inductive bias

set of assumptions about hypotheses as they relate to the data

restriction bias

restrict the set of hypotheses considered

preference bias

prefer certain hypotheses over others

overfitting

fitting a model too closely to the training data; not generalizing

joint distribution

the distribution of probabilities over the cross product of two or more variables

estimation bias

bias resulting from testing the accuracy of a model on training examples

estimation variance

variance from true accuracy based on the makeup of test examples

perceptron

basic unit of an ANN

weak learner

a learner that does better than chance

concept

a function that maps input to output

randomized optimization

optimization methods that can be used on functions that are not differentiable

version space

the set of hypotheses that is consistent with the training data

consistent learner

an algorithm that outputs a hypothesis in the version space

ensemble learning

combines simple rules to form a complex rule

early stopping

a method used by ANNs to avoid overfitting by halting training when the test error increases

pre-pruning

stops tree growth before overfitting occurs

post-pruning

allows overfitting to occur before pruning the tree for generalization

sigmoid

a perceptron-like unit that is based on a differentiable threshold function

Occam's razor

prefer the simplest hypothesis that fits the data

Bayes optimal classifier

finds the most probable classification of an instance by combining the weighted predictions of all hypotheses

shatter

for every dichotomy of S, there is some hypothesis in H consistent with this dichotomy

VC dimension

the size of the largest finite subset of instances shattered by the hypothesis space

Bayesian network

represents the joint probability distribution for a set of variables

epsilon-exhaustion

all hypotheses in the version space have error less than epsilon

PAC-learnable

a concept that can be learned by L with probability 1-delta and error less than epsilon in time that is polynomial in 1/epsilon, 1/delta, n, and size(c)

mutation

small, random changes to a bit string, applied after crossover

crossover

creates two new offspring from two parent string by copying selected bits from each parent

single-point crossover

crossover mask begins with n contiguous 1s, followed by 0s

two-point crossover

crossover mask begins with 0s, then 1s, then 0s

uniform crossover

crossover mask bits sampled uniformly from both parents

crowding

fit hypotheses dominate and reproduce, taking over the population

computational complexity

the computation effort required to find a successful hypothesis

sample complexity

the number of training examples required to find a successful hypothesis

mistake bound

the number of misclassifications that will occur before finding a successful hypotheses

instance

the input vector of attributes

target concept

the answer to the problem

conditional independence

the probability distribution governing X is independent of the value of Y given a value for Z

unbiased learner

a learner that makes no a priori assumptions about the identity of the target concept

information gain

the expected reduction in entropy caused by partitioning the examples according to an attribute

entropy

characterizes the impurity of a collection of examples

no free lunch theorem

a generalized, universal-purpose optimization solution is impossible

expected value

mean

sample error

the fraction of a set of examples that is misclassified

true error

the probability that a hypothesis will misclassify an instance randomly drawn from D

binomial distribution

gives the probability of observing r heads in a sample of n independent coin tosses, when the probability of heads on a single toss is p

confidence interval

an interval that is expected to contain some parameter with probability N%

minimum description length

choose the hypothesis that minimizes the sum of the description length of the hypothesis and the description length of the data given the hypothesis

central limit theorem

given a large number of independent variables with a finite variance, the distribution will be normal

mutual information

a measure of dependence of the relationship between two variables

curse of dimensionality

as the number of dimensions grow, the amount of data we need to generalize grows exponentially