# STATISTICAL LEARNING THEORY

Summer Term 2025

**Book:** Shai Shalev-Schwartz
Understanding Machine Learning

**Online Notes:** http://rkwitt.org → Teaching
(All my handwritten notes will be available there!)

## OTHER COURSE NAMES:

1) Advanced machine learning
2) Machine learning

# MOTIVATION

Example:

| | $\in \mathbb{R}$<br>Weight (in g) | $\in \mathbb{R}$<br>Color ($\in [0,1]$) | Tasty? |
|---|---|---|---|
| Papaya 1 | 800 | 0.1 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| Papaya N | 1200 | 0.7 | 1 |

0 … non-tasty
1 … tasty

Lets call the information in this table our training data.
Based on that, we try to find

$$\boxed{h : \mathbb{R} \times \mathbb{R} \to \{0,1\}}$$

i.e., a function that will take a two-dimensional vector as input (weight & color) and output a prediction of whether a papaya is tasty (1) or not (0).

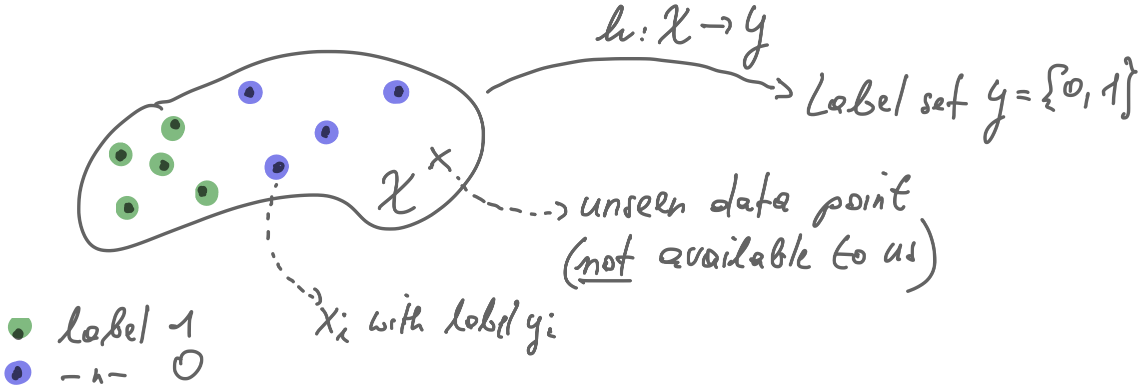
We call such a function a hypothesis.

More formally, the data that is available to us comes in the form

$$\big((x_1, y_1), \ldots\ldots, (x_N, y_N)\big) = S$$

with $x_i \in \underbrace{\mathbb{R}^2}_{\text{Domain } \mathcal{X} = \mathbb{R}^2}$ and $y_i \in \underbrace{\{0, 1\}}_{\text{Label set } \mathcal{Y} = \{0,1\}}$.

## What do we mean by "learning"?




We say a learner receives $S$ and outputs $h$!
(some algorithm)

Two assumptions we will make initially:

(1) All the $x_i$'s are drawn independently and identically (iid) from some (unknown) distribution $D$ over the domain $\mathcal{X}$.

(2) The $x_i$'s are labeled by some (unknown) function

$$f : \mathcal{X} \to \mathcal{Y},$$

called the true labeling function. This means

$$S = \Big( (x_1, \underbrace{f(x_1)}_{y_1}), \ldots\ldots, (x_N, \underbrace{f(x_N)}_{y_N}) \Big)$$

What do we care about?

We care about

$$\{ x \in \mathcal{X} : h(x) \neq f(x) \} = A$$

That is, all the points $x$ in our domain $\mathcal{X}$, where the hypothesis $h$ differs from the true labeling function $f$.

1. Domain set $\mathcal{X}$; we call $x \in \mathcal{X}$ an instance
2. Label set $\mathcal{Y}$, e.g., $\mathcal{Y} = \{0,1\}$
3. Training set $S = ((x_1,y_1), \ldots, (x_m, y_m))$ with
$$(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} =: Z$$
4. A learner receives $S$ and outputs $h: \mathcal{X} \to \mathcal{Y}$ (i.e., a hypothesis)

Assumption: For now, we assume that the $x_i$'s are drawn iid from some probability measure $\mathcal{D}$ over the domain $\mathcal{X}$ and labeled by some function $f: \mathcal{X} \to \mathcal{Y}$, so $y_i = f(x_i)$.

We are interested in:

$$\mathcal{D}\left(\left\{x \in \mathcal{X} : h(x) \neq f(x)\right\}\right) = \mathbb{P}_{x \sim \mathcal{D}}\left[h(x) \neq f(x)\right] = L_{\mathcal{D},f}(h)$$

"Generalization error"

The empirical version of that is

$$\frac{1}{m} \cdot \left|\left\{ i \in \{1, \ldots, m\} : h(x_i) \neq f(x_i)\right\}\right| = L_S(h)$$

"Empirical error"
(Empirical risk)

Notation:

$[m] = \{1, \ldots, m\}$, $S|_x = (x_1, \ldots, x_m)$

**Claim:** $\mathbb{E}_{\underbrace{S|_x}_{(x_1,\dots,x_m)} \sim D^m}\left[L_S(h)\right] = L_{D,f}(h)$

$$\mathbb{E}_{S|_x \sim D^m}\left[L_S(h)\right] = \mathbb{E}_{S|_x\sim D^m}\Big[\overbrace{\frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq f(x_i)}}^{\frac{1}{m}\cdot\left|\{i\in[m]:\ h(x_i)\neq f(x_i)\}\right|}\Big] \qquad \text{// by def.}$$

$$= \frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{E}_{x_i\sim D}\left[\mathbb{1}_{h(x_i)\neq f(x_i)}\right] \qquad \text{// by linearity of } \mathbb{E}[\cdot]$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x\sim D}\left[\mathbb{1}_{h(x)\neq f(x)}\right] \qquad \text{// as all } x_i\text{'s are drawn iid from } D$$

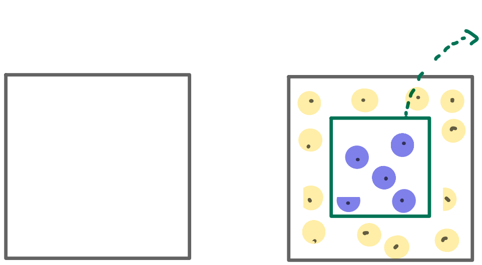
$$= \frac{1}{m}\cdot\sum_{i=1}^{m} \mathbb{P}\left[h(x)\neq f(x)\right]$$

$$= \frac{1}{\cancel{m}}\cdot\cancel{m}\cdot\mathbb{P}\left[h(x)\neq f(x)\right] = \underbrace{L_{D,f}(h)}_{\text{generalization error}}$$

## Our first learning paradigm

**Empirical Risk Minimization (ERM).** As a learner only has access to the training data ($S$), it is "natural" to try to select $h$ (our hypothesis) such that the empirical risk (emp. error) is minimized. We call such a $h$ an **empirical risk minimizer** (an ERM hypothesis).

## Example (of a problematic case):




Domain $\mathcal{X}$

Say the distribution on $\mathcal{X}$ is uniform.

label 1
label 0

Also, assume that the area of the domain (square) is 2 and the area of $\square$ is 1.

Now, say we have an ERM algorithm that returns $h_S$ such that

$$h_S(x) = \begin{cases} y_i, & \text{if } \exists i \in \{1, \dots, m\} : x_i = x \\ 0, & \text{else} \end{cases}$$

(sort of a lookup table)

Obviously, $h_S$ is correct on all instances in $S \to h_S$ is an emp. risk minimizer, meaning

$$L_S(h_S) = 0!$$

But, on unseen instances from $D$ (which is uniform on $\mathcal{X}$), $h_S$ is only correct 50% of the time $\to$

$$L_{D,f}(h_S) = \frac{1}{2}!$$

That's what we call overfitting!

**Hypothesis class ($H$)**: We restrict searching for $h$ to $H$, i.e., a class of functions from $\mathcal{X} \to \mathcal{Y}$ and we write

$$\boxed{ERM_H(S) \in \underset{h \in H}{\arg\min}\ L_S(h)}$$

## ERM over finite hypothesis classes ($|H| < \infty$)

**Assumption (realizability)**: $\exists\, h^* \in H$ with $L_{D,f}(h^*) = 0$.

Now, any ERM hypothesis $h_S$ will attain $0$ empirical error ($L_S(h_S) = 0$) as $h_S$ competes against $h^*$ (which has $L_{D,f}(h^*) = 0$ and, obviously, $L_S(h^*) = 0$).

We know that $\underline{L_{D,f}(h_S) > \varepsilon}$, $\varepsilon \in (0,1)$, can only happen if our learner selects a hypothesis $h_S$ with $L_S(h_S) = 0$, BUT $L_{D,f}(h_S) > \varepsilon$.

We define

$$H_{BAD} = \{ h \in H : L_{D,f}(h) \geq \varepsilon \}$$

set of bad hypothesis!

Also, we define

$$M = \{ S|_x : \exists\, h \in H_{BAD},\ L_S(h) = 0 \}$$

## Observation:

$$\{S|_x : L_{\mathcal{D},f}(h_S) \geq \varepsilon\} \subseteq \{S|_x : \exists h \in H_{BAD},\ L_S(h) = 0\} = M$$

$\uparrow$ ERM hypothesis (Empirical risk minimizer)

Since $\{S|_x : \exists h \in H_{BAD},\ L_S(h) = 0\} = \bigcup_{h \in H_{BAD}} \{S|_x : L_S(h) = 0\}$

we have

$$\mathcal{D}^m(\{S|_x : \exists h \in H_{BAD},\ L_S(h) = 0\}) = \mathcal{D}^m\left(\bigcup_{h \in H_{BAD}} \{S|_x : L_S(h) = 0\}\right)$$

(by σ-sub-additivity 'union' bound) $\leq \sum_{h \in H_{BAD}} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$

## Lets fix some $h \in H_{BAD}$:

$$\mathcal{D}^m(\{S|_x : L_S(h) = 0\}) = \mathcal{D}^m(\{S|_x : \forall i \in \{1, \ldots, m\} : h(x_i) = \underset{\text{True labeling function}}{f}(x_i)\})$$

by iid assumption

$$= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\})$$

$$= \prod_{i=1}^{m} \mathcal{D}(\{x : h(x) = f(x)\})$$

(by definition of $L_{\mathcal{D},f}$) $= \prod_{i=1}^{m} (1 - \underbrace{L_{\mathcal{D},f}(h)}_{> \varepsilon})$ (remember that $h \in H_{BAD}$)

$\hookrightarrow$

$$\le \prod_{i=1}^{m} (1-\varepsilon)$$

$$= (1-\varepsilon)^m$$

$$\le e^{-\varepsilon m} \qquad \text{(without proof)}$$

Overall, we have

$$D^m\left(\{S|_x : L_{\mathcal{D},f}(h_S) \ge \varepsilon\}\right) \le \sum_{h \in H_{BAD}} D^m\left(\{S|_x : L_S(h) = 0\}\right)$$

$$\le \sum_{h \in H_{BAD}} e^{-\varepsilon m}$$

$$= |H_{BAD}| \cdot e^{-\varepsilon m}$$

$$\le |H| \cdot e^{-\varepsilon m} \qquad \left(\text{because } H_{BAD} \subseteq H\right)$$

if we want $|H| \cdot e^{-\varepsilon m}$ to be less than some $\delta \in (0,1)$, we can solve for m and get:

$$|H| \cdot e^{-\varepsilon m} < \delta$$

$$\Rightarrow \boxed{m > \frac{1}{\varepsilon} \cdot \log\left(\frac{|H|}{\delta}\right)}$$

error

confidence