# ML and Optimization Lecture 2 9/1/2023

- What is ML?
- Example of ML: support vector machines
- Probability Theory*

* next time

## what is ML?

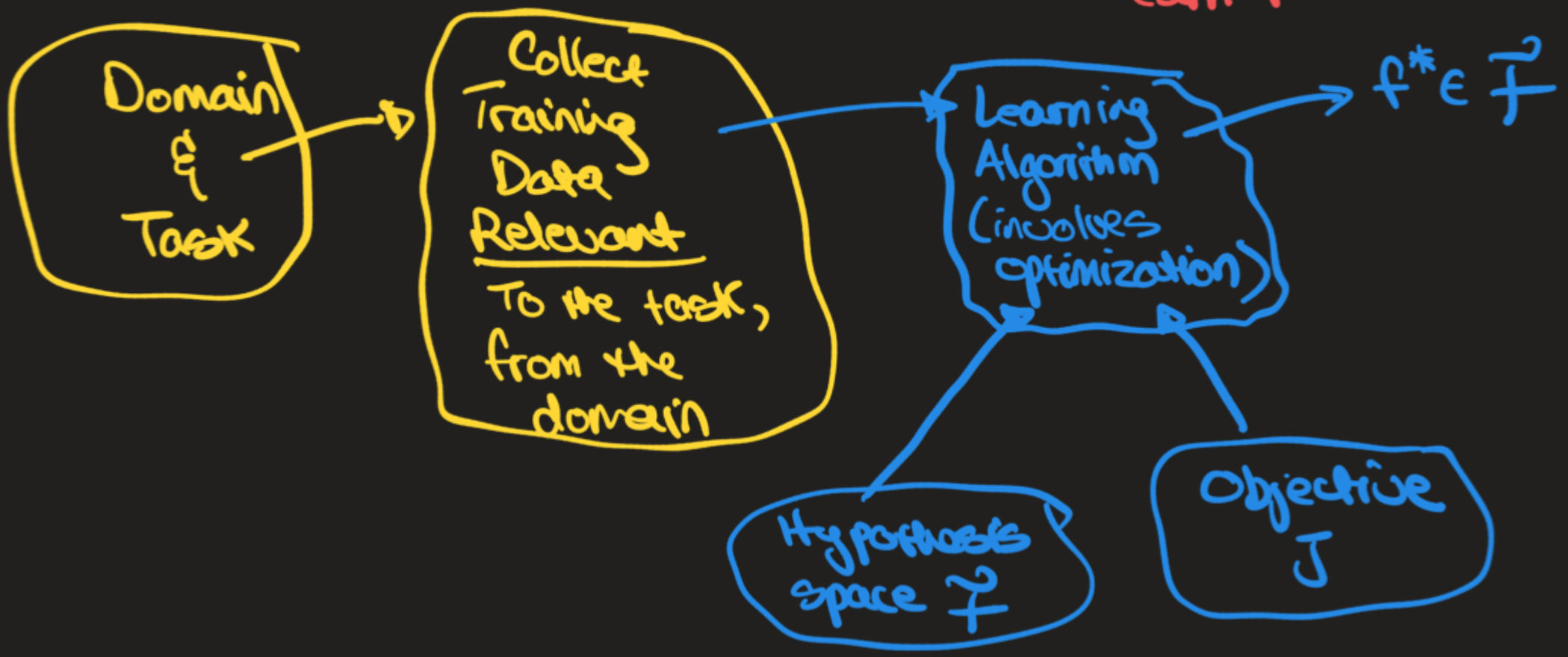
(supervised learning)

"finding a predictable pattern from a set of data"

predictable → accurate

pattern → $f: x \mapsto y$

set of data → observations of $(x, y)$ used to learn $f$

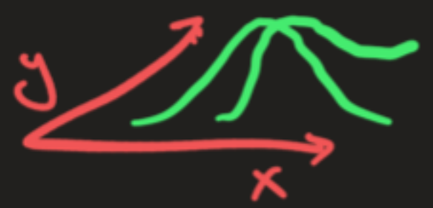


## What does it mean for $f$ to be optimal?

First: $f$ needs to minimize some error over the domain.

We describe domains as sets of $(x, y)$ with a probability measure, ie. a function that maps subsets of my domain to numbers in $[0, 1]$

write $(x, y) \sim \mathbb{P}$




Now define/choose

$\ell$ – loss function $\ell(\cdot, \cdot)$ measures discrepancy between prediction and true value
this depends on the task

$f$ – function in our hypothesis space $\mathcal{F}$
(e.g. the linear functions, quadratic functions, NNs)
this depends on the task & amount of training data

# Formalize ML as Empirical Risk Minimization

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathbb{E}_{(x,y)\sim \mathbb{P}} \, \ell(f(x), y)$$

Population Risk

$$= \int \ell(f(x), y) \, p(x,y) \, dx\,dy$$

by the law of large numbers, as $n \to \infty$

$$\approx \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Empirical Risk

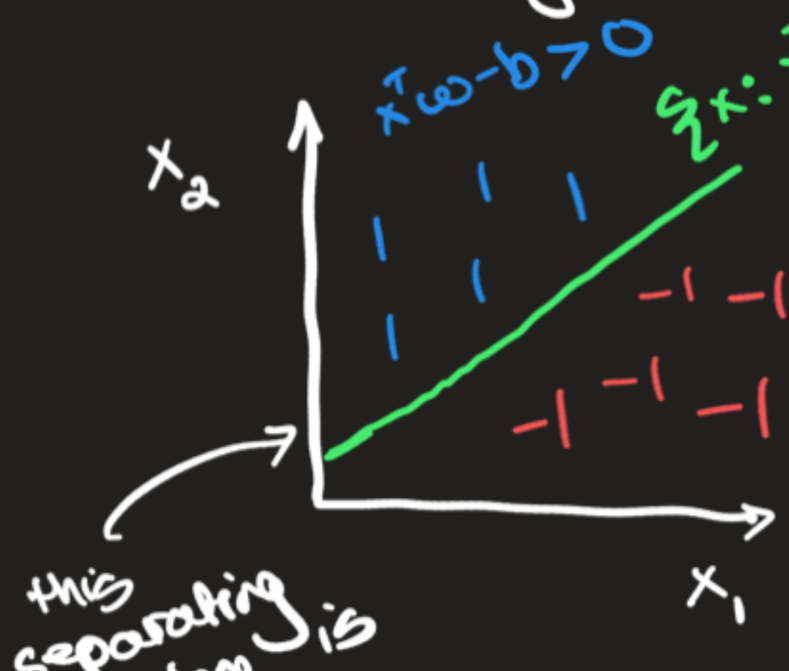
where $(x_i, y_i) \sim \mathbb{P}$ are independent, identically distributed samples from our domain

**Empirical Risk Minimization (ERM)** consists of choosing $f$ by minimizing the empirical risk

## High-level workflow for ML (supervised ML)

- pick a domain & formulate a task (choose an appropriate loss function)
- collect i.i.d. samples from our domain
- choose an appropriate hypothesis class $\mathcal{H}$
- solve the ERM
- iterate

**Ex:** Binary classification using SVMs (support vector machines)




Here $y \in \{\pm 1\}$

Assumption: the two classes are linearly separable

The corresponding hypothesis space is

$$\mathcal{F} = \{f : f(x) = \text{sign}(x^T\omega - b)\}$$

Our model is confident/accurate when

$$y(x^T\omega - b) \gg 1$$

- the prediction $f(x) = \text{sign}(x^T\omega - b)$ must equal $y$
- the model must be confident: $x$ is far from the separating hyperplane

Models with low confidence can be made arbitrarily highly confident: if

$$\min_i \; y_i (x_i^T \omega - b) = d > 0$$

then

$$\min_i \; y_i (x_i^T (\alpha\omega) - \alpha b) = \alpha d > 0$$

for any $\alpha > 0$

To avoid this, we restrict the norm of $(\omega, b)$

We choose our ERM to satisfy these desiderata by:

1) choose a loss function to ensure confidence

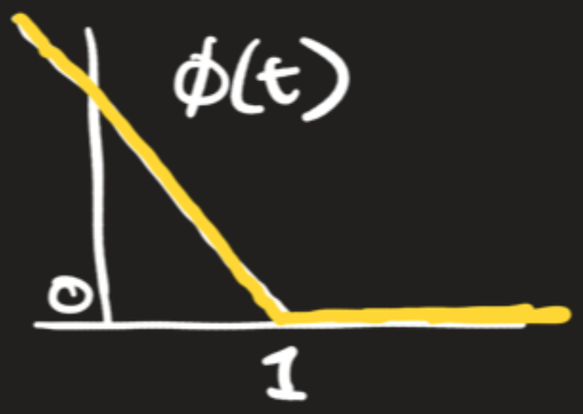
2) restrict our hypothesis space to ensure $(\omega, b)$ is bounded

Our hypothesis space becomes

$$\mathcal{F} = \{ f : f(x) = \text{sign}(w^T x + b), \text{ where } \|w\|_2 \leq C \}$$

For our loss function, we use the hinge loss:

$$\phi(t) = (1-t)_+ = \begin{cases} 1-t & t<1 \\ 0 & \text{otherwise} \end{cases}$$




$$l(y, f(x)) = \phi(y(w^T x - b))$$

The final ERM for the soft-margin SVM is

$$(C) \quad \min_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R} \\ \|w\|_2 \le C}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i(x_i^T w - b))_+$$

constrained ERM

or

$$(L) \quad \min_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i(x_i^T w - b))_+ + \lambda \|w\|_2^2$$

regularization coefficient

Lagrangian multiplier or regularized version

These are equivalent: for any $C$ there is a corresponding $\lambda$ so the solutions to (C) and (L) are the same.