Machine Learning
Introduction

Tech Lead Data Science

Master en Data Science
2022-2023

Assembler
Institute of Technology

Index

1 Introduction

2 Types of ML

3 Phases of ML

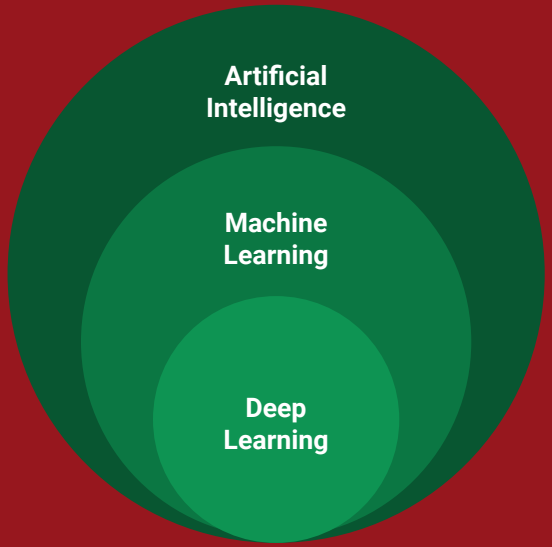
Introduction

What is Machine Learning ?

INTRODUCTION

Definition of Machine Learning

- Machine Learning is a discipline in the field of Artificial Intelligence in which, through massive data (Big Data), computers learn to identify patterns to offer prediction, all through algorithms.



Assembler
Institute of Technology

Everything is Artificial Intelligence

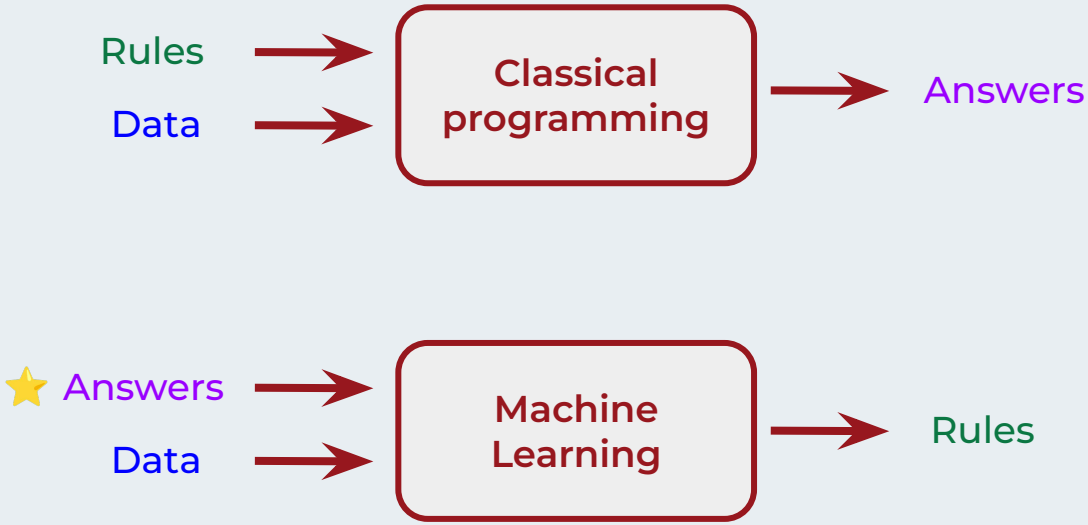
Machine Learning

Machine Learning requires the training of algorithms to provide a computer with "intelligence", in addition, this learning is carried out from experience (understanding as experience the data and parameters of an algorithm)

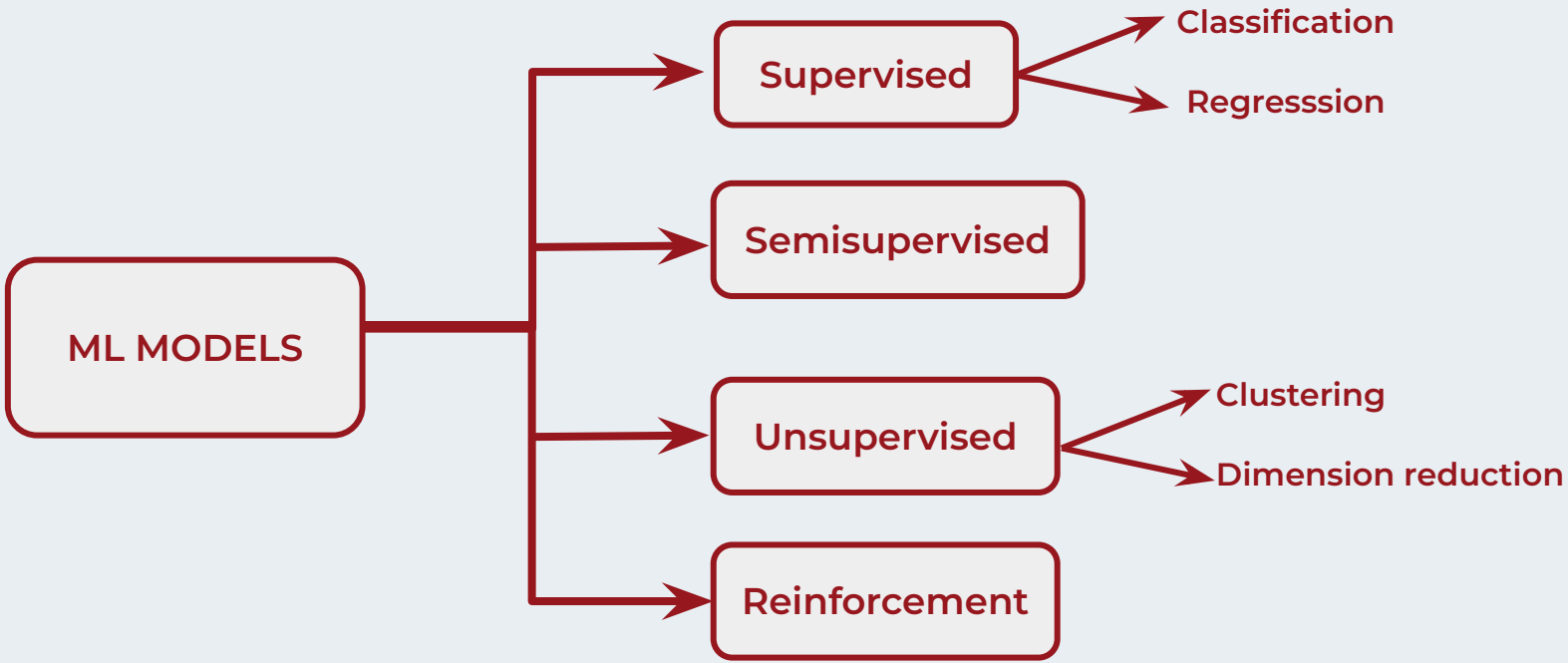
Artificial Intelligence

Artificial Intelligence, on the other hand, is simply that a computer imitates a human capacity, or makes it appear intelligent, this intelligence may require Machine Learning, Big Data, Statistics or Mathematics or simply programming.

ML OVERVIEW


Rules
Data
Classical programming
Answers
⭐ Answers
Data
Machine Learning
Rules

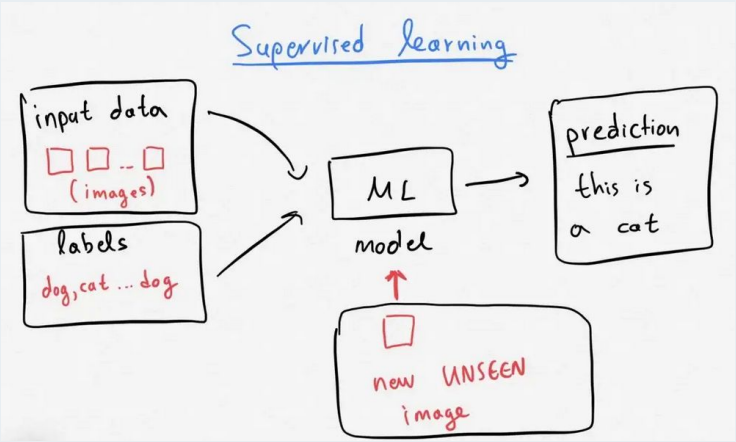
ML OVERVIEW



Assembler
Institute of Technology

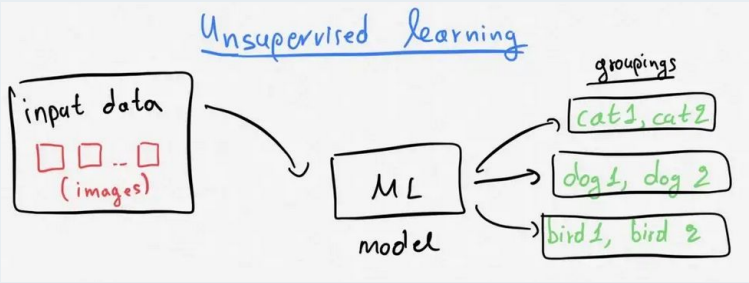
Supervised

- We have a data set with a response variable (target) previously labeled.
- Learning the algorithm will return one of the possible values of the target variable.
- They can be further grouped into:
  - **Classification:** A classification problem is when the output variable is a category e.g. “disease” / “no disease”.
  - **Regression:** A regression problem is when the output variable is a real continuous value e.g. stock price prediction



Unsupervised

- We do not have a target variable with the response data previously labeled
- The algorithm will be in charge of grouping the data by their similarity (or distance)
- The output of the algorithm will be a variable that labels the data sets.
- They can be further grouped into:
    - **Clustering:** Clustering involves grouping sets of similar data (based on defined criteria). It's useful for segmenting data into several groups and performing analysis on each data set to find patterns.
    - **Dimension reduction:** Dimension reduction reduces the number of variables being considered to find the exact information required.

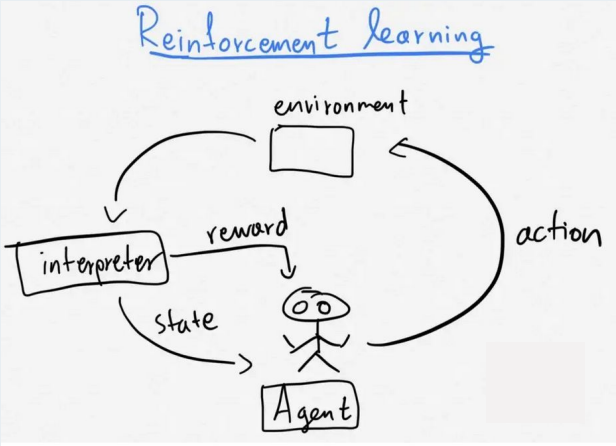


Semisupervised

- There are some labeled data and a large amount of unlabeled data.
- The algorithm tries to predict the labels of the unlabeled data with the aim of improving the quality of the data and improve the quality of the algorithms.

Reinforcement learning

- It does not have a target variable (so it is not supervised)
- The algorithm does not return a classified data set (so it is not unsupervised).
- This family of models consists of algorithms that use the **estimated errors as rewards or penalties.**
    - If the error is big, then the penalty is high and the reward low.
    - If the error is small, then the penalty is low and the reward high.
- Reward feedback is required for the model to learn which action is best and this is known as “the reinforcement signal”.



Phases of ML Project

Phases of ML Project

Modeling Phase

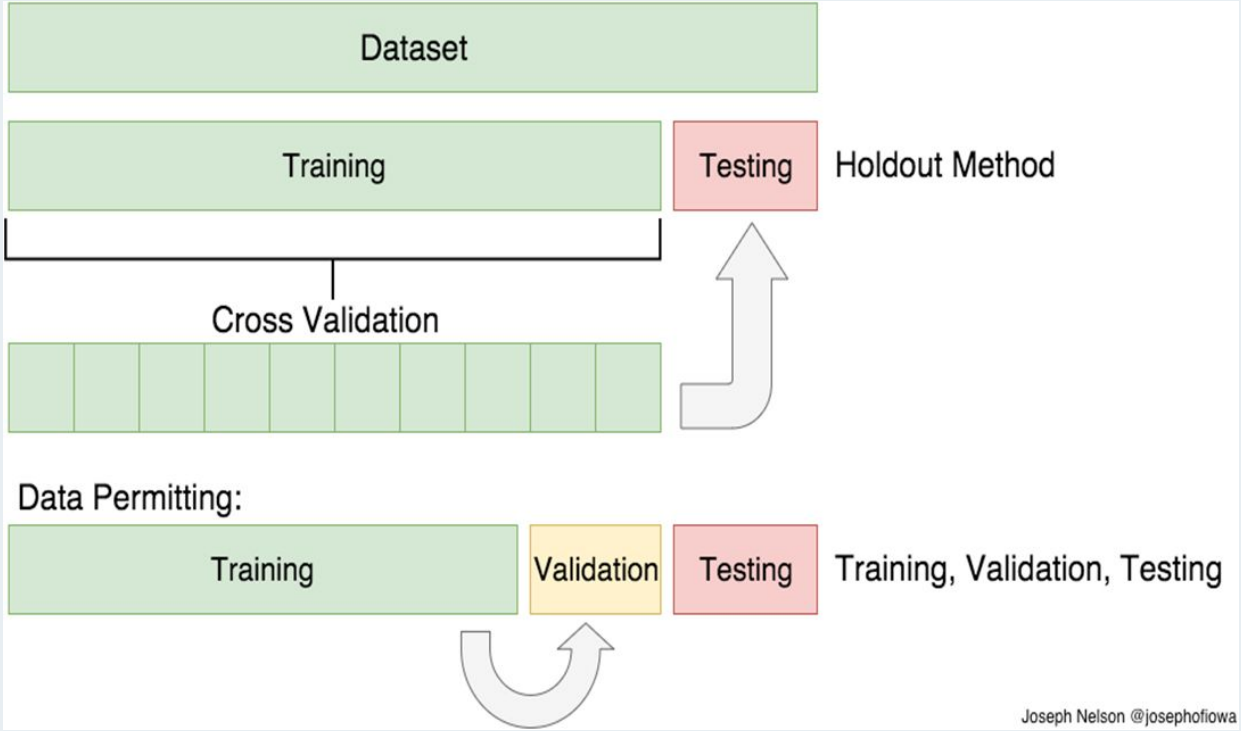
- We'll have a series of characteristics, attributes, columns or simply variables. They will be the **X variables** and they are called **independent** variables.
- In supervised learning, they will also be accompanied by a **dependent variable Y**.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	class
a	52.17	0	y	p	ff	ff	0	f	f	0	f	g	0	0	-
a	48.17	1.335	u	g	i	o	0.335	f	f	0	f	g	0	120	-
a	20.42	10.5	y	p	x	h	0	f	f	0	t	g	154	32	-
b	50.75	0.585	u	g	ff	ff	0	f	f	0	f	g	145	0	-
b	17.08	0.085	y	p	c	v	0.04	f	f	0	f	g	140	722	-
b	18.33	1.21	y	p	e	dd	0	f	f	0	f	g	100	0	-
a	32	6	u	g	d	v	1.25	f	f	0	f	g	272	0	-
b	59.67	1.54	u	g	q	v	0.125	t	f	0	t	g	260	0	+
b	18	0.165	u	g	q	n	0.21	f	f	0	f	g	200	40	+
b	37.58	0					0	f	f	0	f	p		0	+
b	32.33	2.5	u	g	c	v	1.25	f	f	0	t	g	280	0	-

Modeling Phase

- The goal of machine learning models is to find the function that best estimates an output from some new data inputs.
- As algorithms learn (by imitating) they need to be trained first, before they can predict any event.
- The first thing we will do is divide the dataset into:
  - **Training set:**
    The set to which the machine learning model fits
  - **Test set:**
    The one that is not used throughout the process until the final part, simply to check that the model generalizes correctly. It is left aside until the end, as if it were new data.
  - **Validation set:**
    It is almost always used. This set allows you to compare different models and choose the one that performs best.

Modeling Phase



Modeling Phase

- These splits of the data are done to avoid the most important problem in machine learning, which is called overfitting.
- **Overfitting** is the act of making a model that fits the training data so well that it does not generalize correctly to the test data. That is, it fits correctly to the past, but not to the future, which is the objective of the models.