Machine Learning

Introduction

Souhaib Ben Taieb

February 22, 2021

University of Mons

Overview

About this course

Introduction to machine learning

About this course

S-INFO-256: Machine Learning

- **Instructor**
  - Prof. Souhaib BEN TAIEB
  - De Vinci Building, second floor, room 2.15
  - Email: souhaib.bentaieb@umons.ac.be
- **Github page**
  - `https://github.com/bsouhaib/ML21`
  - Lecture notes, project details, etc.
- **Moodle**
  - `https://moodle.umons.ac.be/course/view.php?id=2785`
  - Forum for asking questions, submissions, etc.
  - **No email please — use the Moodle forum**

Assessment

- Exam (E) (*open book*): **70%**
- Project (P) (group of 2 students): **30%**
- Final mark:
  - If $E \geq 45\%$ and $P \geq 45\%$: Final mark $= E \times 0.7 + P \times 0.3$
  - If $E < 45\%$ or $P < 45\%$: Final mark $= \min(E, P)$

Prerequisites

- Probability and statistics
- Multivariate calculus
- Linear algebra
- Optimization (non-linear)
- Computer programming: Python and/or R

About this course

- **What this course is**:
  - *Fundamentals of machine learning*: bias/variance tradeoff, overfitting, parametric and non-parametric models, regression, classification, model selection, dimensionality 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.
- **What 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, etc.

References

- *An Introduction to Statistical Learning*. James, Witten, Hastie and Tibshirani. [Link]
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Trevor Hastie, Robert Tibshirani, Jerome Friedman. [Link]
- *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Bradley Efron, Trevor Hastie. [Link]
- *Understanding Machine Learning: From Theory to Algorithms*, Shai Shalev-Shwartz, Shai Ben-David. [Link]
- *Machine Learning: A Probabilistic Perspective*, Kevin Murphy. [Link]

Other references

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

Introduction to machine learning

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

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

Learning from data

"Machine learning is a **scientific discipline** that explores the **construction and study of algorithms** that can **learn 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 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 problems?

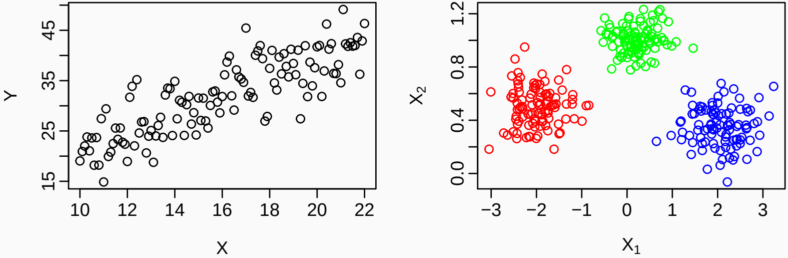
Which of the following problems are best suited for Machine Learning?

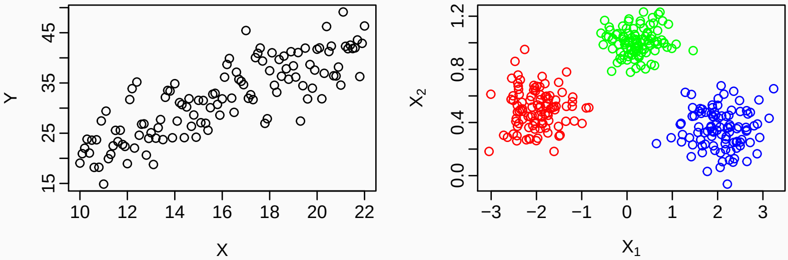
1. Classifying numbers into primes and non-primes.
2. Detecting potential fraud in credit card charges.
3. Determining the time it would take a falling object to hit the ground.
4. Determining the optimal cycle for traffic lights in a busy intersection.

Supervised learning

- Input: $x \in \mathcal{X}$ where $\mathcal{X}$ is the input space
  - Example: $\mathcal{X} = \mathbb{R}^p$
- Output: $y \in \mathcal{Y}$ where $\mathcal{Y}$ is the output space
  - Regression: $\mathcal{Y} \subseteq \mathbb{R}$.
  - Classification (with K classes): $\mathcal{Y} = \{C_1, C_2, \ldots, C_K\}$.
- Data: $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} = \{(x_i, y_i)\}_{i=1}^n$
- Task to solve: predict the output y for new inputs x

Supervised learning

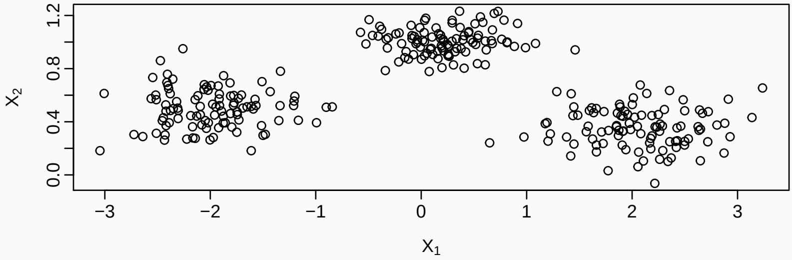




- Left figure: $\mathcal{X} = \mathbb{R}$ (one-dimensional input) and $\mathcal{Y} \subseteq \mathbb{R}$
- Right figure: $\mathcal{X} = \mathbb{R}^2$ (two-dimensional input) and $\mathcal{Y} = \{RED, GREEN, BLUE\}$

Unsupervised learning

- No explicit output to predict
- Data: $\mathcal{D} = \{x_1, x_2, \ldots, x_n\} = \{x_i\}_{i=1}^n$
- Task to solve: clustering (partition data in groups), feature extraction (learn meaningful features automatically), dimensionality reduction (learn a lower-dimensional representation of input), etc.



$\mathcal{X} = \mathbb{R}^2$ (two-dimensional input)

Different learning problems

- Supervised learning
  - (input, output)
- Unsupervised learning
  - (input)
- Semi-supervised learning
  - (input, output) for some observations, and only (input) for others.
- Reinforcement learning
  - (input, *some* output, grade for this output)
  - (state, action, reward)
- Other types of learning: online learning, active learning, etc.

In practice, it is important to identify which learning problem is best suited for the application and the data available.

Different learning problems

For each of the following tasks, identify which type of learning is involved (supervised, unsupervised or reinforced) and the training data to be used. If a task can fit more that one type, explain how and describe the training data for each type.

- Recommending a book to a user in an online bookstore
- Playing tic-tac-toe
- Categorizing movies into different types
- Learning to play music

Probability and statistics review

See board.