eleven - Hackathon

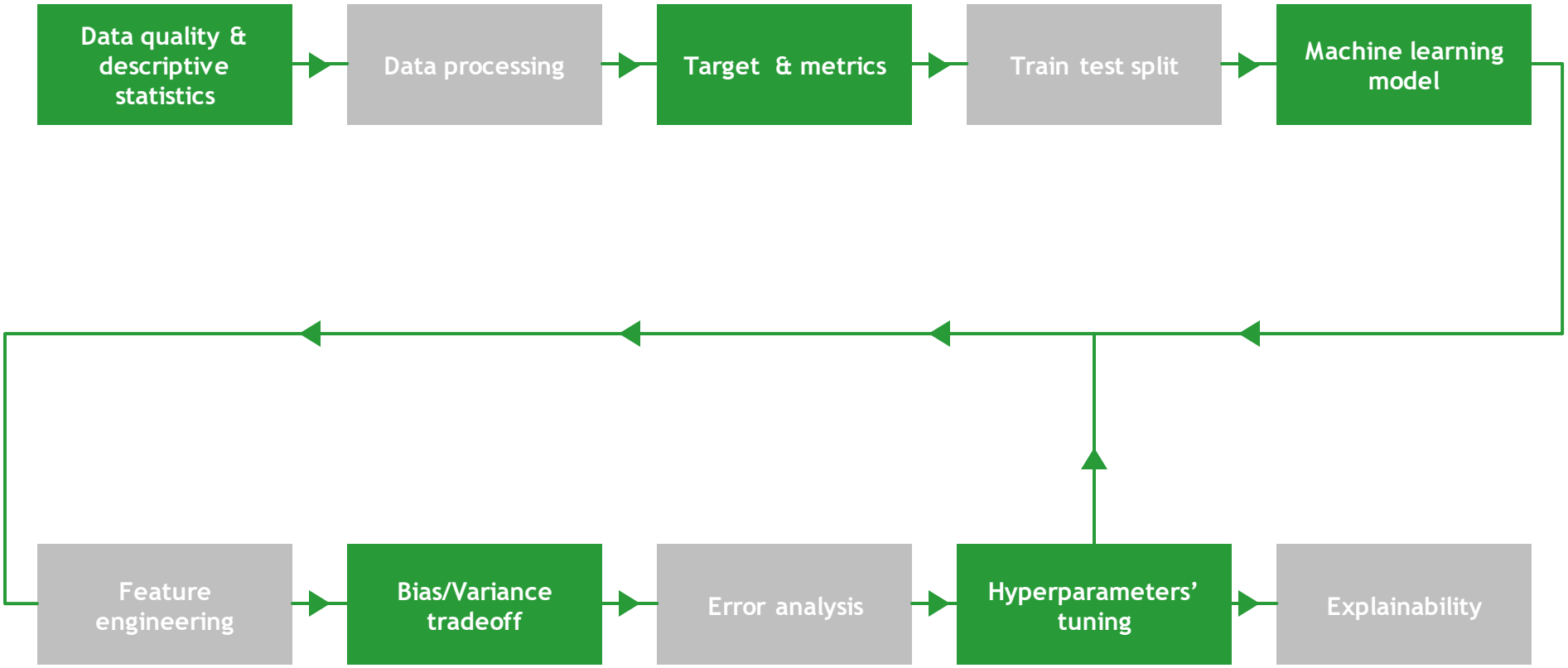
Machine Learning Basics

To the attention of the Ingénierie Mathématique & Informatique students

September 6th, 2022

eleven
strategy • data • digital

Building a rigorously proper machine learning model requires to follow a precise pipeline whereby the richness of the data is fully exploited



Artificial intelligence can be organized around three families of machine learning models

	Approach	Model	Goal
Machine learning	Supervised learning	Regression	Learn a function that maps an input to an output based on example input-output pairs (labeled data)
		Classification	
	Unsupervised learning	Clustering	Identify and uncover previously undetected patterns with no preexisting labeled data
		Collaborative filtering	
	Reinforcement learning	Q-table based	Find the optimal solution based on: An environment A set of rules A playing agent
		Deep learning based	

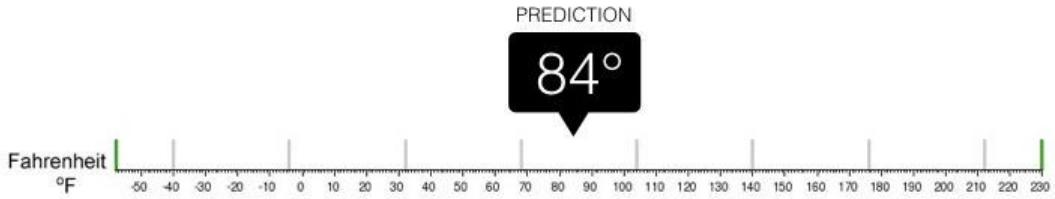
Not necessary for the hackathon

Regression models aim at modeling continuous variables whereas classification models aim at modeling categorical variables

Example of the difference between regression and classification

Regression

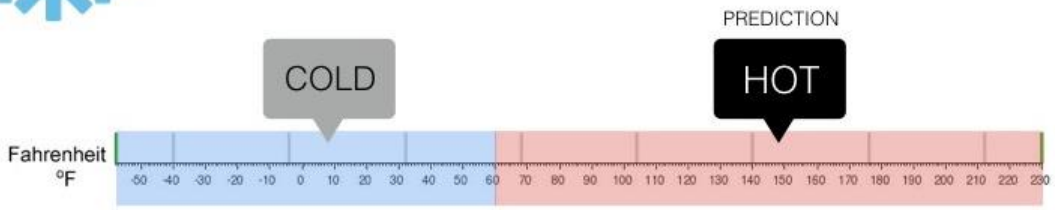
What is the temperature going to be tomorrow?



The output of the model is the temperature which is a continuous variable

Classification

Will it be Cold or Hot tomorrow?



The output of the model is a category: cold or hot

Machine learning models can be leveraged to fulfill different goals

Understand

- Understanding the behavior or a given process, e.g.,
  - The relation between temperature and a building infrastructure
  - The relation between the intensity of a press and the quality of the animal food
  - …

Predict

- Predicting the output of a process given a new occurrence, e.g.,
  - What will the energy price be in one week?
  - What is this task's time to completion?
  - …

Optimize

- Optimizing a process given a certain amount of information and constraints, e.g.,
  - Generating an efficient sprinkler network based on the building's blueprint and the involved regulatory principles
  - …

…

- If the goal is to understand a behavior, then the value is in the estimation of the parameters
  (the causal parameters)

- If the goal is to predict an output or optimize a process, then the value is in the goodness of the output estimation
- It would be enough to have a « good » estimation of the parameter to be able to make accurate predictions or efficient optimizations

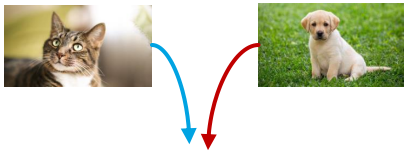
A machine learning model refers to both a group of functions to solve a task, and the chosen function within that group

Illustration of machine learning models

Is there a cat in this picture?

What could be an individual's revenue

Input space: Images



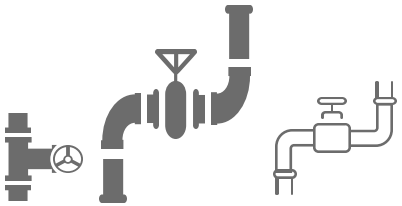
Education level

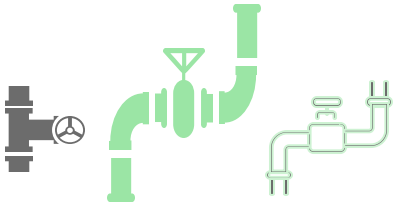
Professional experience

Age

Input space: Tabular data

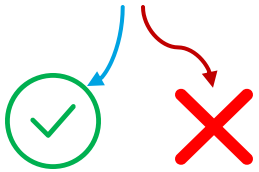
Model: Maps data from the input space into the output space





Model: Maps data from the input space into the output space

Output space: Yes or No



£ 35k

Output space: A predicted revenue

Machine Learning tasks consist in making a statistical model learn a mapping function from the inputs to the desired outputs from the data available

Formally, a model (e.g., Linear Regression) is a class of functions to consider for that task (e.g., all linear functions)

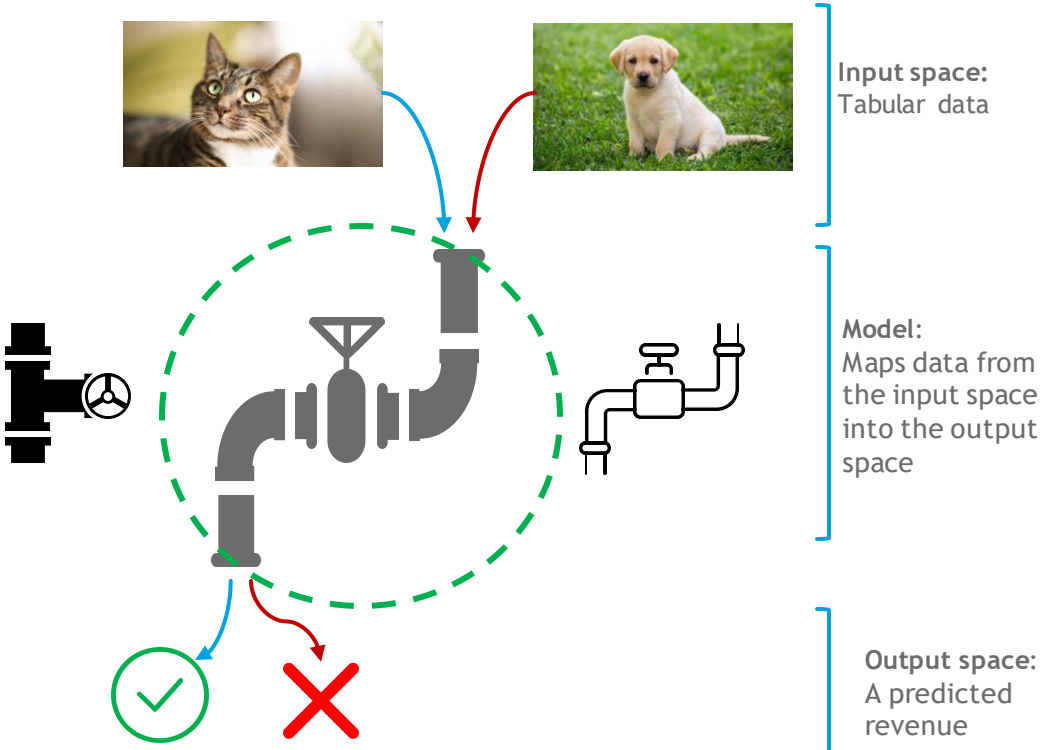
$$S = \{f_\theta, \theta \in \Theta\}$$

In other words, a group of mappings to choose from:

A machine learning model refers to both a group of functions to solve a task, and the chosen function within that group

Illustration of machine learning models

Is there a cat in this picture?



Training a model amounts to finding the optimal function with that class:

$$\text{Find best } s \in S$$

By extension, we call « model » both the ensemble of considered mappings (before training) and the one selected after the training process

- The goal of ML model training is to find the configuration of parameters' values of the model that most suit the observed data
- Once the model is "trained" it can then make proper inferences about a previously unknown observation (test set, real-life data)

ML approaches can be categorized as instance-based, model-based, and ensemble-based

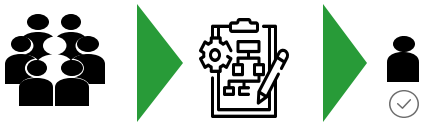
Instance-based learning

Instance-based models directly memorize (parts of) the training dataset, and use this memory to make predictions on new samples

Examples:
K-Nearest-Neighbors
Self-Organizing Map

Model-based learning

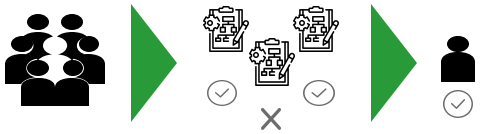
Model-based approaches deduce rules and parameters from the training dataset, without storing it. They apply these rules to make predictions on new samples



Examples:
Linear Regression
Decision Tree
Neural Network

Ensemble-based learning

Ensemble-based models use a collection of smaller models, each making its decisions. Individual decisions are aggregated to make the final prediction



Examples:
Random Forest
XGBoosting

ML approaches can be categorized as instance-based, model-based, and ensemble-based

Instance-based learning

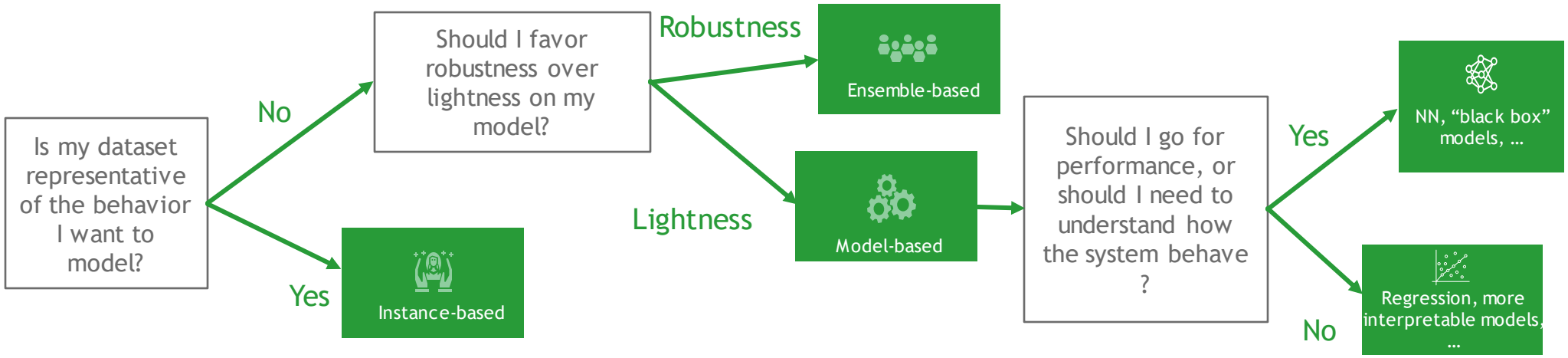
Instance-based models directly memorize (parts of) the training dataset, and use this memory to make predictions on new samples

Model-based learning

Model-based approaches deduce rules and parameters from the training dataset, without storing it. They apply these rules to make predictions on new samples

Ensemble-based learning

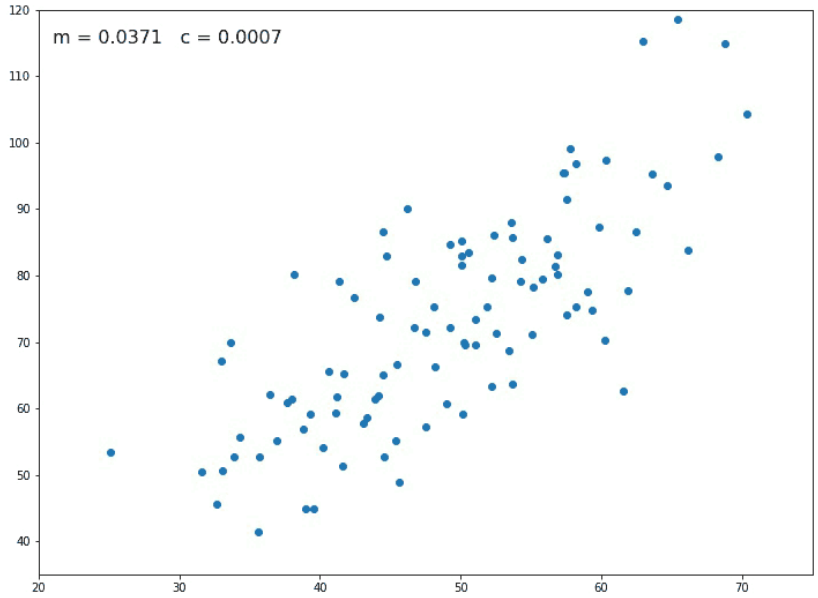
Ensemble-based models use a collection of smaller models, each making its decisions. Individual decisions are aggregated to make the final prediction



Training or fitting a model amounts to finding the parameters making the model closer to the dataset with regards to a metric

Illustration of machine learning models

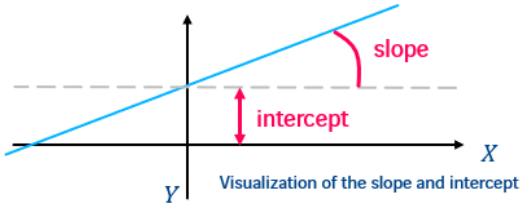
Example of linear regression



Linear regression example

Our objective is to find the line that "best represents" (fits) the available observations (the point cloud).

That line is defined by two parameters: the slope θ_1 and the intercept θ_2



Visualization of the slope and intercept

- We must find the intercept and slope (the parameters) that best fit the points
- We need a metric (the Loss) in order to tell apart the different parameters. We choose the distance between each point and the line.
- We find the best parameters by minimizing this distance.

Regression attempts to estimate the mapping function from the input variables to numerical or continuous output variables

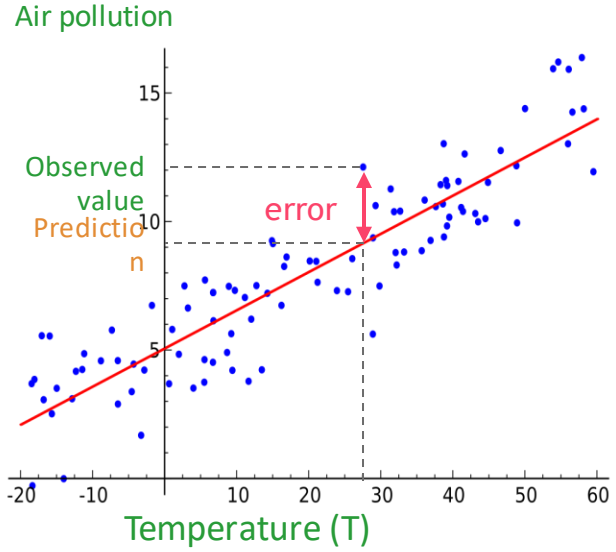
Description : Linear Regression

- Find the relation between input variables and continuous output variable
- Example : Understanding the relation between air pollution and temperature
- Class of model selected: linear regression

$$\hat{y} = f_\theta(x) = \theta_0 + \theta_1 x_1$$

- Parameter $\boldsymbol{\theta} = (\theta_0, \theta_1)$

Illustration



Metric

The error is the difference between model's prediction and the real value

$$error = (y - \hat{y})$$

A classical metric for regression is the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$$

- N is the number of observations
- Yi the observed output values
- $\hat{Y}_i$ represents the predicted values

The goal of regression algorithms is to minimize the RMSE

How can we make sure that the parameters found are also optimal for new samples ?