THE MACHINE LEARNING FRAMEWORK

"A computer program is said to learn from experience E with respect to some task T, and some performance measure P if its performance on T, as measured by P, improves with experience E"

— T.M. Mitchell, 1997

For Machine Learning we need:

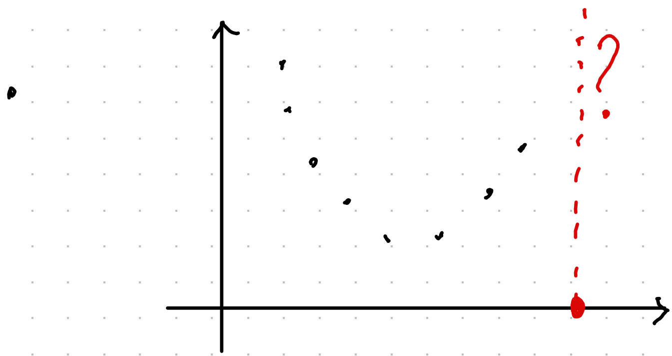
- a task (e.g., regression, classification, anomaly detection, prediction, etc.)
- experience (a dataset)
- a model
- a performance measure (a loss function)
- improvement

Typical aim of a machine learning problem:

find a function $y = f(x)$ (the task)

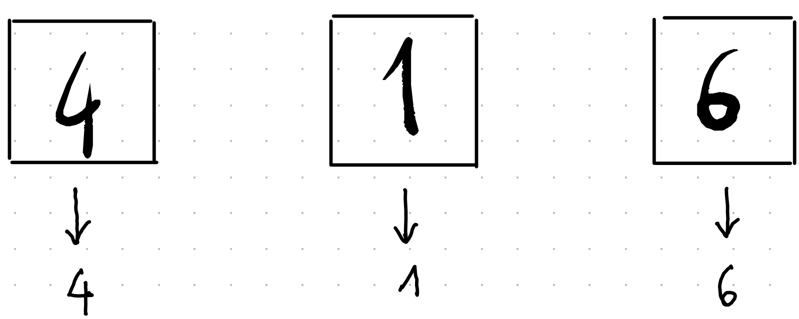
$f : \mathbb{R}^{M_{in}} \longrightarrow \mathbb{R}^{M_{out}}$

Prediction



- The pen is on the ______ .

Classification



Anomaly detection

Finding $y = f(x)$ among all possible functions is not feasible.

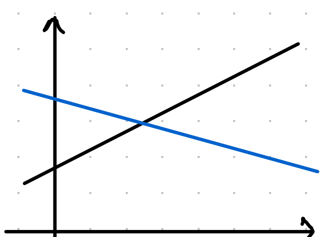
A model is chosen:

$$f(x; w)$$

a class of functions described by some parameters.

Examples:

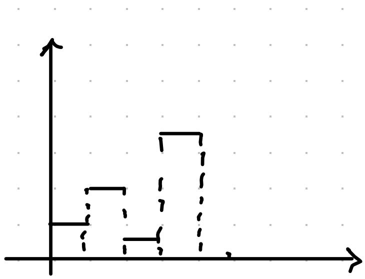
1)



$y = ax + b$

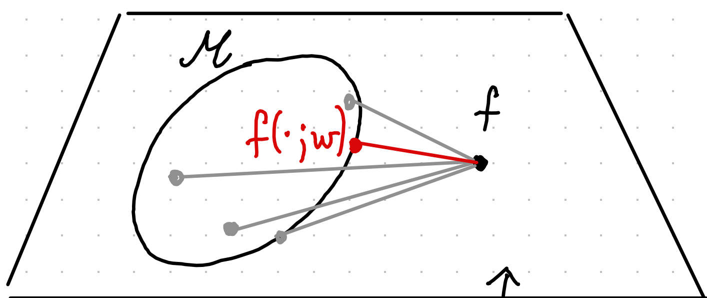
parameters: a, b

2)



parameter: heights

<u>New aim</u>: Among all possible functions in the model class, find an approximation $f(x;w)$ of $f(x)$.



all possible functions

Which one to choose? We want the "best" approximation.
"Best" according to a functional that measures how far we are from f.

To define this functional, first of all we have a loss function

$$\ell(y_{pred}, y_{true})$$

that allows us to measure how much a prediction is "distant" from the true values.

In this way, given a model $f(x;w)$, we are able to measure

$$\ell(\underbrace{f(x;w)}_{\text{predicted output}}, \underbrace{f(x)}_{\text{true output}})$$

In some sense, we want to sum (or integrate) over all possible input x's.

Unfortunately, we don't have access to <u>all</u> possible inputs, but we have data.

A single instance of a datum can be thought as a realization of a random variable $X : (\Omega, \mathbb{P}) \longrightarrow \mathbb{R}^{M_{in}}$

When we do a measurement in an experiment, we collect a datum x, which means that we are observing the event $X = x$, i.e., a realization of the random variable X.

(We do another experiment, we observe a different event).

This realization gives the loss

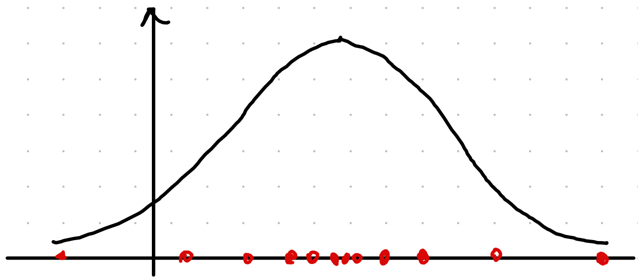
$$\ell(f(x;w), f(x))$$

This is nothing but a realization of the random variable

$$\ell(f(X;w), f(X))$$

We want to "sum" over all possible realizations of this random variable.

However, the random variable X has some probability distribution, i.e., data is distributed according to some law.



observed data can be distributed according to some law

When we sum the losses of predictions, we have to weight the possible observations with the probability that they actually occur!

Case of discrete distribution:

$$\sum_x \ell(f(x;w), f(x))\, \mathbb{P}(X=x)$$

Case of continuous distribution:

there is a probability density function $p(x)$

$$\int_{\mathbb{R}^{M_{in}}} \ell(f(x;w), f(x))\, p(x)\, dx$$

Concretely, we are computing

$$\mathbb{E}[\ell(f(X;w), f(X))]$$

typically called <u>risk</u>.

The <u>new aim</u> becomes: find an approximation $f(\cdot;w)$ of f with lowest risk.

(Interpretation: such an approximation is such that, typically, the loss for using $f_w(x)$ on an observation is low).

Problem 1: we don't know $f(x)$! So we could never compute this loss.

Way out: when we measure a datum, we observe both input x and output y.

We relax the hypothesis that $y = f(x)$ and think of (x, y) as an observation of a random variable $(X, Y) : (\Omega, \mathbb{P}) \longrightarrow \mathbb{R}^{M_{in}} \times \mathbb{R}^{M_{out}}$.

Discrete case:

$$\sum_{(x,y)} \ell(f(x;w), y)\, \mathbb{P}(X = x, Y = y)$$

Continuous case:

$$\int_{\mathbb{R}^{M_{in} \times M_{out}}} \ell(f(x;w), y) \underbrace{p(x,y)}_{\text{probability density function}} dx\, dy$$

This means that the risk is:

$$\mathbb{E}[\ell(f(X;w),Y)]$$

<u>Aim</u> (in mathematical terms):

$$\min_{w} \mathbb{E}[\ell(f(X;w),Y)]$$

If we are able to find this minimum, we commit an error given by

$$\mathbb{E}[\ell(f(X;w),Y)]$$

<u>Problem 2</u>: We don't know the probability distribution of data...
We can never compute the risk.

The only thing we can do is estimating it.

To do so, we use data.

A dataset $(x_0, y_0), \ldots, (x_{N-1}, y_{N-1})$ is the realization of a random sample $(X_0, Y_1), \ldots, (X_{N-1}, Y_{N-1})$, i.e., i.i.d. random variables, all distributed with the data distribution.

The random variable (<u>empirical risk</u>)

$$\frac{1}{N} \sum_{i=0}^{N-1} \ell(f(X_i; w), Y_i)$$

is an unbiased estimator of the risk, i.e.,

$$\mathbb{E}\left[\frac{1}{N} \sum_{i=0}^{N-1} \ell(f(X_i; w), Y_i)\right] =$$

$$= \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{E}\left[\ell(f(X_i; w), Y_i)\right] =$$

$$= \mathbb{E}\left[\ell(f(X; w), Y)\right]$$

But, as the empirical average, has low variance

$$\mathrm{Var}\left[\frac{1}{N} \sum_{i=0}^{N-1} \ell(f(X_i; w), Y_i)\right] =$$

$$= \frac{1}{N} \mathrm{Var}\left[\ell(f(X; w), Y)\right]$$

We can estimate the risk with the realization of the empirical risk on the dataset

$$\frac{1}{N}\sum_{i=0}^{N-1} \ell(f(x_i;w), y_i)$$

New aim: Given the dataset $\{(x_i, y_i)\}_{i=0,\dots,N-1}$ find the approximation $f(\cdot\,;w)$ in the model class that minimizes the empirical risk.

By finding the minimum of the empirical risk, we make a statistical error, on top of the modeling error

$$\mathbb{E}\left[\ell(f(X;w), Y)\right] +$$

$$+ \left|\frac{1}{N}\sum_{i=0}^{N-1} \ell(f(x_i;w), y_i) - \mathbb{E}\left[\ell(f(X;w), Y)\right]\right|$$

Problem 3: Computing the minimum

$$\min_{w} \frac{1}{N} \sum_{i=0}^{N-1} \ell(f(x_i; w), y_i)$$

cannot be done explicitly.

We need to resort to a numerical method to find an approximation w^* of this error.

Hence, the full error is

$$\mathbb{E}[\ell(f(X; w), Y)] +$$

$$+ \left| \frac{1}{N} \sum_{i=0}^{N-1} \ell(f(x_i; w), y_i) - \mathbb{E}[\ell(f(X; w), Y)] \right|$$

$$+ \left| \frac{1}{N} \sum_{i=0}^{N-1} \ell(f(x_i, w^*), y_i) - \frac{1}{N} \sum_{i=0}^{N-1} \ell(f(x_i; w), y_i) \right|$$