

STATISTICAL LEARNING THEORY

Summer Term 2024

Resource: Shai Shalev-Schwartz

Understanding Machine Learning

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

OTHER COURSE NAMES:

- 1) Advanced machine learning
- 2) Machine learning

MOTIVATION

A motivating example:

	Weight (in g)	Color ($\in \{0, 1\}$)	Tasty
Papaya 1	800	0.1	0
⋮	⋮	⋮	⋮
Papaya N	1200	0.9	1

0... Not tasty
1... Tasty

Let's call the information on this table our **training data**. Based on that we aim to find

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

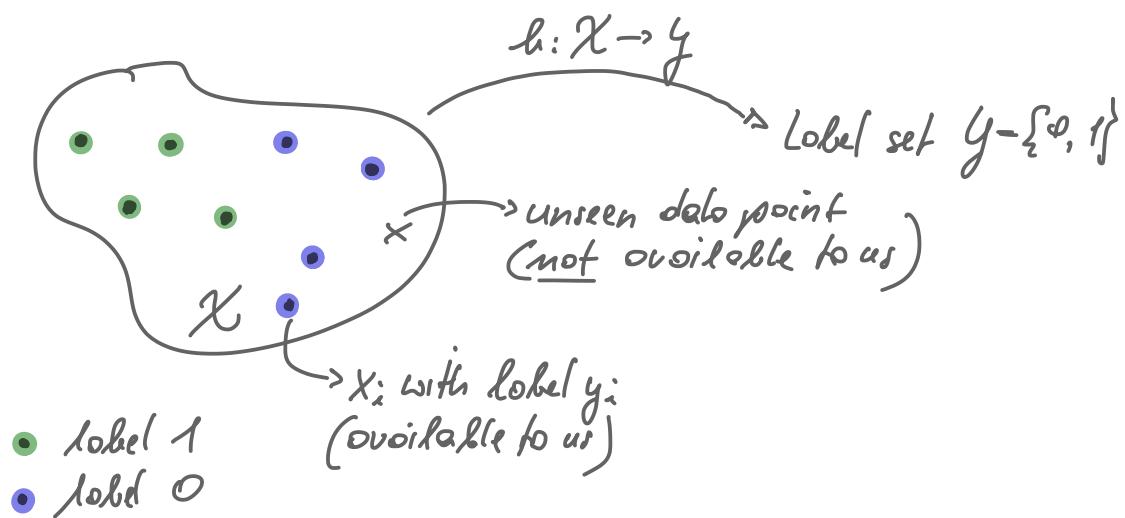
i.e., a function that will take a two-dim. vector as input (weight and color) and output a prediction of whether a papaya is tasty (1) or not (0). We call such a function a **hypothesis**.

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

$$((x_1, y_1), \dots, (x_N, y_N)) = S$$

with $x_i \in \mathbb{R}^2$ and $y \in \{0, 1\}$. \hookrightarrow Label set $Y = \{0, 1\}$
 \hookrightarrow Domain $X = \mathbb{R}^2$

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

identically and independently (iid)

from some (unknown) distribution D over the domain \mathcal{X} .

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

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

which we call the true labeling function, meaning

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

What do we care about?

We care about

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

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

BACKGROUND

(a really short overview
on required concepts)

In the following, $\mathcal{P}(X)$ will denote the power set of X .

Remember that we care about

$$A = \left\{ x \in X : h(x) + f(x) \right\}$$

\hookrightarrow domain

In case of $X = \mathbb{R}^n$ and $A \in \mathcal{P}(X)$, we already have a problem.

Measure Problem: find $\mu: \mathcal{P}(\mathbb{R}^n) \rightarrow [0, \infty]$ with properties:

1. $A_i \in \mathcal{P}(\mathbb{R}^n), i \in \mathbb{N}$, pairwise disjoint, we want

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

2. if $A, B \in \mathcal{P}(\mathbb{R}^n)$ are congruent, we want

$$\mu(A) = \mu(B)$$

3. we want

$$\mu([0, 1]^n) = 1$$

The measure problem is unsolvable for all $n \in \mathbb{N}$!

Solution: We will constrain ourselves to elements of some σ -Algebra over the domain X .

Def. (σ -Algebra): Let S be a non-empty set. A family of sets $\mathcal{F} \subset P(S)$ is called a σ -Algebra over S , if

1. $S \in \mathcal{F}$

2. from $A \in \mathcal{F}$, it follows that $A^c = S \setminus A \in \mathcal{F}$

3. from $A_i \in \mathcal{F}$ with $i \in \mathbb{N}$, it follows that

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \quad (\text{i.e. for any countable collection})$$

Example(s) (smallest σ -Alg. that contains A):

$$\{\emptyset, A, A^c, S\}$$

$$S = \{q, b, c, d\} \text{ with } \sigma\text{-Algebra } \left\{ \emptyset, \{q, b\}, \{c, d\}, \{q, b, c, d\} \right\}$$

Def. (generator): Given $\mathcal{E} \subset P(S)$ a family of sets and Σ denoting the set of all σ -Algebras that contain \mathcal{E} , we call

$$\sigma(\mathcal{E}) = \bigcap_{\mathcal{F} \in \Sigma} \mathcal{F}$$

the σ -Alg. generated by \mathcal{E} . Further, if it holds that for a σ -Alg. \mathcal{A}

$$\sigma(\mathcal{E}) = \mathcal{A}$$

we call \mathcal{E} the generator of \mathcal{A} .

Example : $\mathcal{E} = \{\{\{1\}\}\}$, $\mathcal{S} = \{1, 2, 3\}$

$$\begin{aligned}\mathcal{O}(\mathcal{E}) &= \mathcal{O}(\{\{\{1\}\}\}) \\ &= \{\emptyset, \{\{1\}\}, \{\{1, 2\}\}, \{\{1, 2, 3\}\}\}\end{aligned}$$

Def. (Topological Space) : A topological space is a tuple (X, τ) with X a set and τ a collection of subsets of X s.t.

1. $\emptyset \in \tau$, $X \in \tau$

2. closed under union, i.e., $\{\overline{T_i}\}_{i \in I} \subseteq \tau \Rightarrow \bigcup_{i \in I} \overline{T_i} \in \tau$

3. closed under finite intersection, i.e.,

$$\{\overline{T_i}\}_{i=1}^n \subseteq \tau \rightarrow \bigcap_{i=1}^n \overline{T_i} \in \tau$$

Remark: The elements of τ are called open sets.

Example: $X = \{1, 2, 3, 4\}$

$$\tau = \{\emptyset, \{1, 2, 3, 4\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$

Def. (BOREL σ-Alg.): Given a topological space (S, \mathcal{Q}) with \mathcal{Q} denoting the system of open sets, then we call

$$\sigma(\mathcal{Q}) =: \mathcal{B}(S)$$

the BOREL σ-Algebra over S. Its elements are called BOREL sets. For $S = \mathbb{R}^n$, we write

$$\mathcal{B}^n := \mathcal{B}(\mathbb{R}^n)$$

Importantly, each of the following systems of sets are generators for \mathcal{B}^n :

① $\{U \subset \mathbb{R}^n : U \text{ open}\}$

② $\{A \subset \mathbb{R}^n : A \text{ closed}\}$

③ $\{[q, b] : q, b \in \mathbb{R}^n \text{ with } q \leq b\}$

④ $\{[-\infty, c] : c \in \mathbb{R}^n\}$

meantif
 $[q, b] = [q_1, b_1] \times \dots \times [q_n, b_n]$
 for $q = (q_1, \dots, q_n)$
 $b = (b_1, \dots, b_n)$

Convention: We extend \mathbb{R} by symbols " $+\infty$ ", " $-\infty$ " as

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$$

$$\overline{\mathcal{B}} = \sigma(\mathcal{B} \cup \{-\infty\} \cup \{+\infty\})$$

Def. (Measurable space): if \mathcal{F} is a σ -Algebra over S , we call (S, \mathcal{F}) a measurable space.

Example: $(\mathbb{R}^n, \mathcal{B}^n)$

Def. (Measurable function): Given (S_1, \mathcal{F}_1) and (S_2, \mathcal{F}_2) measurable spaces, we call

$$f: S_1 \rightarrow S_2$$

a measurable function if

$$\forall E \in \mathcal{F}_2 : f^{-1}(E) \subset \mathcal{F}_1$$

Example: $\mathbb{1}_A: S \rightarrow \mathbb{R}$ // characteristic function

$$S \ni \omega \mapsto \mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{else} \end{cases}$$

Consider $\mathbb{1}_A$ as a function mapping to \mathbb{R} :

$$\text{if } B \leq 0, \text{ then } \{\omega : \mathbb{1}_A(\omega) < B\} = \emptyset$$

$$\text{if } B > 1, \text{ then } \{\omega : \mathbb{1}_A(\omega) < B\} = S$$

$$\text{if } 0 < B \leq 1, \text{ then } \{\omega : \mathbb{1}_A(\omega) < B\} = S \setminus A = A^c$$

$\Rightarrow \mathbb{1}_A$ is measurable if $A \in \mathcal{F}$.

↑
all
measurable

Def. (Measure): let (S, \mathcal{F}) be a measurable space.
A function $\mu: \mathcal{F} \rightarrow \overline{\mathbb{R}}$ is called a measure if the following conditions hold:

$$1. \mu(\emptyset) = 0$$

$$2. \mu(A) \geq 0 \text{ for all } A \in \mathcal{F}$$

3. for every sequence $(A_n)_{n \in \mathbb{N}}$ of disjoint sets from \mathcal{F} , we have

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \quad // \theta\text{-Additivity}$$

Example (counting measure): $\mu_{\text{count}}: \mathcal{F} \rightarrow \overline{\mathbb{R}}$

$$A \mapsto \begin{cases} |A|, & \text{if } A \text{ is finite} \\ \infty, & \text{else} \end{cases}$$

Def. (Measure space): if (S, \mathcal{F}) is a measurable space and $\mu: \mathcal{F} \rightarrow \overline{\mathbb{R}}$ is a measure, we call (S, \mathcal{F}, μ) a measure space.

Some elementary properties : Let (S, \mathcal{F}, μ) be a measure space. Also, let $A, B \in \mathcal{F}$ and $A_n \in \mathcal{F}$ with $n \in \mathbb{N}$. Then, it holds that

1. if A and B are disjoint, then

$$\mu(A \cup B) = \mu(A) + \mu(B)$$

2. if $A \subset B$ and $\mu(A) < \infty$, then

$$\mu(\underbrace{B \setminus A}_{\text{set difference}}) = \mu(B) - \mu(A)$$

3. if $A \subset B$, then

$$\mu(A) \leq \mu(B)$$

4. $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n)$ // called σ -Additivity



Def. (Probability space): Given (S, \mathcal{F}, D) a measure space and $D(S) = 1$,

then we call (S, \mathcal{F}, D) a probability space, and D a probability measure.

Remarks: S ... interpreted as the set of outcomes of a random experiment

\mathcal{F} ... interpreted as the set of events to which we want to assign probabilities to.

D ... the prob. measure assigns to each event $E \in \mathcal{F}$ a probability $D(E) \in [0, 1]$

Another remark: Based on $(S_1, \mathcal{F}_1, \mu)$ a measure space, a measurable space (S_2, \mathcal{F}_2) and a measurable function $f: S_1 \rightarrow S_2$, we can easily construct a new measure ν :

$$\nu_f: \mathcal{F}_2 \rightarrow [0, \infty]$$

$$B \mapsto \nu_f(B) = \mu(f^{-1}(B))$$

pre-image of B under f

We call ν_f the "push-forward" measure of μ under f .

Def. (Random Variable): Given $(S_1, \mathcal{F}_1, \mathbb{P})$ a prob. space and (S_2, \mathcal{F}_2) a measurable space, we call a measurable function

$$X: S_1 \rightarrow S_2$$

a random variable. If $S_2 = \mathbb{R}$, we say X is a real random var.

Also, the push-forward measure \mathbb{P}_X on S_2 is a prob. measure, since

$$\mathbb{P}_X(S_2) = \mathbb{P}\left(\underbrace{X^{-1}(S_2)}_{S_1}\right) = \mathbb{P}(S_1) = 1$$

We call \mathbb{P}_X the distribution of X .

CONVENTION(s) on notation:

X ... random variable

$$\{X \in A\} = \{\omega \in S : X(\omega) \in A\}$$

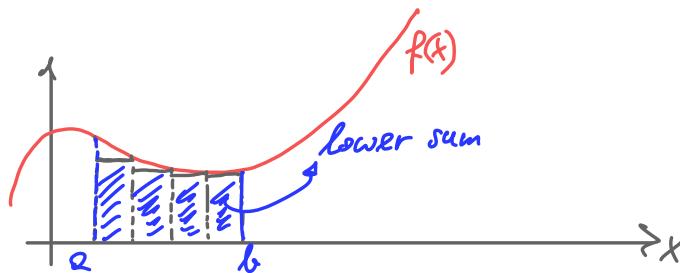
$$\{X = c\} = \{\omega \in S : X(\omega) = c\}$$

$$\mathbb{P}_X(A) = \mathbb{P}\left(\{\omega \in S : X(\omega) \in A\}\right)$$

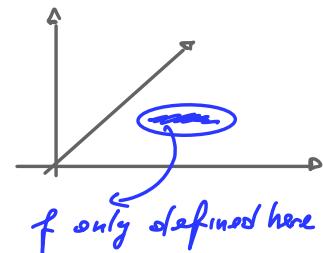
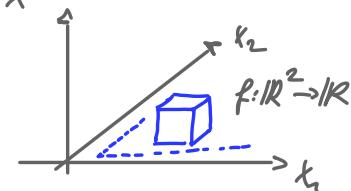
$$\mathbb{P}_X(\{c\}) = \mathbb{P}\left(\{\omega \in S : X(\omega) = c\}\right)$$

A very short primer on the LEBESGUE INTEGRAL

Key idea of Riemann integral: $f: \mathbb{R} \rightarrow \mathbb{R}$

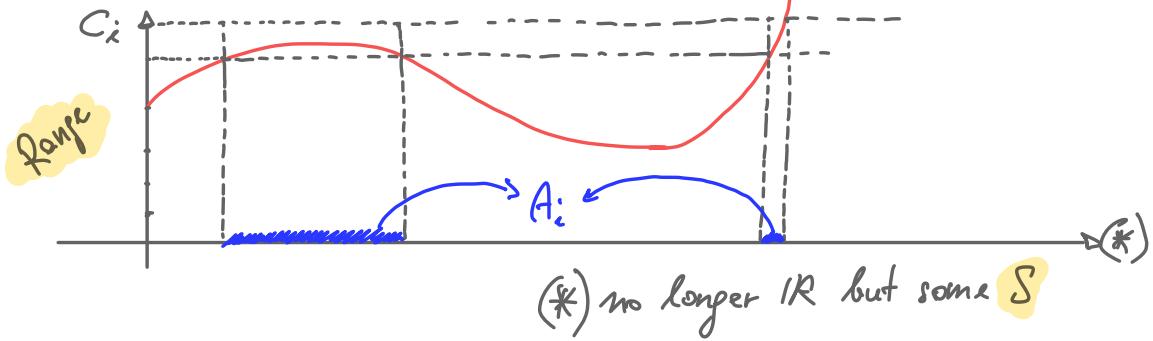


- 1) Difficult to extend to higher dimensions
- 2) Reliance on continuity
- 3) :



Question: How to partition this domain?

Idea of LEBESGUE interpretation: Partition the range of a function $f: S \rightarrow \mathbb{R}$ into intervals to get to an approximation by so called "elementary functions".



We need a way to measure A_i ; If we can do that, we can write

$$c_i \cdot \mu(A_i)$$

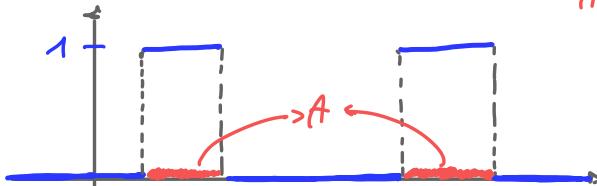
measure

.... this allows us to eventually write

$$\sum_i c_i \cdot \mu(A_i) \quad \text{OR} \quad \int_S f d\mu$$

A little bit more detail: Let's look at the case of "simple" functions.
Soy, we have (S, \mathcal{F}, μ) , $\mu: \mathcal{F} \rightarrow [0, \infty]$

$$\mathbb{1}_A: S \rightarrow \mathbb{R}, A \in \mathcal{F}$$

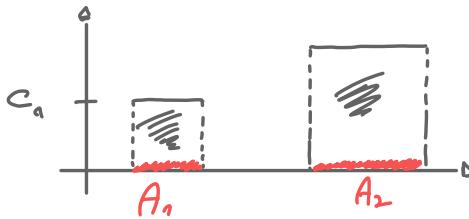


Any reasonable notion of interpretation of $\mathbb{1}_A$ should return $\mu(A)$:

$$\text{Integral} \quad \mathbb{I}(\mathbb{1}_A) = \mu(A)$$

finite here!

Now, a "simple" function can be written as $f(x) = \sum_{i=1}^n c_i \cdot \mathbb{1}_{A_i}(x)$ with $A_i \in \mathcal{F}$ and $c_i \in \mathbb{R}$. Visually:



We can define the integral of f via: $\mathbb{I}(f) = \sum_{i=1}^n c_i \mu(A_i)$. But, since $\mu: \mathcal{F} \rightarrow [0, \infty]$ and $c_i \in \mathbb{R}$, we could get something like

$$10 \cdot \infty - 8 \cdot \infty$$

Solution: $\{f: S \rightarrow \mathbb{R} \mid f \text{ is "simple" and } f \geq 0\} = T^+$. For $f \in T^+$ we have a well-defined representation of f as

$$f(x) = \sum_{i=1}^n c_i \cdot \mathbb{1}_{A_i}(x) \text{ with } c_i \geq 0$$

and we define the LEBESGUE integral (for this class) as:

$$\mathbb{I}(f) := \sum_{i=1}^n c_i \cdot \mu(A_i) \quad \text{or} \quad \int f d\mu \quad \text{or} \quad \int f(x) d\mu(x)$$

Some properties: (that follow immediately)

for $f, g \in T^+$ and $\alpha, \beta \geq 0$

- $\int(\alpha f + \beta g) d\mu = \alpha \cdot \int f d\mu + \beta \cdot \int g d\mu$ LINEARITY

for $f, g \in T^+$ and $f \leq g$

- $\int f d\mu \leq \int g d\mu$ MONOTONICITY

Overall, even though we just looked at "simple" functions, the construction extends to all measurable functions $f: S \rightarrow \mathbb{R}$.

Def. (Expected value of a RV): Given a prob. space (S, \mathcal{F}, D) and a quasi-integrable real RV $X: S \rightarrow \mathbb{R}$, then

$$\mathbb{E}[X] = \int X dD$$

is called the expected value of X .

We have:

- If $X \geq 0$, then $\mathbb{E}[X] \geq 0$
- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for RV's X and Y
- ⋮

Def. (MARKOV inequality): Given prob. space (S, \mathcal{F}, D) and a non-negative RV $X (\geq 0)$, we have for $q > 0$

$$\underbrace{\mathbb{P}[X \geq q]}_{\mathcal{D}(\{\omega \in S: X(\omega) \geq q\})} \leq \frac{\mathbb{E}[X]}{q}$$

Proof: Define $\varphi: S \rightarrow \mathbb{R}$ with $\varphi(x) = \begin{cases} q, & \text{if } X(x) \geq q \\ 0, & \text{else } (X(x) < q) \end{cases}$

$\rightarrow 0 \leq \varphi(x) \leq X(x)$. Hence, by monotonicity

we have $\int X dD \geq \int \varphi dD = q \cdot \mathcal{D}(\{\omega \in S: X(\omega) \geq q\})$.

\Rightarrow as $q > 0$, we have $\frac{1}{q} \cdot \int X dD \geq q \cdot \mathcal{D}(\{\dots\})$

and since $D(\{\omega \in S : X(\omega) \geq Q\}) = P[X \geq Q]$, we
get $P[X \geq Q] \leq \frac{\int X dD}{Q} = \frac{E(X)}{Q}$ □

CONTINUING ML MATERIAL