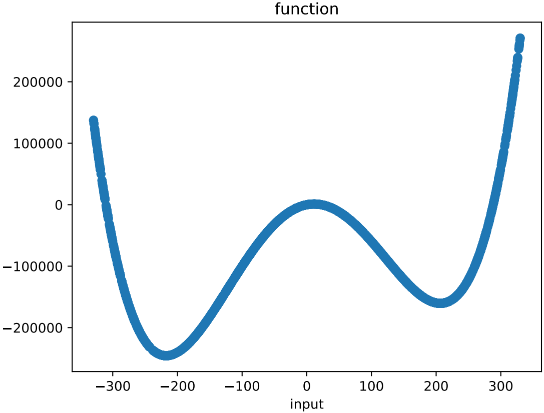
Machine learning I, supervised learning: problem statement



Machine learning (ML)

- (proposed definition of) **learning** : "Modification of a behavior, based on a life experiment"
- a **machine learning** system is programmed to learn in a **semi-automatic** way.

Classical programming vs ML

- Classical program : predict the total amount of money spent, based on the number of fruits bought and the price of each individual fruit (a summation is enough)
- ML program : predict the probability that a person buys some given fruit in a month, based on a database of customers and on some information about this person (e.g. buying log).

Use cases of ML

ML is useful for problems :

- that we cannot solve direcly thanks to an explicit representation (such as the amount of money spent in the previous first example).
- that we can solve in practice, but without a complete understanding (face recognition)
- that we could solve explicitely, but the computationnal ressources would be too heavy (molecular dynamics)

Denomination

- The name "machine learning" is rather deceitful !
- It is no more machine driven than any algorithm or coffee machine.
- The denomination "statistical learning" is used in some contexts.

Ingredients of machine learning

ML's main ingredients are

- optimization
- statistics and probabilities

Other tools come from :

- graph theory
- information theory
- statistical physics

DL / ML / AI :

- ML is a subset of AI.
- Deep learning is a subset of ML.

Learning paradigms

- **supervised learning** : learn to predict an output as a function of an input (predict the energy production of a wind farm based on sensors)
- **unsupervised learning** : learn information about the structure of data (density estimation, clustering, dimensionality reduction)
- **reinforcement learning** : learn to perform actions in order to maximize a reward (game player, alphago)

Why is there a hype around machine learning ?

Machine learning has received attention and funding because it has reached state-of-the-art efficiency on several problems, such as :

- computer vision
- spam classification
- machine translation
- speech recognition
- self-driving cars

Deep learning is involved in several of these setups.

ML revolution

- Technical progress in the computing and storage capacities
- Increase in amount of available data. According to IBM, 10^{18} bytes are created each day.
- Progress in algorithmic methods to analyze the data.

Example 1 : ImageNet

- A database of images (more than 15M), hand-annotated in order to indicate what objects are present in the image. More than 20000 categories of images.
- Contest : ImageNet Large Scale Visual Recognition Challenge.
- The best top 5 score (a measure of the classification error) went from 25% in 2011 to $\simeq$ 15.3% in 2012.

Example 1 : ImageNet

- A database of images (more than 15M), hand-annotated in order to indicate what objects are present in the image. More than 20000 categories of images.
- Contest : ImageNet Large Scale Visual Recognition Challenge.
- The best top 5 score (a measure of the classification error) went from 25% in 2011 to $\simeq$ 15.3% in 2012.
- The technology used was deep learning, exploiting GPUs (AlexNet).

Example 2 : AlphaGo

- In 2015 : beats a professionnal player. In 2017 : beats the world champion.
- Uses several technologies : among them **Deep reinforcement learning.**
- Improvements : AlphaGo Zero, trained without a database of played games. In 20217, AlphaZero beats AlphaGo Zero after 3 days of learning.

Example 3 : AlphaFold

- Goal : to predict the spatial configuration of proteins, from their DNA sequence.
- Achieves a breakthgough performance on the CASP challenge :
  - 2018 : more than 50% GDT (Global distance test), whereas it was $\leq$ 40% before then.
  - 2020 : $92,4\%$ GDT. At a $\geq$ 90% score, the method is considered competitive with experimental methods..
- `https://alphafold.ebi.ac.uk/`
- Also based on Deep learning.

Supervised Learning : formalization

- For a certain input x, you want to **predict** an output y : for instance,
  - x : contains the age, and the height of a person, so here x is a **vector** containing **two features**.
  - y : best record on a 100 meters track
- To do so, you learn from a number of **labeled examples** (x_i, y_i)
- In the case where what you want to predict is a **class**, it is a **classification problem**
- In the case where what you want to predict is a general function $y = f(x)$, it is a **regression problem** (example : 100 meters track time)

Supervised learning

- To do so, you learn from a number of **labeled examples** (x_i, y_i)
- In the case where what you want to predict is a **class**, it is a **classification problem** : $y \in \mathbb{N}$. (example : MNIST)
- In the case where what you want to predict is a general function $y = f(x)$, it is a **regression problem** : $y \in \mathbb{R}$.
- **Objective :** find a good estimation $\tilde{f}$, of f.

Important question

- **Objective :** find a good estimation $\tilde{f}$, of f.
- We have to define what it means that a function is a good estimation of another function.
- In order to measure the quality of $\tilde{f}$, we use **loss functions**.

Example loss function for a regression problem

- The loss function should be a measure of the discrepancy between our prediction and the correct label.
- For an individual sample, **a** discrepancy is the least-square loss

$$(f(x_i) - y_i)^2 \tag{1}$$

Loss function

- Taking into account the whole dataset, the **loss function** writes :

$$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{2}$$

- Several other loss functions are possible :

$$\sum_{i=1}^{n}|f(x_i) - y_i| \tag{3}$$

Loss function

- The loss function is a **real number** measuring the relevance of a **collection** of parameters ($\tilde{f}$ is defined by these parameters.)
- The number of parameters depends on the situation, and varies between 1 (e.g. for a simple linear model) and millions (e.g. for some deep neural networks).

Important question II

- To what subset of functions does $\tilde{f}$ belong ?

More supervised learning examples : I

Predict the winning team of an NBA game at half-time.

- Dataset : 15 years of games (comments, text) : approximately 17000 games.
- The dataset is preprocessed to have as an input a time-series : each time contains the score **and** 10 technical features (rebounds, etc.). So for each time the dimension is 11. Each game is a matrix of size 1440×11, reorganized as a line vector.
- Output : Receiving team wins or looses (classification)
- Evaluation metric : classification error ("0-1" loss).

Example II

Predict the quantity of oil in a rock.

- Input : tomographic image of a rock.
- Output : material of the rock, average presence of residual oil in the rock (regression).

Example III

Detect issues in wind farms.

- Input : sensors on the wind turbine (wind direction, air temperature, electric tension, rotation speed, component temperature, etc.) as a time series. Each step represents 10 minutes (several years).
- Output : Power generated by the turbine (regression)
- Evaluation metric : MAE (mean absolute error).

Difficulties of machine learning

- hard optimization problem
- overfitting / statistical garantees
- curse of dimensionality

Optimization problem

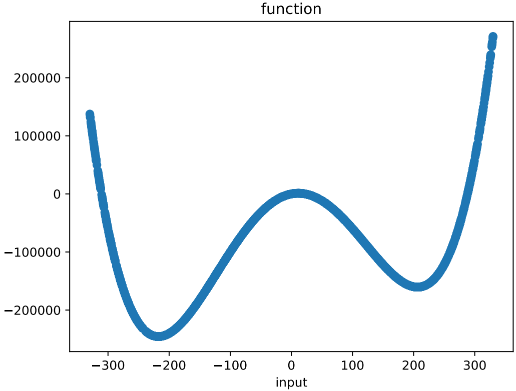
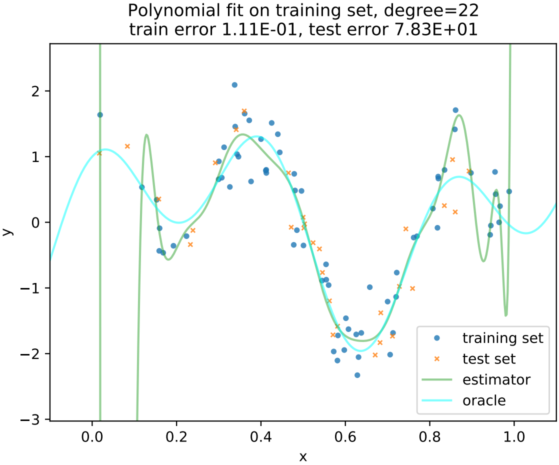


Figure – Loss function

Overfitting



Curse of dimensionality

Two numbers are important in machine learning :

- n : number of samples
- d : dimension (number of features) of a unique sample

Both can be large and prohibitive for some algorithms.

Curse of dimensionality

- n is large when the dataset has many samples.
- d is large if each sample has many features :
  - image
  - DNA sequence
  - text
  - audio/video file