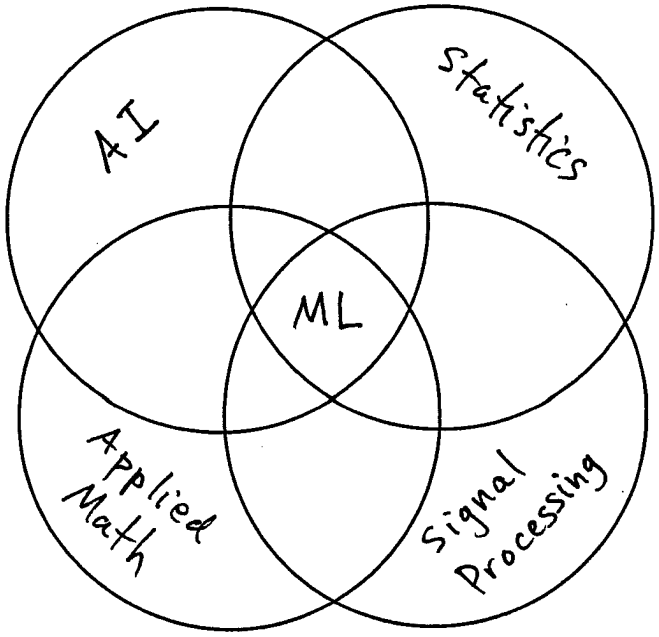
STATISTICAL MACHINE LEARNING

Machine Learning

Machine learning is a multi-disciplinary field of study concerned with the design of algorithms that allow computers to "learn."

The term "machine learning" comes from the artificial intelligence community, but is now a focus area in many branches of applied math and computer science.



In this class, "learning" refers to learning from examples, or learning from <u>data</u>

Notation for data:

X random variable

x realization of X

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

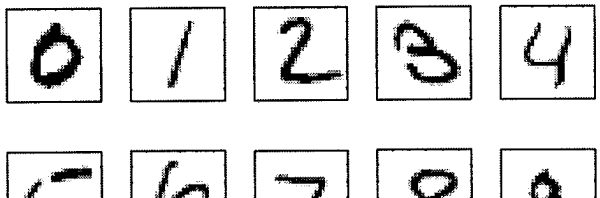
X represents a measurement or observation of some natural or man-made phenomenon, and may be called a _____. The coordinates of X are called _____.

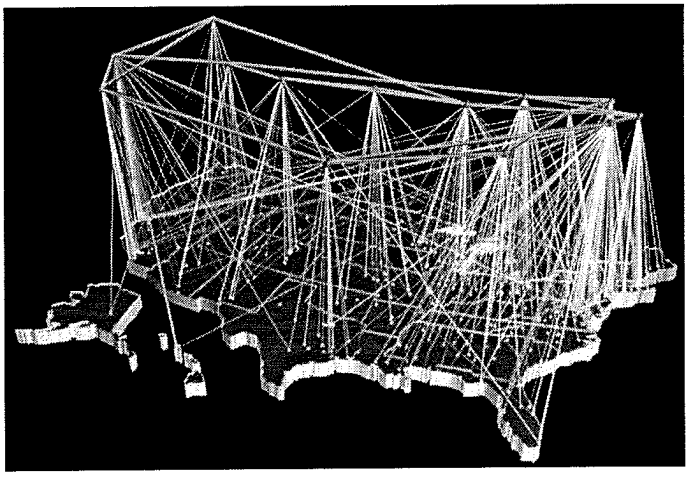
- pattern ↑
- signal
- feature vector
- input
- instance
- independent variable

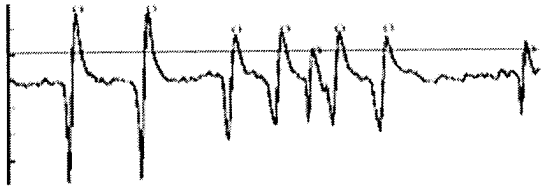
- attributes ↑
- features
- predictors
- variates

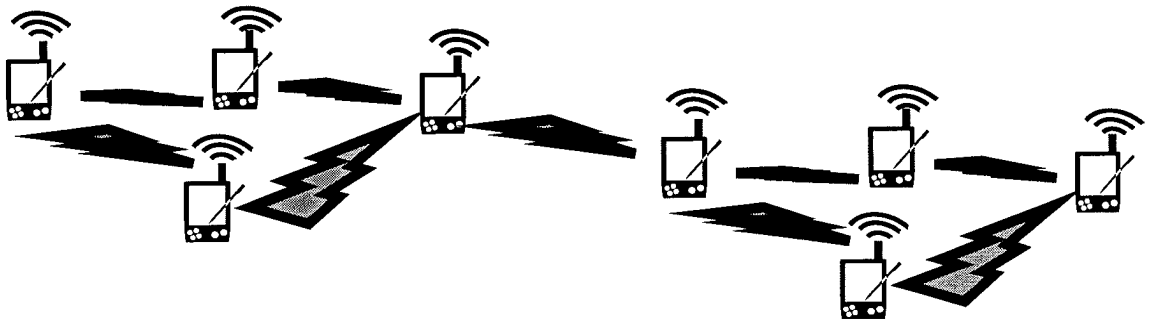
<u>Training data</u>: We usually observe several patterns, $X_1, \ldots, X_n$, and need to

- predict some property of a future pattern X_n
- infer some underlying parameter or function characterizing the data
- extract low dimensional structure
- select an action to maximize a payoff
- etc.









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.

The following terms are often used interchangeably:

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

Types of Learning Problems

We will consider three major paradigms for statistical learning.

Supervised Learning

In addition to patterns

$$X_1, \ldots, X_n$$

we also have access to variables

$$Y_1, \ldots, Y_n.$$

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

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

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

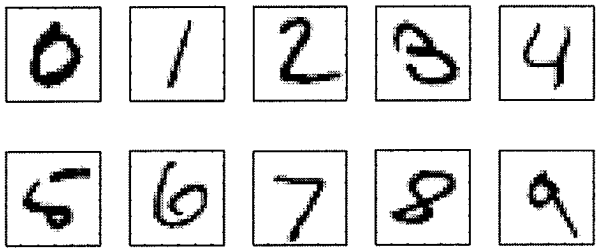
The Y variable may be called a

- response
- output
- label
- dependent variable

The primary supervised learning problems are

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

Examples of patterns



Training data (suppose correct labels are provided)

Goal: predict label of a future pattern

Unsupervised Learning

The patterns

$$X_1, \ldots, X_n$$

are <u>not</u> accompanied by output variables.

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

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.

Types of Learning Methods

Distributional assumptions

- Generative : full probability model
- Discriminative : models only the desired function or set (e.g., the decision boundary in classification)

Computational Form

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

Complexity

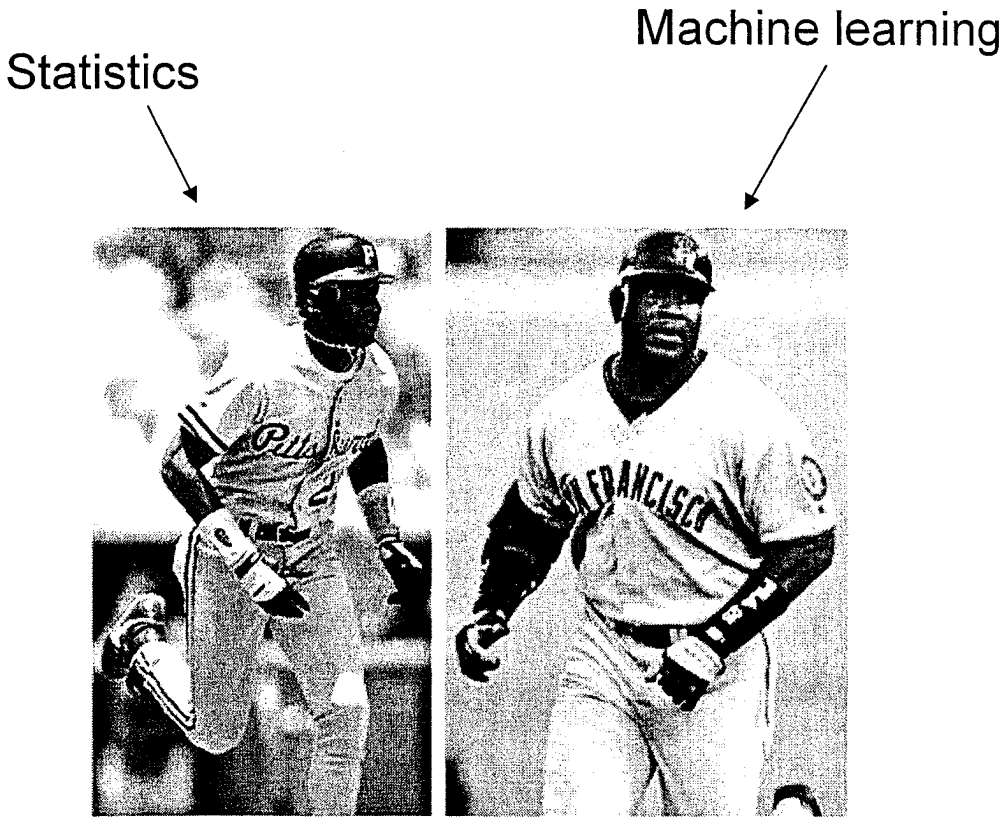
- parametric : # of model parameters independent of sample size
- nonparametric : # of model parameters grows with sample size

"Statistical Machine Learning" vs. "Statistics"

So what's the difference between SML and good ol' multivariate statistical data analysis?

SML tends to emphasize large scale data:
high dimensionality (d) and/or large sample size (n)

> machine learning = statistics on steroids



Other Learning Problems

- Semi-supervised learning / Transductive learning
- Active learning / Query learning
- Online learning
- Learning with structured outputs
- Anomaly prediction / one-class classification
- Co-clustering
- Distributed learning
- Co-training
- Ranking
-
-