

# Machine Learning

## Module 3.0 - Models: Introduction

Arnold Vialfont

Master in Management, Business Analytics, HEC UNIL

Spring 2026

# Context

In ML, models are mainly used for **supervised learning**, the aim is

- Predict a response  $y$ : regression if numerical, classification if categorical.
- From features  $x = \{x_1, \dots, x_p\}$ : available at the moment of prediction,
- With the best possible quality: built from the data in an optimal way.

The  $n$  observed features and responses are denoted

$$(y_1, x_1), \dots, (y_n, x_n).$$

# Elements

In ML, a model consists mainly of three elements:

- A **prediction** formula  $f$ , taking the features  $x$ , returning a prediction  $f(x)$  for  $y$  at  $x$ ,
- A **loss function**  $\mathcal{L}(y, \hat{y})$  measuring how "wrong" a prediction  $\hat{y}$  is for  $y$ .
- A **training algorithm** which optimizes the prediction formula  $f$  with respect to the loss function using the observed data.

# The prediction formula

The prediction formula  $f$  is a mathematical formula, possibly a complex algorithm, using **parameters**<sup>1</sup>  $\theta$ . The prediction is written

$$f(x; \theta).$$

Thus,  $\theta$  must be chosen carefully to obtain good predictions of  $y$ .

---

<sup>1</sup>Also called **weights**, especially for Neural Networks.

# The loss function

The loss function indicates how wrong is a prediction  $\hat{y}$  of the corresponding  $y$ : a large  $\mathcal{L}(y, \hat{y})$  indicates a poor prediction of  $y$  by  $\hat{y}$ .

A classical example for regression is the square of the error:

$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2.$$

The larger  $\mathcal{L}(y, \hat{y})$ , the further  $\hat{y}$  is from  $y$ .

# The optimal parameters

"Good parameters"  $\theta$  must achieve a low loss on the overall data base. We want  $\mathcal{L}(y, f(x; \theta))$  to be small for all  $(y, x)$ . To achieve an overall quality on the whole available data base, we want  $\theta$  achieving a small:

$$\bar{\mathcal{L}}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}\{y_i, f(x_i; \theta)\}.$$

Example: with the square of the error, this is

$$\bar{\mathcal{L}}(\theta) = \frac{1}{n} \sum_{i=1}^n \{y_i, f(x_i - \theta)\}^2.$$

# The optimization algorithm

Finding the optimal  $\hat{\theta}$  is done by applying an algorithm, i.e., a procedure that finds

$$\hat{\theta} = \arg \min_{\theta} \bar{\mathcal{L}}(\theta).$$

The algorithm is often a sequential procedure. It builds a sequence  $\theta_1, \theta_2, \theta_3, \dots$  such that

$$\bar{\mathcal{L}}(\theta_1) > \bar{\mathcal{L}}(\theta_2) > \bar{\mathcal{L}}(\theta_3) > \dots$$

Ultimately, this should reach the minimum possible  $\bar{\mathcal{L}}(\theta)$ .

# Mathematical considerations

- More flexible model  $f$  provides better opportunity to minimize  $\bar{\mathcal{L}}$ . Often, this is associated with the size of  $\theta$  (number of parameters).
- The algorithm may not reach the global minimum of  $\bar{\mathcal{L}}$ . Most algorithms cannot guarantee such results except under theoretical assumptions.
- A probabilistic interpretation: the optimal  $\theta$  is obtained by minimizing the expected loss on the population of  $(Y, X)$

$$E [\mathcal{L}\{Y, f(X; \theta)\}].$$

The data base is used to estimate it with an empirical mean

$$\hat{E} [\mathcal{L}\{Y, f(X; \theta)\}] = \frac{1}{n} \sum_{i=1}^n \mathcal{L}\{y_i, f(x_i; \theta)\}.$$

This estimate is minimized in turn to find an estimate of the optimal  $\theta$ .