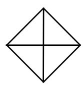


## Why

With a probabilistic data-generation model, it is natural to select a hypothesis which performs well on the training set.

## Definition

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 $S : \Omega \to (\mathcal{X} \times \mathcal{Y})^n$.

For a dataset $D = ((\xi_1, \gamma_1), \dots, (\xi_n, \gamma_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the *empirical error* of a predictor $h : \mathcal{X} \to \mathcal{Y}$ 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 $h : \mathcal{X} \to \mathcal{Y}$ whose empirical error is minimal.

An *empirical risk minimization inductor* or *empirical risk minimization algorithm* is $A : (\mathcal{X} \times \mathcal{Y})^n \to (\mathcal{X} \to \mathcal{Y})$ for which $A((\xi_1, \gamma_1), \dots, (\xi_n, \gamma_n))$ is an empirical risk minimizer.

For the random variable training set $S$, the *training error* of a classifier $h : \mathcal{X} \to \mathcal{Y}$ is a random variable $\mathrm{err}_h : \Omega \to \mathsf{R}$ defined by

$$\mathrm{err}_h(\omega) = (^1/_n) \left|\{i \in \{1, 2, \dots, m\} \mid h(x_i(\omega)) \neq y_i(\omega)\}\right|,$$

and the *empirical error minimizer set* is the random variable $\mathrm{EEM} : \Omega \to (\mathcal{X} \to \mathcal{Y})$ defined by $\mathrm{EEM}(\omega)$ is

$$\{h : \mathcal{X} \to \mathcal{Y} \mid \mathrm{err}_h(\omega) \leq \mathrm{err}_g(\omega) \text{ for all } g : \mathcal{X} \to \mathcal{Y}\}.$$

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*.

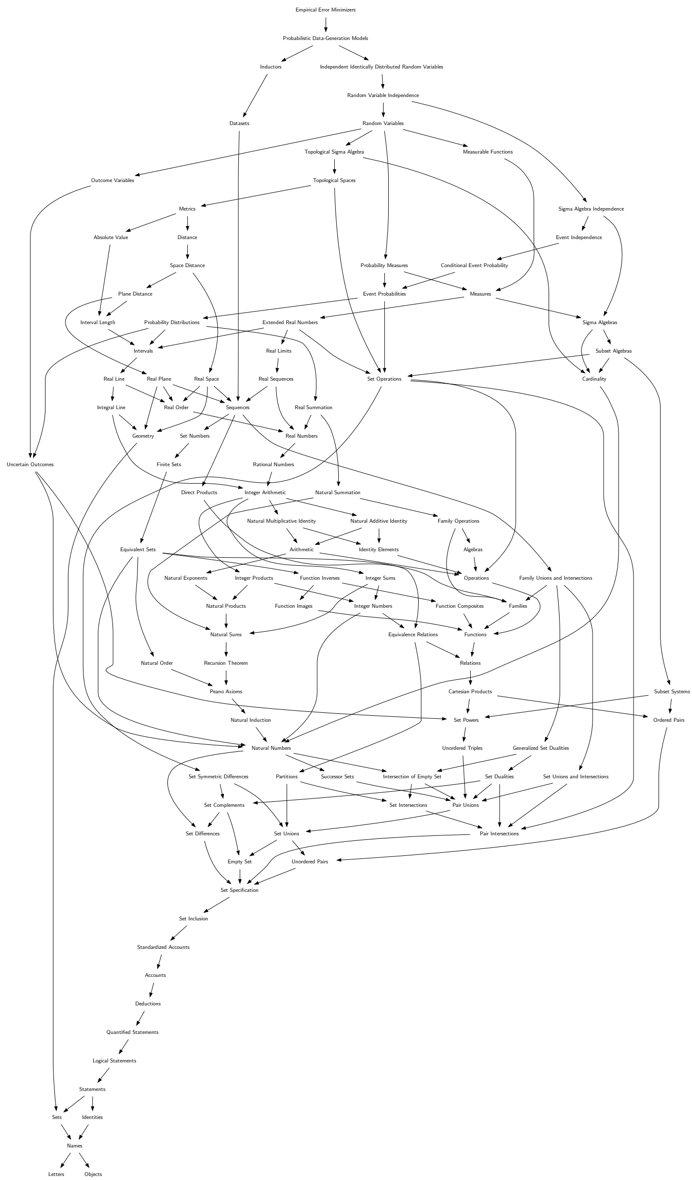
### Overfitting

Although selecting a classifier to minimize the empirical risk seems natural, it can be foolish. Let $A \subset \mathcal{X} \subset \mathsf{R}^2$ and $\mathcal{Y} = \{0, 1\}$. 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 $(\mathcal{X}, \mathcal{A}, \mathbf{P})$ we have $A \in \mathcal{A}$ and $\mathbf{P}(A) = 1/2$.

For the training set $S$ in $\mathcal{X} \times \mathcal{Y}$, the hypothesis $h : \mathcal{X} \to \mathcal{Y}$ defined by

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

achieves zero empirical risk but has error (w.r.t. $\mathbf{P}$) 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."


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