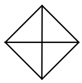


Why

A natural and simple approach is to select a predictor which performs best on the (correctly) labeled training dataset.

Definition

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

$$^1/_n \left|\{i \in \{1, 2, \dots, n\} \mid h(\xi_i) \neq \gamma_i\}\right|,$$

and so an *empirical error minimizer* for the dataset is a hypothesis whose empirical error is minimal.

Let $\mathcal{M}_{X \to Y}$ denote the set of measurable functions from X to Y. An *empirical risk minimization inductor* or *empirical risk minimization algorithm* is an inductor $A : (X \times Y)^n \to \mathcal{M}_{X \to Y}$ for which $A(D)$ is an empirical risk minimizer of D, for all datasets $D \in (X \times Y)^n$,

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

Overfitting

Although selecting a classifier to minimize the empirical risk seems natural, it can be foolish. Let $A \subset X \subset \mathbf{R}^2$, and $Y =$

$\{0, 1\}$. Suppose that the true classifier $f : \mathcal{X} \to \mathcal{Y}$ is $f(x) = 1$ if $x \in A$ and $f(x) = 0$ otherwise. Suppose that for the underlying distribution $(X, \mathcal{X}, \mu)$ we have $A \in \mathcal{X}$ and $\mu(A) = 1/2$.

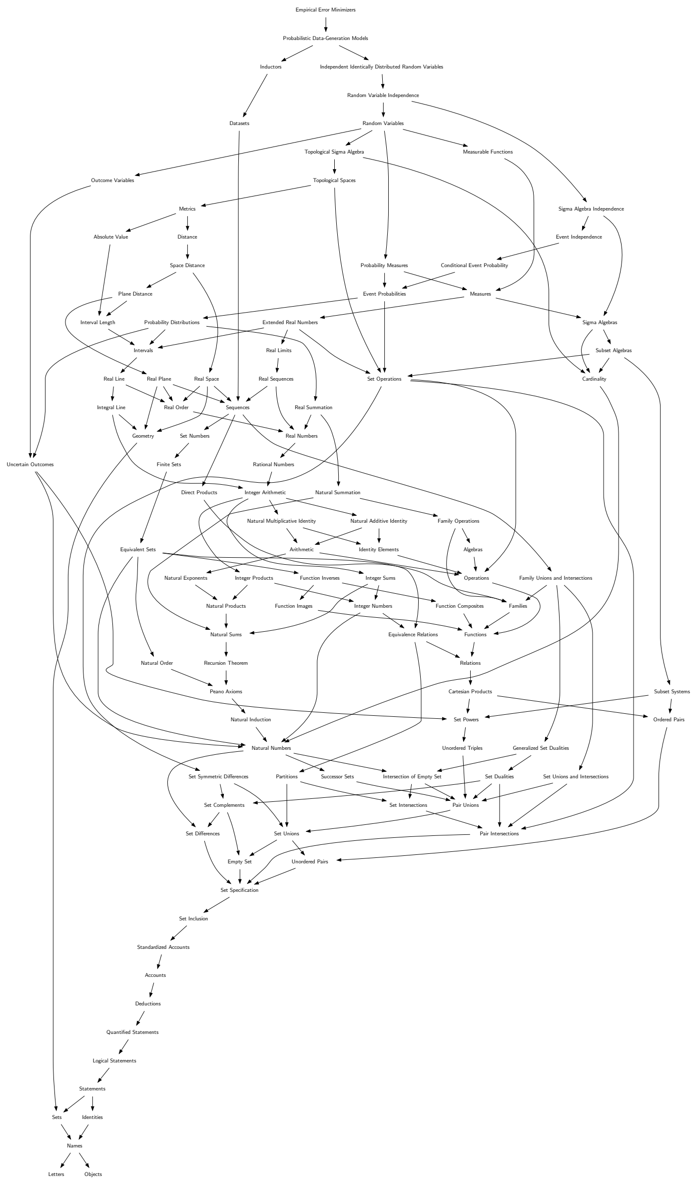
For any training set $(x_1, y_1), \dots, (x_n, y_n)$ in $X \times Y$, the hypothesis $h : X \to Y$ defined by

$$h(x) = \begin{cases} y_i & \text{if } x_i = x \\ 0 & \text{otherwise.} \end{cases}$$

achieves zero empirical risk but has error (w.r.t. μ) of $1/2$. Such a classifier is said to be *overfit* or to exhibit *overfitting*. It is said to fit the training dataset "too well."

Inductive bias

One way to mitigate overfitting for empirical error minimization is to constrain the set of predictors considered to a particular hypothesis class. As mentioned in Hypothesis Classes, we call the hypothesis class an *inductive bias*.



Empirical Error Minimizers
Probabilistic Data-Generation Models
Inductors
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 Measures
Conditional Event Probability
Plane Distance
Event Probabilities
Measures
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