Introduction to Machine Learning and Data Science
S-INFO-256

1. Introduction

Pierre Vandenhove
Course material by Souhaib Ben Taieb

Université de Mons

UMONS
Université de Mons

Faculté
des Sciences

Outline

Teaching staff

Pierre Vandenhove (lectures + exercise sessions)

Theoretical Computer Science Lab
De Vinci Building, second floor, office 2.15
pierre.vandenhove@umons.ac.be

Course material from Souhaib Ben Taieb's course, with practical sessions by Victor Dheur and Tanguy Bosser.

S-INFO-256: Introduction to Machine Learning and Data Science

- **Prerequisites**
  - Probability and statistics
  - Linear algebra
  - Optimization
  - Python programming
- **Moodle**
  - `https://moodle.umons.ac.be/course/view.php?id=2785`
  - Lecture notes, announcements, project details, assignment submissions, etc.
- **To ask questions**: Moodle forum, after or before courses, by email (start your email subject with `[ML1]`).

Assessment

- Written exam ($\mathbf{E}$) (closed book) (/20)
- Project ($\mathbf{P}$) (/20)
- Final mark $= \begin{cases} \mathbf{E} \times 0.7 + \mathbf{P} \times 0.3 & \text{if } \mathbf{E} \geq 50\% \text{ and } \mathbf{P} \geq 50\%, \\ \min(\mathbf{E}, \mathbf{P}) & \text{otherwise.} \end{cases}$

Project

- Groups of **three** students.
- Project statement provided at the end of March.
- Analysis of a dataset using models discussed and not discussed in the course.
- Code, report, and short oral presentation.
- Deadline on **Friday, May 9, 2025**.
- Presentations the following week.

What is this course about?

- **This course is about**:
  - **A broad introduction to machine learning**: regression, classification, linear and nonlinear models, model assessment and selection, dimension reduction, etc.
  - **Preparation for learning**: machine learning is evolving fast; we want you to be able to understand the fundamentals and teach yourself the latest.

- **This course is not**:
  - A **survey/practical course**: list of machine learning algorithms, how to win prediction competitions, how to perform data analysis, how to use ChatGPT, etc.
  - An **easy course**: familiarity with intro probability, statistics and linear algebra are assumed. Start studying early.

References I

There are lots of freely available and high-quality machine learning resources.

- **An Introduction to Statistical Learning**, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. [Website link]
- **CS229 Lecture Notes**, Andrew Ng, Tengyu Ma. [Website link]
- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, Trevor Hastie, Robert Tibshirani, Jerome Friedman. [Website link]
- **Probabilistic Machine Learning: a book series**, Kevin Murphy. [Website link]

References II

- **Linear Algebra Review and Reference**, Zico Kolter and Chuong Do. [Website link]
- **All of Statistics**, Larry Wasserman. [Website link]
- **Numerical Optimization**, Nocedal, Wright. [Website link]
- **Linear Algebra**, David Cherney, Tom Denton, Rohit Thomas and Andrew Waldron. [Website link]

Why Python for machine learning?

- **Python** is a popular programming language for machine learning and data science.
- The main argument is the **established ecosystem** of libraries and tools for machine learning and data science (e.g., NumPy, pandas, scikit-learn, TensorFlow, PyTorch, Jupyter Notebooks, etc).
- Also, you are already proficient in it ☺
- It may evolve; Python was not always the leader for data science in the past (R, Matlab, C++...).
- The **strong library support** (Python and others) is one of the key reasons for the recent success of machine learning.
- Python is an easy-to-use scripting language.

Why Python for machine learning?

- **Pitfall**: not strongly typed and very permissive; sometimes hard to debug because too "friendly"!
- Example: this is code by a student to invert a matrix M (i.e., compute M^{-1}).

```
import numpy as np
M = np.array([[4, 4], [2, 2]])
print(1 / M)
```

- Returns `[[0.25, 0.25], [0.5, 0.5]]`, which is a component-wise $x \mapsto \frac{1}{x}$, not the matrix inverse... Should have used `np.linalg.inv(M)`.
- Even more important to understand the underlying math!

Outline

What is learning?

"The activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something."

(Merriam Webster dictionary)

What is machine learning? (1/2)

The term *machine learning* was coined by Arthur Samuel in 1959, who defined it as...

> *"The field of study that gives computers the ability to learn without being explicitly programmed."*
>
> *(Arthur Samuel, 1959)*

He wrote a computer program that was better than himself at playing checkers.

https://en.wikipedia.org/wiki/Checkers#/media/File:CheckersStandard.jpg

What is machine learning? (2/2)

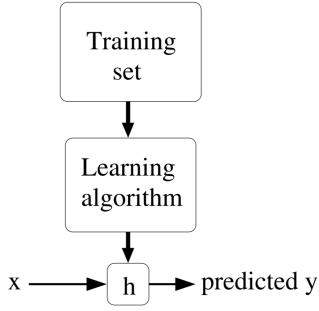
More modern definitions:

> *"The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data."*
>
> *(Oxford Languages)*

> *"A* **computer program** *is said to* **learn** *from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*
>
> *(Tom Mitchell)*

Why learn from data?

- **Better understand** (**inference**) or **make predictions** about a certain phenomenon under study.
- **Construct a model** of that phenomenon by finding relations between several variables.
- If phenomenon is complex or depends on a large number of variables, an **analytical solution** might not be available.
- However, we can **collect data** and learn a model that **approximates** the true underlying phenomenon.



Learning from data

- **The essence of machine learning**
  - A pattern exists.
  - We cannot pin it down mathematically.
  - We have data on it!
- **Learning examples**
  - Spam detection
  - **Product recommendation**
  - Credit card fraud detection
  - Medical diagnosis

Me: *Purchases a toilet seat on Amazon *

Amazon for the next 4 weeks:

Related fields and other views of "learning from data"

"Statistics is the **science of learning from data**, and of **measuring**, **controlling**, and **communicating uncertainty**; [...]"

"Data mining, [...], is the **computational process** of **discovering patterns** in **large data sets** involving methods at the intersection of **artificial intelligence**, **machine learning**, **statistics**, and **database systems**."

"Data science means the **scientific study** of the **creation**, **validation** and **transformation of data** to **create meaning**."

"Artificial intelligence is the theory and development of **computer systems** able to perform tasks normally requiring **human intelligence**, such as **visual perception**, **speech recognition**, **decision-making**, and **translation between languages**."

Machine learning/data science lifecycle

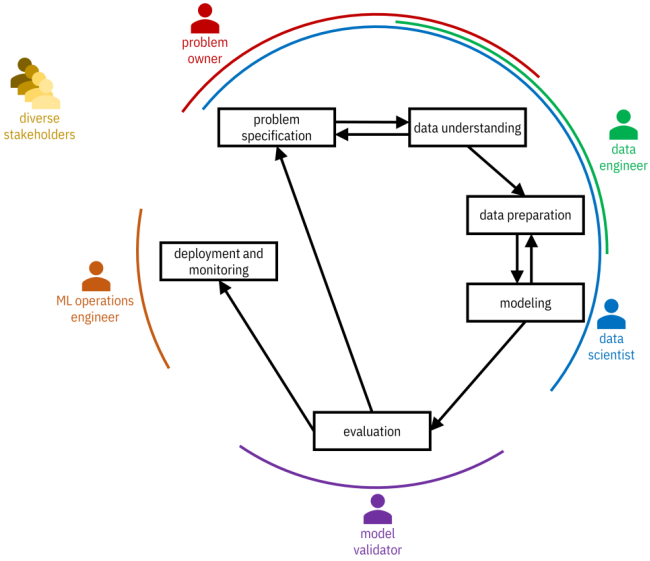


Image source: http://www.trustworthymachinelearning.com

Beyond model accuracy

> *"The full cycle of a machine learning project is not just modeling. It is finding the right data, deploying it, monitoring it, feeding data back [into the model], showing safety—doing all the things that need to be done [for a model] to be deployed. [That goes]* **beyond doing well on the test set**, *which fortunately or unfortunately is what we in machine learning are great at."*
>
> *(Andrew Ng)*

Other challenges:

- Data biases and privacy
- Model reliability (distribution shift, fairness, adversarial robustness)
- Model interpretability, explainability, and transparency

Especially important given the **pervasiveness and accessibility** (thanks to software, hardware, online resources...) of machine learning today (e.g., in healthcare, finance, automation, education, media, surveillance...).

For more details, see `http://www.trustworthymachinelearning.com`.

Outline

Machine learning problems?

Which of the following problems are **best suited** for machine learning?

1. Classifying numbers into primes and non-primes.
2. Detecting potential fraud in credit card charges.
3. Determining the time it would take a falling object to hit the ground.
4. Determining the optimal cycle for traffic lights in a busy intersection.
5. Calculating the maximum load a bridge can support based on its dimensions and the materials used in construction.

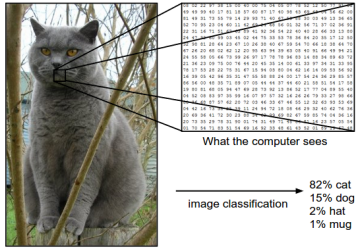
Supervised learning

We are given a training dataset consisting of **inputs** and corresponding **outputs (labels)**.
The goal of supervised learning is learning a function that maps these inputs to their outputs, based on the given input-output pairs.

Supervised learning tasks	Input	Output (label)
object recognition	image	object category
image captioning	image	caption
document classification	text	document category
speech-to-text	audio waveform	text
⋮	⋮	⋮

Input Vectors

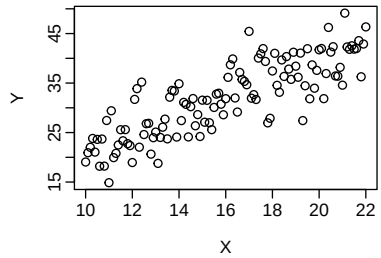
- Machine learning algorithms can be suited to **various types of data** (images, text, audio waveforms, graphs, time series, etc).
- We often **represent** the input as a vector in $\mathbb{R}^p$.
  - Vectors are a useful representation since we can do linear algebra.

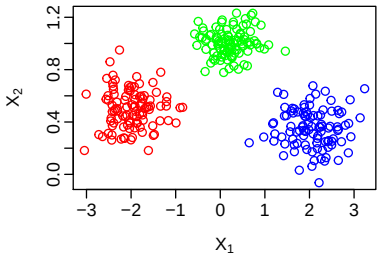


[Image Credit: Andrej Karpathy]

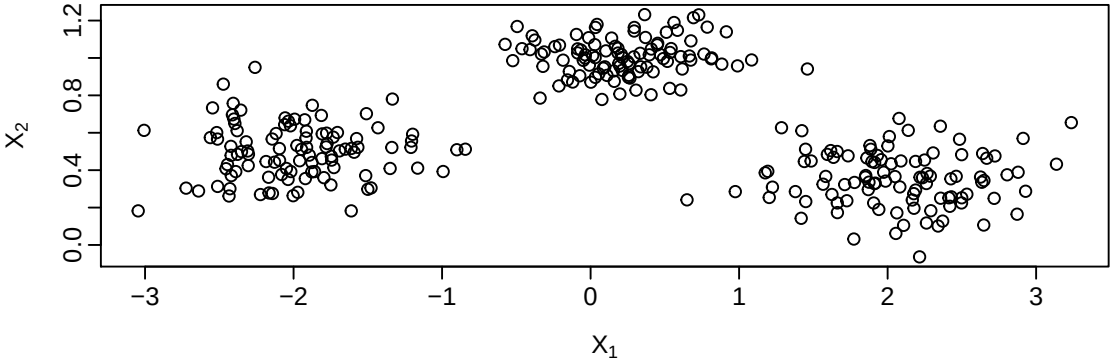
Supervised learning

- **Input**: $\boldsymbol{x} \in \mathcal{X}$ where $\mathcal{X}$ is the input space.
  - Example: $\mathcal{X} = \mathbb{R}^2$.
- **Output**: $y \in \mathcal{Y}$ where $\mathcal{Y}$ is the output space.
  - Regression: $\mathcal{Y} = \mathbb{R}$.
  - Classification (with K classes): $\mathcal{Y} = \{C_1, C_2, \ldots, C_K\}$.
  - The output can also be a structured object (e.g. image, text, etc).
- **Data**: $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.
- **Task**: predict the output y for new inputs $\boldsymbol{x}$.





Unsupervised learning



- **Input**: $X \in \mathcal{X}$ where $\mathcal{X}$ is the input space.
  - Example: $\mathcal{X} = \mathbb{R}^2$.
- **No explicit output to predict**.
- **Data**: $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} = \{\boldsymbol{x}_i\}_{i=1}^n$.

Examples of unsupervised learning

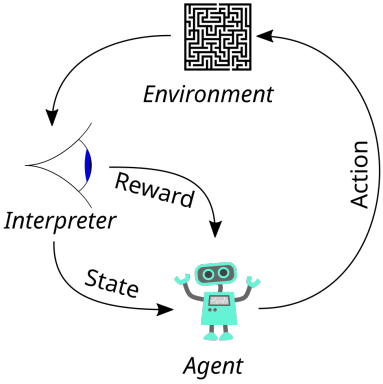
- **Clustering**, i.e., partitioning the data into groups of similar objects.
  - Grouping similar news articles ("You may also like...").
  - Identifying customer segments ("Users who bought this also bought...").
- **Feature extraction**, i.e., finding a compact representation of the data.
  - Reducing the dimensionality of data.
  - Edge detection in images.
- Other concrete problem: **Cocktail party problem**: given a set of audio recordings of people talking simultaneously, separate the voices of the different speakers.

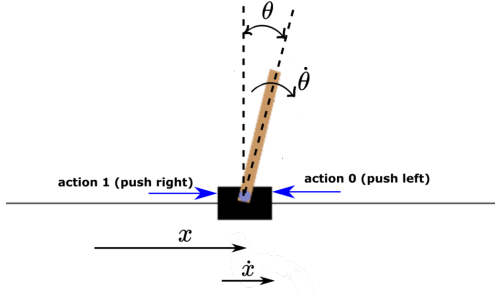
Different learning problems (1/3)

- **Supervised learning**
  - (input, output)
- **Unsupervised learning**
  - (input)
  - **Self-Supervised Learning**
    - Learning representations by predicting parts of the input.
    - Example: Learning to make sentences by predicting missing words.
- **Semi-supervised learning**
  - (input, output) for some observations, and only (input) for others.

Different learning problems (2/3)

- **Reinforcement learning**
  - An agent learns to make decisions (actions) in an environment (state) to maximize a reward.
  - Sequence of (state, action, reward).





See *Gymnasium* for more examples.

Different learning problems (3/3)

- **Transfer learning**
  - Leveraging a *pre-trained* model on a new, related task.
  - Example: Using a model trained on a large image dataset to perform a specific image recognition task with a much smaller dataset.
  - For exemple, from a neural network that can already distinguish dog breeds, retrain it to distinguish cat breeds.
- Other types of learning:
  - **Online learning**, where the data arrives sequentially and models must be updated on-the-fly (e.g., stock market prediction),
  - **Active learning**, where the algorithm can query the user for labels (e.g., in medical diagnosis),
  - ...

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved (supervised, unsupervised, semi-supervised, reinforcement, online, active),
2. identify the **training data** to be used (inputs, and outputs if relevant).

(If a task can fit more that one type, explain how.)

- Recommending a book to a user in an online bookstore
- Playing tic-tac-toe
- Categorizing movies into different types
- Optimizing delivery routes in real-time
- Predicting the next word in a sentence
- Identifying fraudulent transactions
- ChatGPT (GPT = *Generative Pre-trained Transformers*)