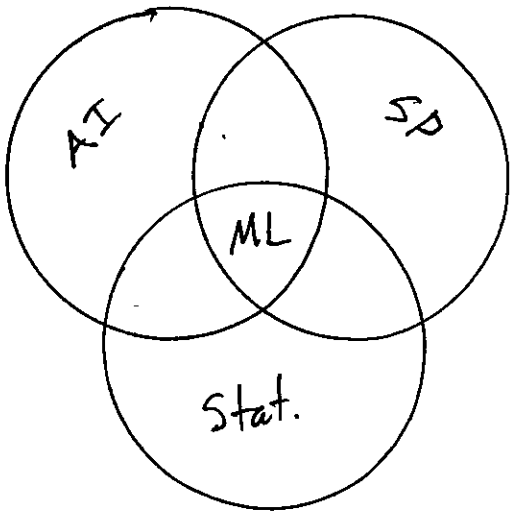
STATISTICAL MACHINE LEARNING

Machine Learning

Machine learning is a multi-disciplinary field concerned with the study of algorithms that learn from examples.

ML theory and methodology emerged historically out of three areas: artificial intelligence, signal processing, and statistics.

By now, the best practices of these areas have spread to the others, and ML is its own well-defined field.



Learning from examples $\Leftrightarrow$ learning from <u>data</u>

Notation for data:

X random variable

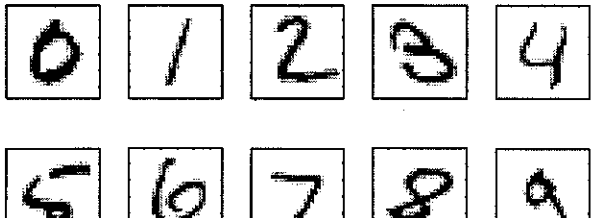
x realization of X

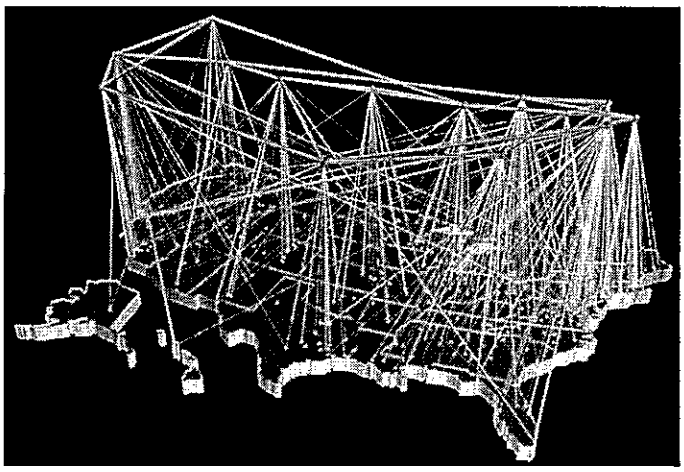
Typically, $X \in \mathbb{R}^d$.

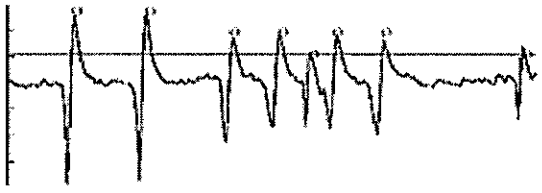
X represents a measurement or observation of some natural or man-made phenomenon, and may be called a _____. The coordinates of X are called _____.

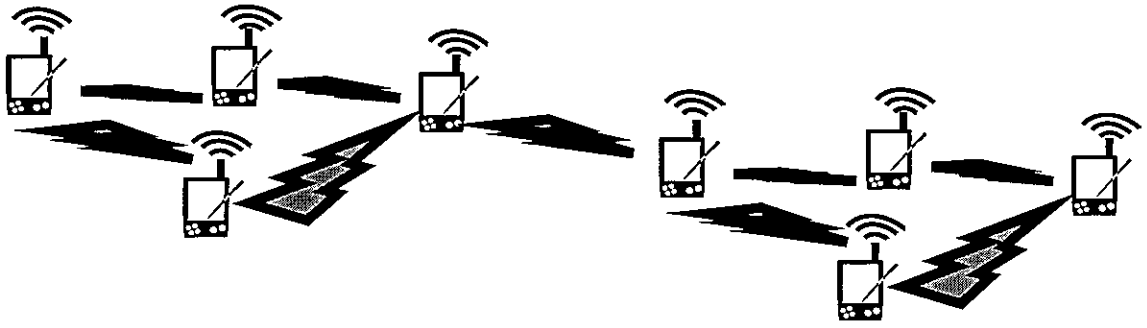
- pattern
- signal
- feature vector
- input
- instance
- independent variable

- features
- attributes
- predictors
- covariates









Statistical Machine Learning

In most applications, there is some uncertainty or randomness inherent in the data.

In statistical machine learning, we will

- view a pattern X as a random variable
- use the tools of probability and statistics to provide a mathematical framework for
  - → posing machine learning problems
  - → formulating solutions to those problems.
  - → evaluation

The following terms are often used interchangeably:

- Machine learning
- Statistical machine learning
- Statistical learning
- Pattern recognition
- Multivariate data analysis

Types of Learning Problems

We will consider various paradigms for statistical learning.

Supervised Learning

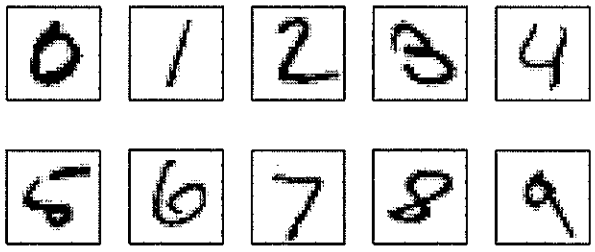
In addition to patterns

$$X_1, \ldots, X_n$$

we also have access to variables

$$Y_1, \ldots, Y_n.$$

Examples of patterns



Training data (suppose correct labels are provided)

Goal: predict label of a future pattern

We may think of the pair (X, Y) as obeying a (possibly noisy) input-output relationship.

The goal of supervised learning is usually to <u>generalize</u> the input-output relationship, facilitating the <u>prediction</u> of the output associated with previously unseen inputs X.

The data

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

are called <u>training data</u>.

The Y variable may be called a

- response
- output
- label
- dependent variable

The primary supervised learning problems are

- classification : $Y \in \{1, \ldots, M\}$
- regression : $Y \in \mathbb{R}$

Unsupervised Learning

The patterns

$$X_1, \ldots, X_n$$

are not accompanied by output variables.

The goal of unsupervised learning is typically not related to future observations. Instead, one seeks to understand structure in the data sample itself, or to infer some characteristic of the underlying probability distribution.

The primary unsupervised learning problems are

- clustering
- density estimation
- dimensionality reduction
  ↑ can also be supervised

Reinforcement Learning

The patterns $X_1, X_2, \ldots$ are observed sequentially. After each X_i is observed, the learner must take an action. After each action, the learner receives a reward from the environment. The goal of the learner is to determine a policy (for selecting actions based on observations) to maximize long-term reward.

RL is important in robotics (navigation, path planning), economics, and other areas.

Other Learning Problems

- Semi-supervised learning
- Active learning
- Online learning
- Novelty detection
- Ranking
- Transfer learning
- Multi-task learning
-
-

Types of Learning Methods

Distributional assumptions

- Generative: full probabilistic model
- Discriminative: partial or no probabilistic model; typically models only the desired function or set (e.g. the decision boundary in classification)

Computational Form

- Linear
- Nonlinear
  - polynomial
  - partition-based
  - kernel based
  - ⋮

Complexity

- parametric: # of model parameters is independent of sample size
- nonparametric: # of model parameters grows with sample size

Prerequisites

Probability

- joint distributions, multivariate densities and mass functions, expectation, independence, conditional distributions, Bayes rule, the multivariate normal (Gaussian) distribution

Ex] In classification, we view (X, Y) as a jointly distributed pair. We will show that the optimal classifier, in terms of probability of error, is

$$x \longmapsto \arg\max_y P(Y=y \mid X=x)$$

Linear Algebra

Linear methods are widely used in all learning problems. They are typically simpler to understand, and also form the basis of more sophisticated nonlinear methods

- rank, nullspace, linear independence, span, dimension, inner product, orthogonality, positive (semi-) definite matrices, eigenvalue decomposition, projection

Parameter Estimation

Suppose ($\overset{iid}{\sim}$: independent and identically distributed)

$$X_1, \ldots, X_n \overset{iid}{\sim} f(x; \theta)$$

where θ is an unknown parameter to be estimated.

The maximum likelihood estimate (MLE) of θ is

$$\hat{\theta} = \arg\max_{\theta} f(X_1, \ldots, X_n ; \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} f(X_i ; \theta)$$

If θ is viewed as a random variable, we write $X_i \sim f(x \mid \theta)$, and base our estimates on the posterior distribution $f(\theta \mid X_1, \ldots, X_n)$

$$\propto f(X_1, \ldots, X_n \mid \theta) \cdot f(\theta)$$

($\propto$: Bayes' rule; $f(\theta)$: prior)

Ex | MAP estimate: $\hat{\theta} = \arg\max_{\theta} f(\theta \mid X_1, \ldots, X_n)$