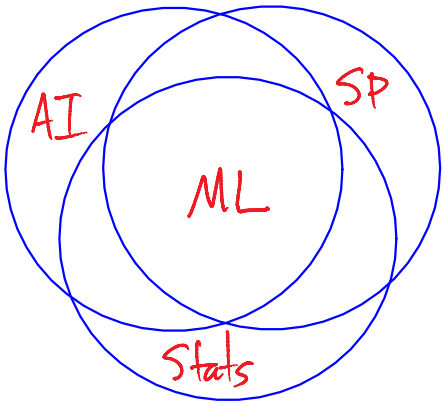
STATISTICAL MACHINE LEARNING

Machine learning is a field of study concerned with making quantitative inferences and predictions based on data.

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 has an independent identity.



The term "statistical machine learning" recognizes

the fact that most of modern machine learning has its foundation in probability and statistics as a framework for handling the uncertainty or randomness inherent in data.

In this course, we will use the framework of probability and statistics to

- ▷ pose machine learning problems
- ▷ formulate solutions to those problems
- ▷ evaluate performance

Data: Notation and Terminology

We will typically denote a measurement by x.

In this course we will assume $x \in \mathbb{R}^d$, and write

$$x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} \quad \text{or} \quad x = \begin{bmatrix} x_1 \\ \vdots \end{bmatrix}$$

$\begin{bmatrix} x^{(d)} \end{bmatrix}$ $\begin{bmatrix} x_d \end{bmatrix}$

depending on which notation is best suited for a given situation. x is called a

pattern, signal, input, instance, feature vector

and the coordinates of x are called

features, attributes, predictors, covariates.

In the statistical setting, we view x as a realization of a random variable X.

Supervised Learning

In supervised learning, we observe training data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

where $x_1, \ldots, x_n$ are patterns and $y_1, \ldots, y_n$ are associated outputs. The goal in supervised learning is to predict the output

y associated to a new test pattern x.

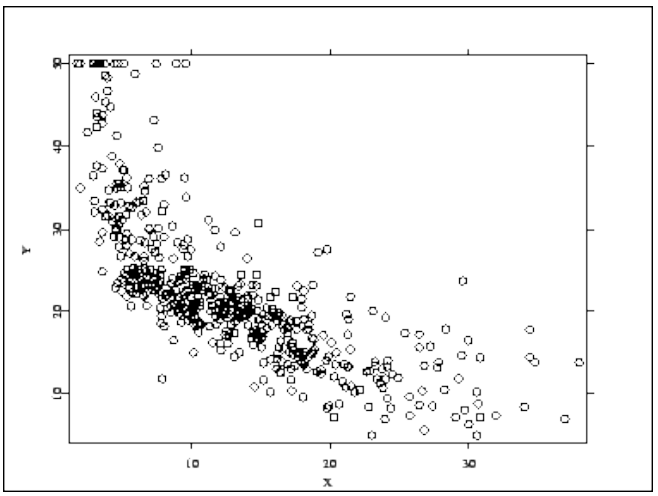
There are two basic kinds of supervised learning problems:

- classification: $y \in \{1, \ldots, C\}$

  (y: label)

- regression: $y \in \mathbb{R}$

  (y: response variable)



In supervised learning, the central task is generalization: learning a general input-output relationship (i.e., a function) from a finite number of measurements.

Unsupervised Learning

In unsupervised learning, the patterns $x_1, \ldots, x_n$ are not accompanied by output variables.

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

underlying probability distribution.

The primary unsupervised learning problems are

- clustering
- density estimation
- dimensionality reduction

Other Machine Learning Problems

- reinforcement learning
- semi-supervised learning
- active learning
- online learning
- novelty detection
- ranking
- transfer learning
- :

Types of Learning Methods

Machine learning methods are often categorized based on various factors. The terms below will make more sense after we have covered some methods in detail.

Distributional assumptions

- generative: full probabilistic model
- discriminative: partial or no probabilistic model

Computational form

- linear: the output is a linear/affine function of the input
- nonlinear

Complexity

Complexity

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