UniGe | MaLGa

Lab 0 – Data Generation

Basic definitions

Given a training set

$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$$S_n \subseteq X \times Y$$

$$X \subseteq \mathbb{R}^D$$

We want to estimate a function such that

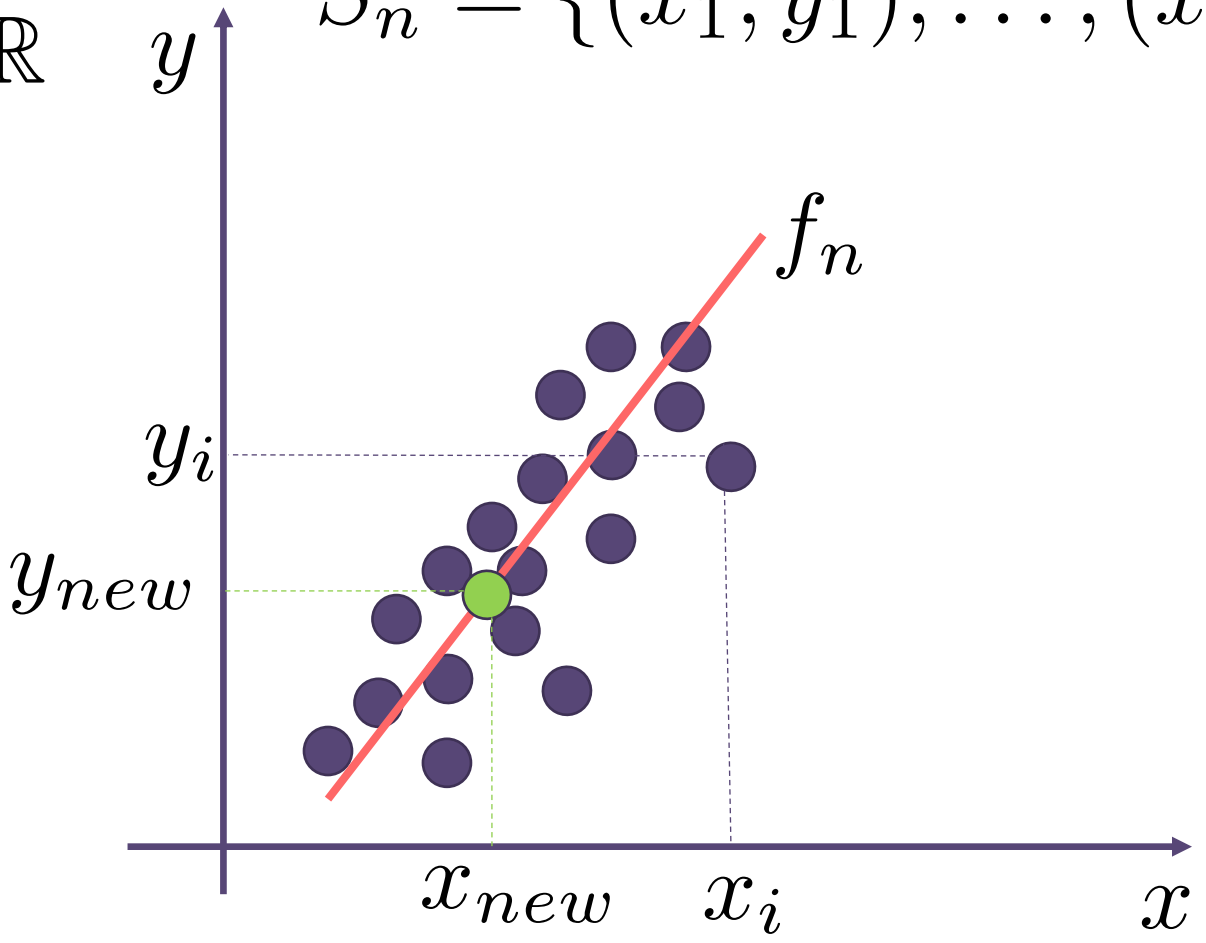
$$f_n(x_i) \sim y_i$$

What about Y?

Basic setting: regression

$Y \subseteq \mathbb{R}$

$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

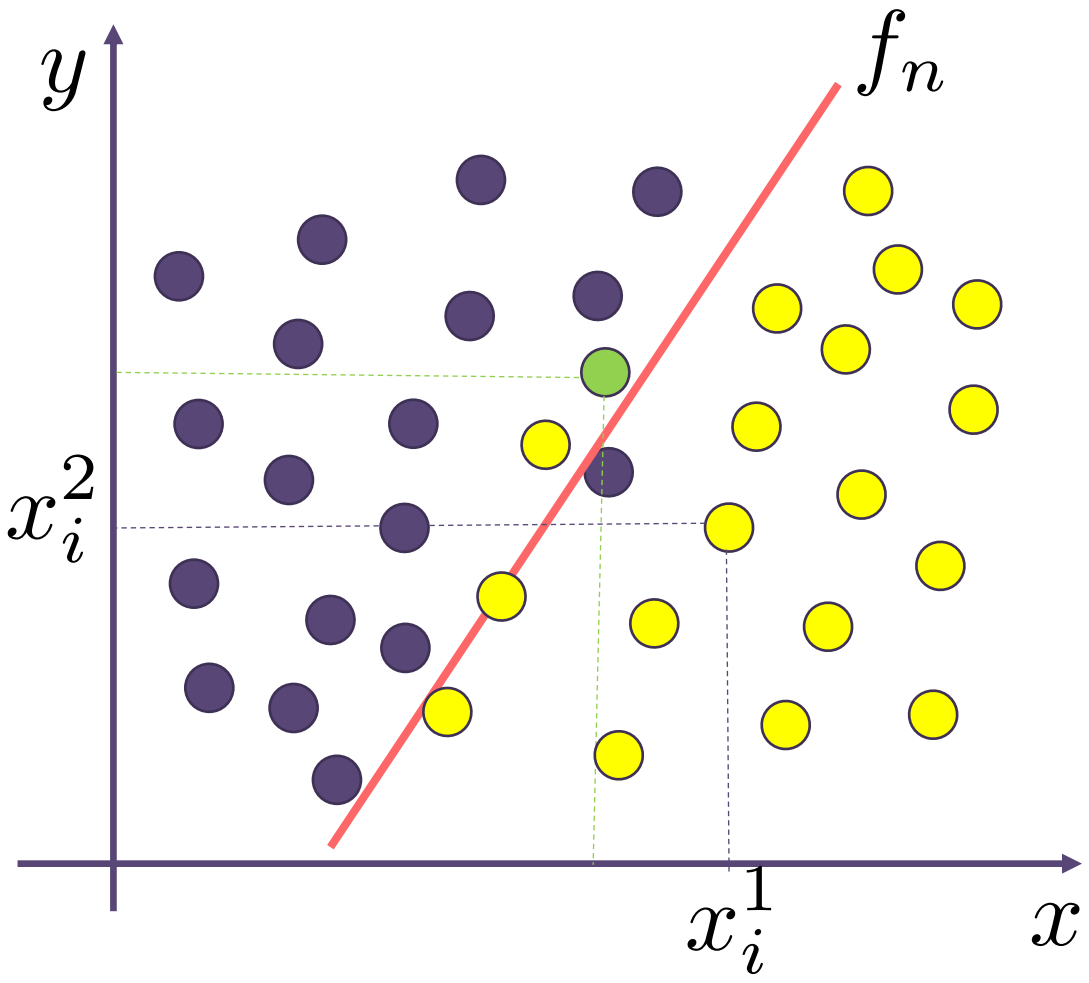


Basic setting: classification

$$Y = \{-1, 1\}$$

$$X \subseteq \mathbb{R}^2$$

$$x_i = [x_i^1, x_i^2]$$



$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

Where is the magic?

ASSUMPTION 1

The samples are i.i.d. according to a joint distribution p(x,y)

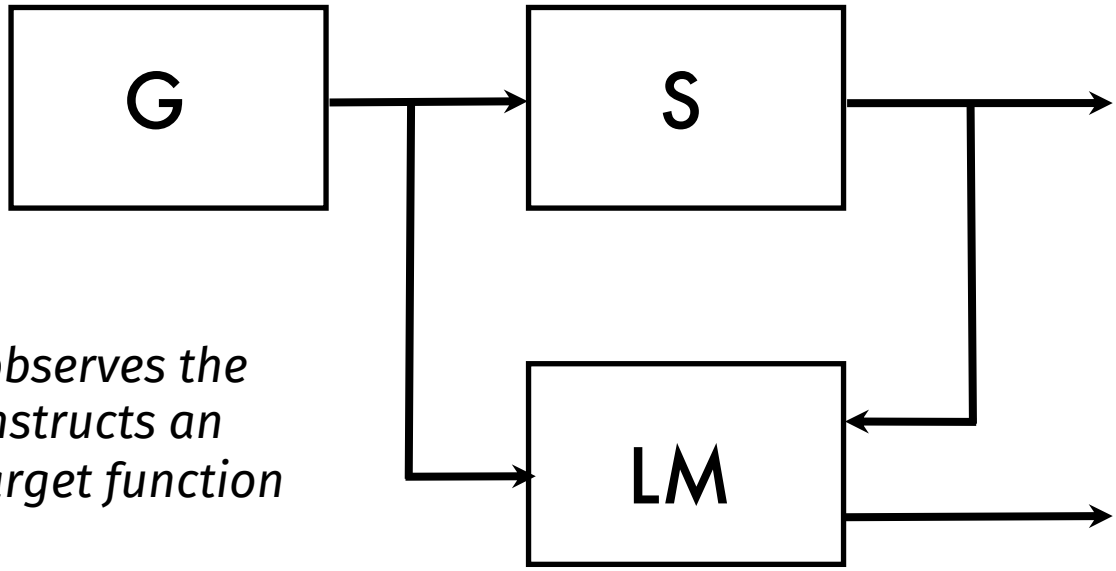
ASSUMPTION 2

The probability can factorize as $p(x,y) = p(x)\,p(y|x)$, including the case when $p(y|x)$ depends on a function f such that $f(x)=y$

Supervised Learning

GENERATOR: *generates vectors according to "certain rules" which are* ***unknown*** *but* ***fixed***

SUPERVISOR: *transforms the vectors into output values. It is* ***unknown*** *but* ***it exists*** *and* ***does not change***

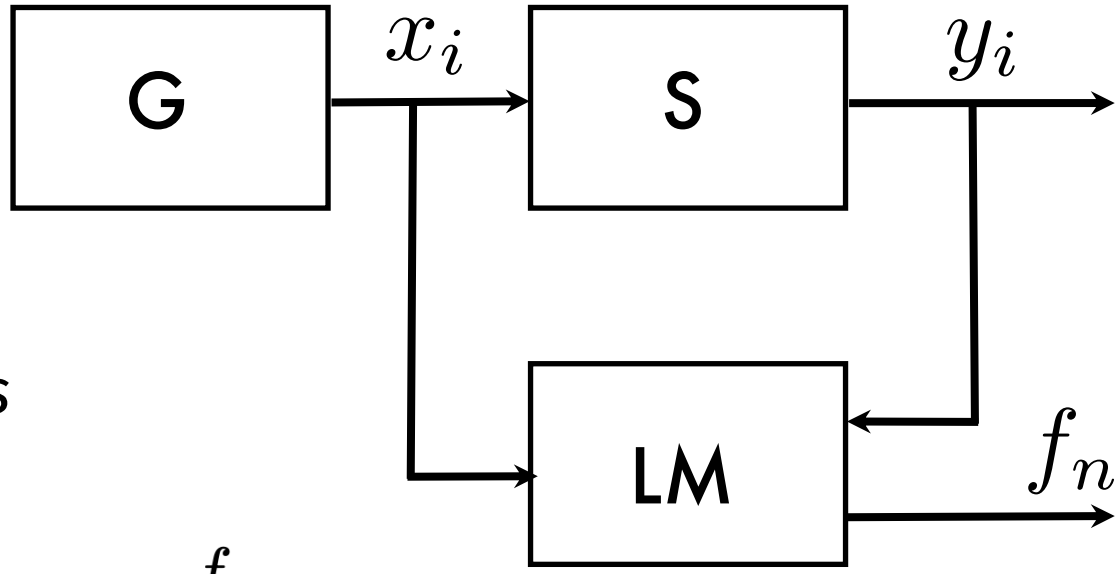


LEARNING MACHINE: *it observes the pairs of x and y and constructs an approximation of the target function*

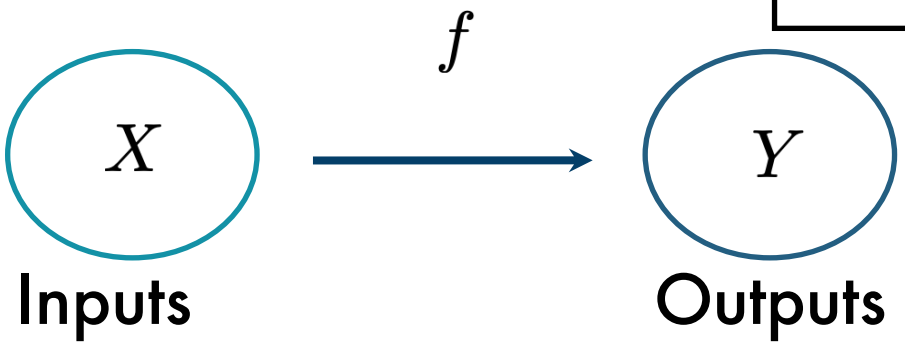
Supervised Learning

$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$$p(x) \qquad p(y|x) \rightarrow f(x) = y$$



We are happy if the estimated function is close to the target f



In this lab…

Play with data generation

- Explore different strategies for data generation for regression and classification problems
- Observe the effect of changing the number of sampled points (and other factors)
- Notice how samples change when noise becomes part of the game

Give a face to many names

- The role of some of the concepts mentioned so far (training set, probability distributions, input-output function) will become clear

How to proceed

- Go to 2021.aulaweb.unige.it and access to the Machine Learning module
- Download the notebook file for Lab 0
- Run the Jupyter Notebook and open the downloaded file
- In case you use Google Colab, you may simply load the notebook on your Google drive and open it

UniGe