Machine Learning I

Introduction

Souhaib Ben Taieb

University of Mons

UMONS
Université de Mons

Faculté
des Sciences

Outline

About this course

Introduction to machine learning

Learning problems

Teaching staff

Souhaib BEN TAIEB (Instructor)
Big Data and Machine Learning Lab
De Vinci Building, second floor, room 2.15
souhaib.bentaieb@umons.ac.be

Victor DHEUR (TA)
PhD candidate
De Vinci Building, ground floor
victor.dheur@umons.ac.be

Tanguy BOSSER (TA)
PhD candidate
De Vinci Building, ground floor
tanguy.bossera@umons.ac.be

S-INFO-075: Machine Learning I

- Prerequisites
    - Probability and Statistics
        - S-PHYS-100: Probabilités
        - S-PHYS-101: Probabilités et statistique
    - Linear algebra
    - Optimization
    - Python programming
- **Course Webpage**
    - `https://github.com/bsouhaib/ML1-2024`
    - Lecture notes, project details, etc.
- **Moodle**
    - `https://moodle.umons.ac.be/course/view.php?id=2785`
    - Forum for asking questions, assignment submissions, etc.
- **No email please — use the Moodle forum**

Assessment

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

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 fast-moving; we want you to be be able to understand the fundamentals and teach yourself the latest.

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

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 fast-moving; we want you to be be able to understand the fundamentals and teach yourself the latest.

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

References I

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

- **An Introduction to Statistical Learning**. James, Witten, Hastie and Tibshirani. [Website link]
- **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Trevor Hastie, Robert Tibshirani, Jerome Friedman. [Website link]
- **Computer Age Statistical Inference: Algorithms, Evidence and Data Science**. Bradley Efron, Trevor Hastie. [Website link]
- **Understanding Machine Learning: From Theory to Algorithms**, Shai Shalev-Shwartz, Shai Ben-David. [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 Algeba**, David Cherney, Tom Denton, Rohit Thomas and Andrew Waldron. [Website link]

Outline

About this course

Introduction to machine learning

Learning problems

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?

"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)

Learning from data

- **Better understand** 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

$$\boxed{\text{Data} \longrightarrow \text{Learning model} \longrightarrow \text{Knowledge/Decision}}$$

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

Related fields and other views of “learning form data”

“Statistics is the **science of learning from** <u>**data**</u>, and of **measuring**, **controlling**, and **communicating** **uncertainty**; [...]”

“Data mining, [...], is the **computational process** of **discovering** **patterns** in **large** <u>**data**</u> **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** <u>**data**</u> 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

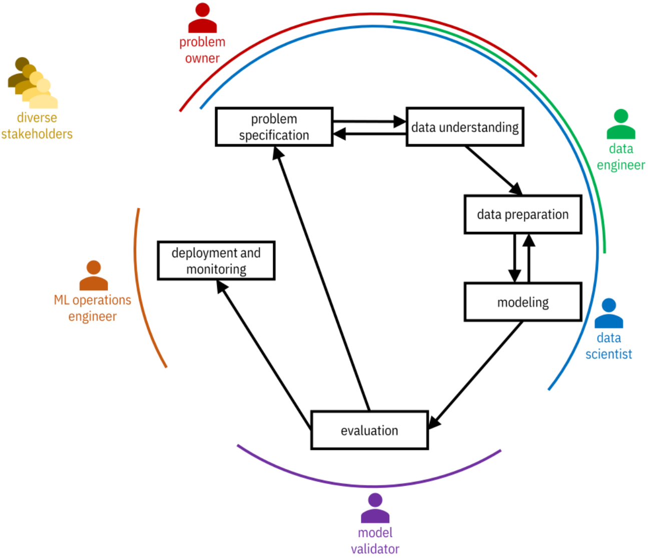


Image source: `http://www.trustworthymachinelearning.com`

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 and explainability
- Model transparency

For more details, see `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 and explainability
- Model transparency

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

Outline

About this course

Introduction to machine learning

Learning problems

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.

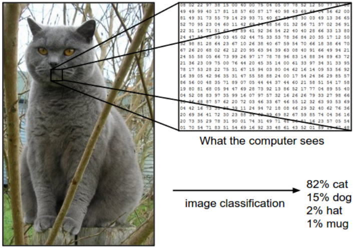
Supervised learning

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

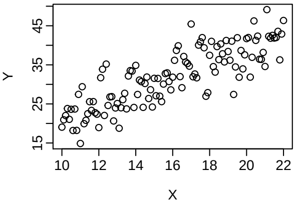
- Machine learning algorithms must be able to handle **various types of data** (images, text, audio waveforms, graphs, time series, etc)
- We often **represent** the input as a vector in $\mathbb{R}^p$
  - Vectors are a useful representation since we can do linear algebra.

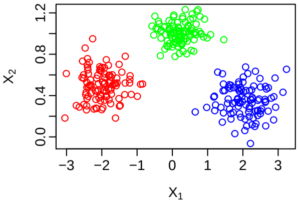


[Image Credit: Andrej Karpathy]

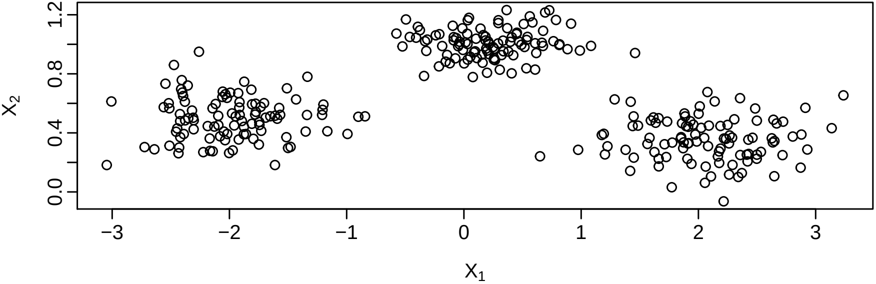
Supervised learning

- **Input**: $X \in \mathcal{X}$ 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 tasks**: clustering (partition data in groups), feature extraction (learn meaningful features automatically), etc

Different learning problems

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

Different learning problems

- **Reinforcement learning**
  - (input, *some* output, grade for this output)
  - (state, action, reward)
  - An agent learns to make decisions (actions) in an environment (state) to maximize a reward.
- **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.
- Other types of learning: **online learning**, **active learning**, etc.

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)

Different learning problems

For each of the following tasks,

1. identify which **type of learning** is involved
2. identify the **training data** to be used.

(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)