Why

Are there conditions under which every empirical error minimizer does not overfit? One route is to ask whether there are situations in which good performance on empirical risk implies (or makes it highly likely to achieve) good performance on the underling distribution.

In other words, we want to connect performance on the training dataset with performance on the underlying model distribution.

Discussion

Let $((\Omega, \mathcal{A}, \mathbf{P}), \{x_i : \Omega \to \mathcal{X}\}_{i=1}^n, f : \mathcal{X} \to \mathcal{Y})$ be probabilistic data-generation model with training set $S : \Omega \to (\mathcal{X} \times \mathcal{Y})^n$.

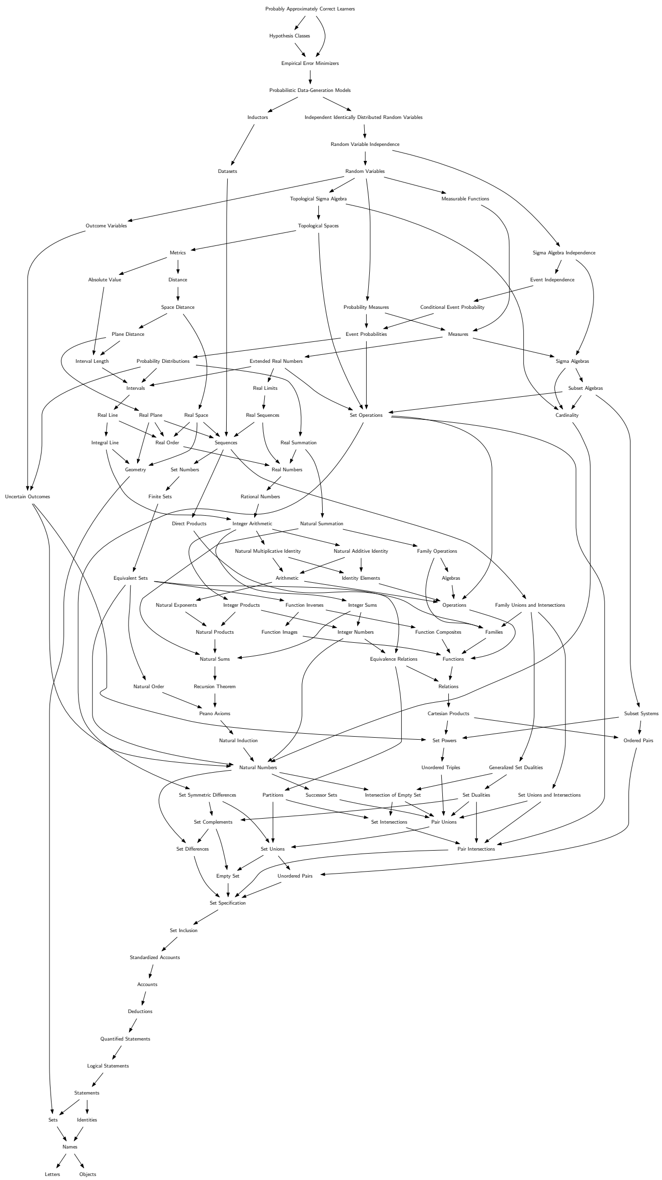
Since the empirical error minimization algorithm selects a hypothesis on the basis on the training set S, and this dataset is random, it can happen that the dataset does not "reflect well" the underlying distribution. For this reason, we will speak probabilistically about whether the error of an empirical error minimizer is large. Let $((\Omega, \mathcal{A}, \mathbf{P}), \{x_i : \Omega \to \mathcal{X}\}_{i=1}^n, f : \mathcal{X} \to \mathcal{Y})$ be probabilistic data-generation model with training set $S : \Omega \to (\mathcal{X} \times \mathcal{Y})^n$.

Moreover, there is always some nonzero probability that we will not see a particular input in $\mathcal{X}$ and so we can not hope to achieve zero generalizaion error.

Definition

For these reasons, let $\delta, \varepsilon \in (0,1)$.

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space. Let $(\mathcal{D}, f : \mathcal{X} \to \mathcal{Y})$ be a probabilistic data-generation model.


Probably Approximately Correct Learners
Hypothesis Classes
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