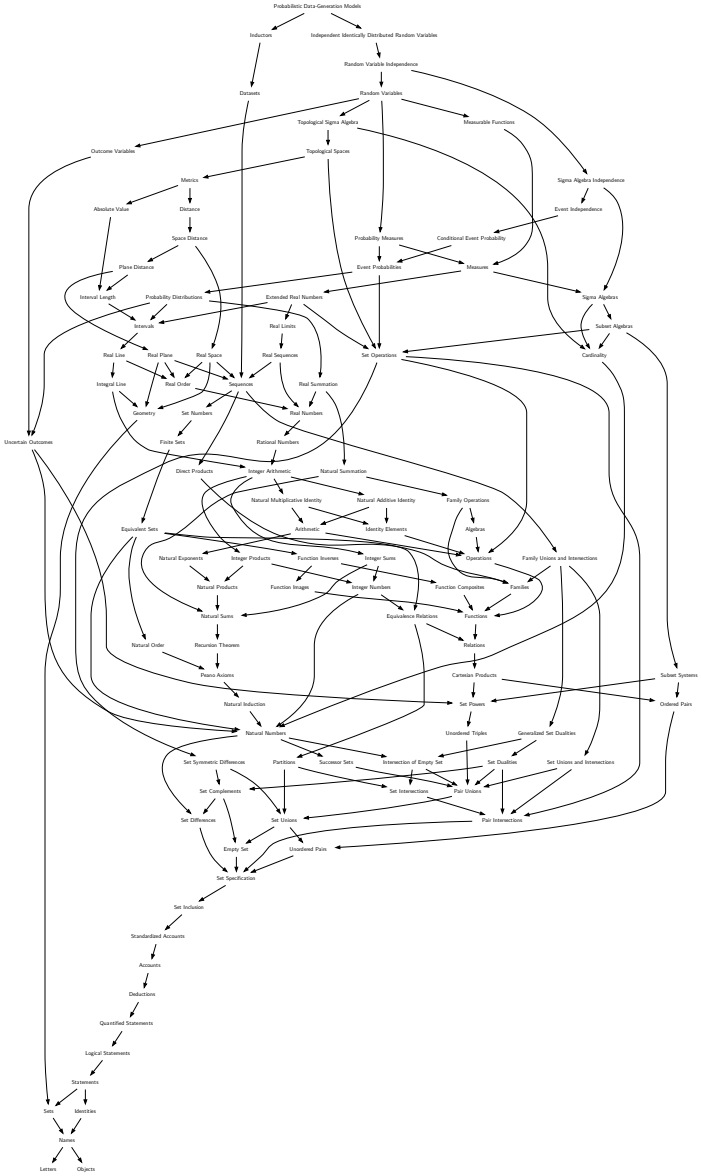
predictor under the probabilistic data-generation model is

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

In other words, the error is the probability (w.r.t. the underling distribution) that the classifier h mislabels a point.

Many authors associate an event $A \in \mathcal{A}$ with a function $\pi : \mathcal{X} \to \{0, 1\}$ so that $A = \{x \in \mathcal{X} \mid \pi(x) = 1\}$ and it is common to write $\mu[\pi(x)]$ for $\mu(A)$. In this case the error of h is $\mu[h(x) \neq f(x)]$.

The error is measured with respect to the underling distribution μ 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.


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