Introduction to machine learning & Chemistry, or how I learned to be trendy

Clabaut Paul, Jiang Tao, Staub Ruben

ENS de Lyon

November 29,2019

Overview

1. Theory & Generalities
   - General principle of machine learning
   - Classification of machine learning
2. Application
   - Usages in chemistry
   - Limitations

Theory & Generalities

LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?

“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 M. Mitchell

LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?

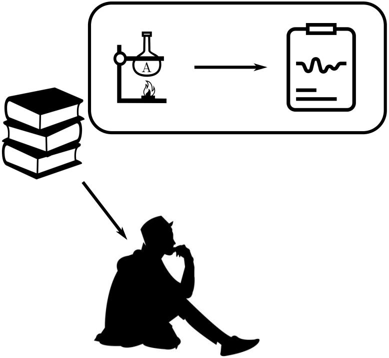
“Machine learning is glorified statistics”

“Machine learning is for Computer Science majors who couldn’t pass a Statistics course.”

“Machine learning is Statistics minus any checking of models and assumptions.”

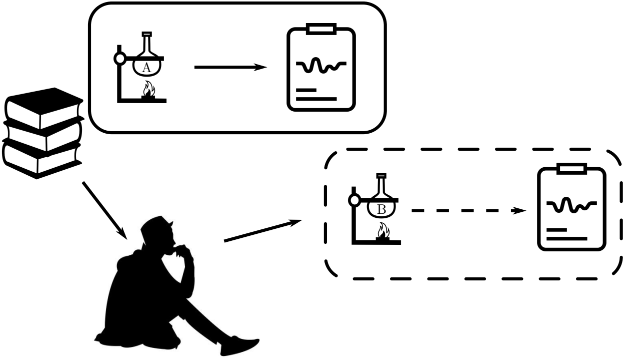
LABORATOIRE
DE CHIMIE
ENS DE LYON

What is machine learning ?



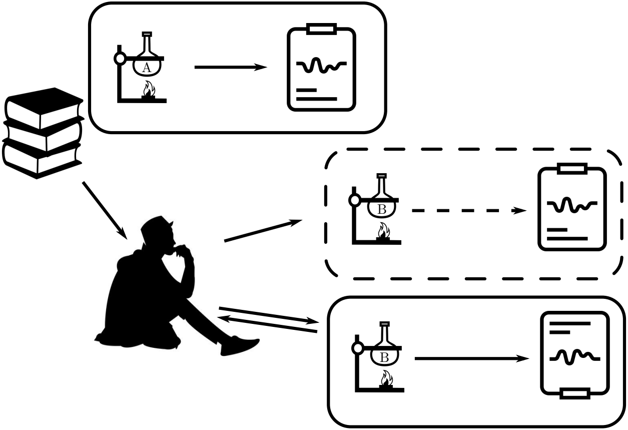
LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?



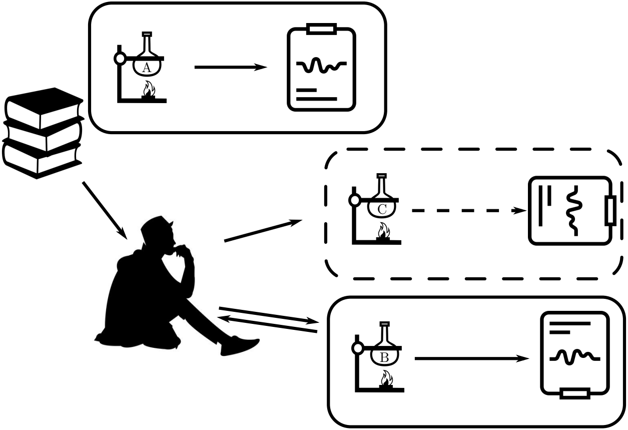
LABORATOIRE
DE CHIMIE
ENS DE LYON

What is machine learning ?



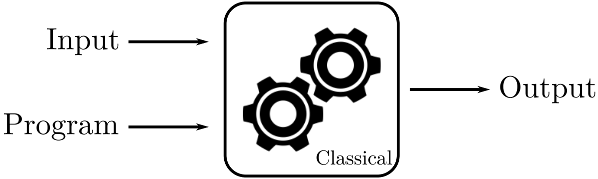
LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?



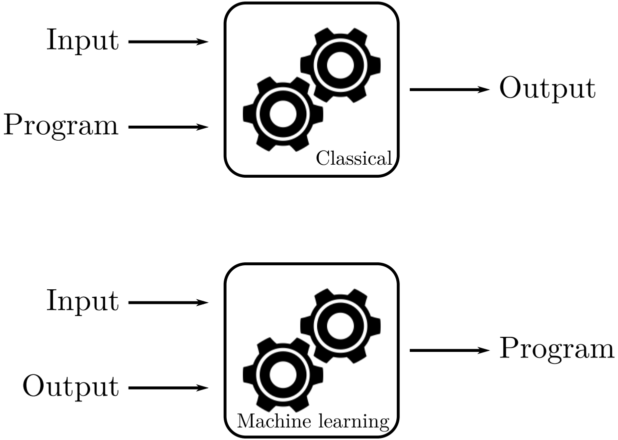
LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?



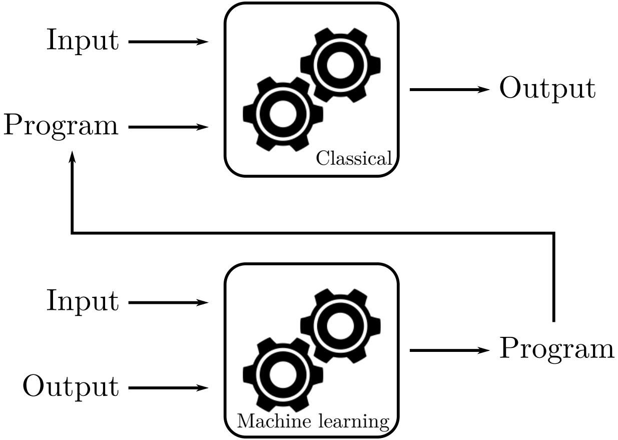
LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?



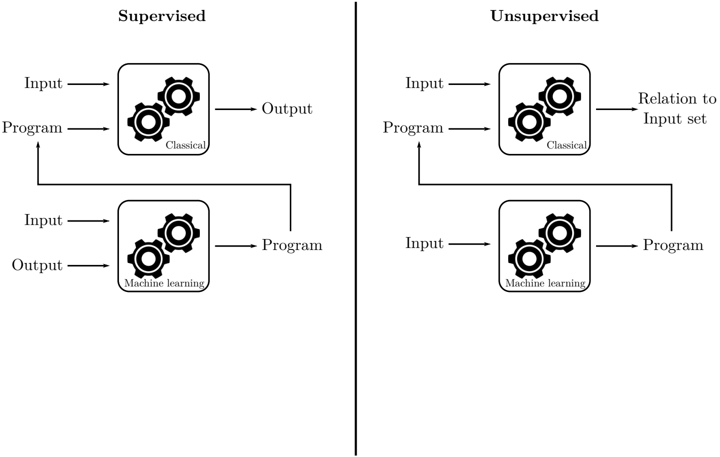
LABORATOIRE DE CHIMIE ENS DE LYON

What is machine learning ?



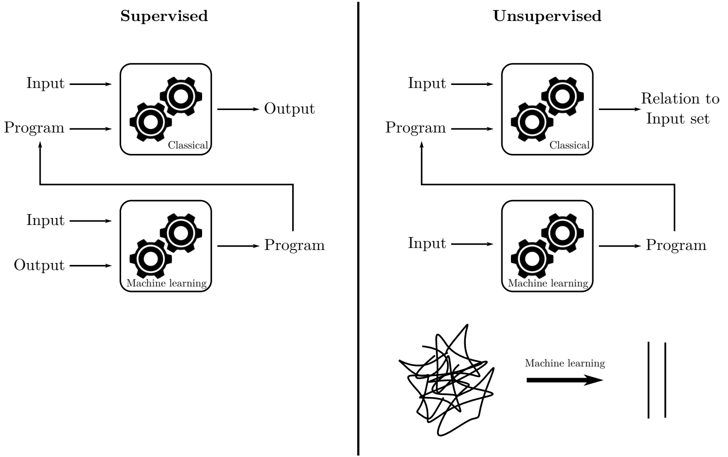
LABORATOIRE
DE CHIMIE
ENS DE LYON

Supervised VS. Unsupervised



LABORATOIRE
DE CHIMIE
ENS DE LYON

Supervised VS. Unsupervised



LABORATOIRE DE CHIMIE ENS DE LYON

Families of Supervised learning

Aim: finding a function (program) mapping features (input) to a target (output)

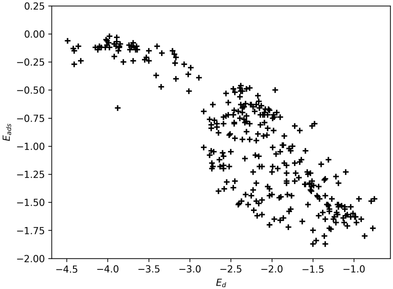
Linguistic distinction based on target type:

- Continuous target ⇒ **Regression** (e.g. E_{ads})
- Categorical target ⇒ **Classification** (e.g. active or not, cat or dog?)

Choose type of mapping function

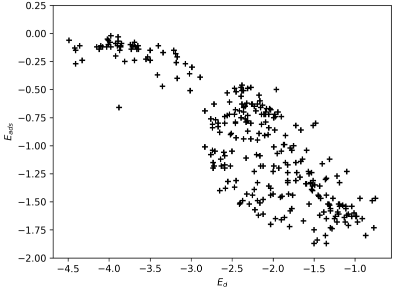
LABORATOIRE DE CHIMIE ENS DE LYON

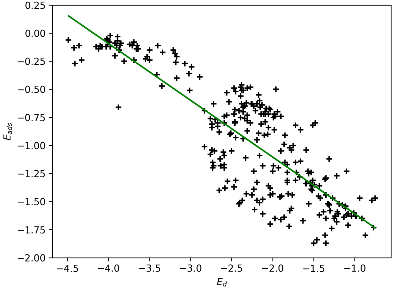
We all have done it a little before



LABORATOIRE DE CHIMIE ENS DE LYON

We all have done it a little before





LABORATOIRE DE CHIMIE ENS DE LYON

Linear model (regression)

Input: m features, 1 target

$$
\begin{array}{cccc}
 & \text{Features} & & \text{Target} \\
 & (x_1\ x_2\ \cdots\ x_m) & \rightarrow & y \\
\hline
 & (x_{11}\ x_{12}\ \cdots\ x_{1m}) & \rightarrow & y_1 \\
\text{Training set} & (x_{21}\ x_{22}\ \cdots\ x_{2m}) & \rightarrow & y_2 \\
 & \vdots & & \\
 & (x_{n1}\ x_{n2}\ \cdots\ x_{nm}) & \rightarrow & y_n
\end{array}
$$

Model: m parameters $w_1, w_2, \ldots, w_m$

$$\sum_i^m w_i x_i \approx y$$

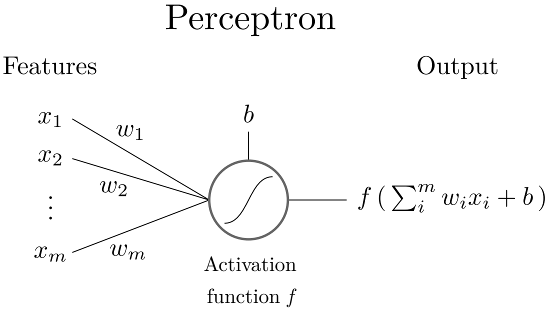
Pros: easily to understand

Cons: restricted to linear relations

LABORATOIRE DE CHIMIE ENS DE LYON

Neural network (regression)

Simplest neural network



Parameters: $w_1, w_2, \ldots, w_m, b$

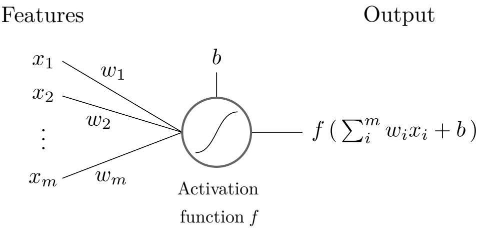
Equivalent to linear model if we use $f : x \mapsto x$

LABORATOIRE DE CHIMIE ENS DE LYON

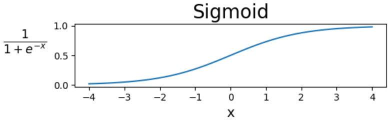
Neural network (regression)

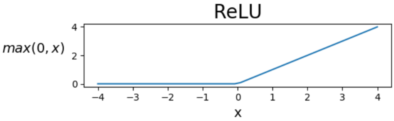
Simplest neural network

Perceptron



Common activation functions:

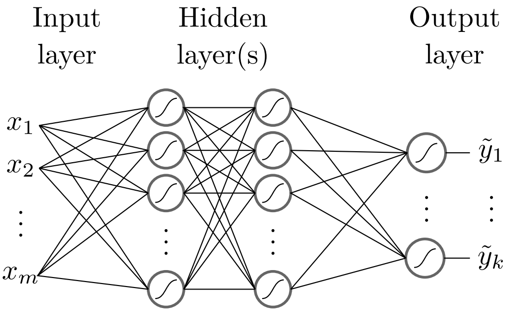




LABORATOIRE DE CHIMIE ENS DE LYON

Neural network (regression)

Input: *m* features, *k* targets



Parameters: weights + bias of each neuron

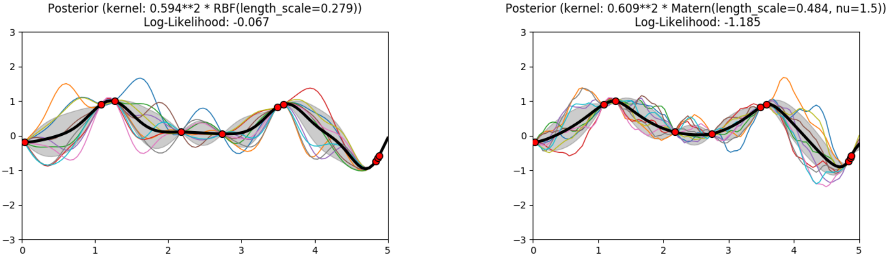
Pros: fast to train, fast to predict, good fitting properties, trendy

Cons: black box, design tuning

LABORATOIRE DE CHIMIE ENS DE LYON

Gaussian process regression (regression)

Assume fitted target is a Gaussian process with given smoothness, use Bayesian inference to estimate probability distribution for prediction



Intuition: consider all possible fitted targets and extract probability distribution

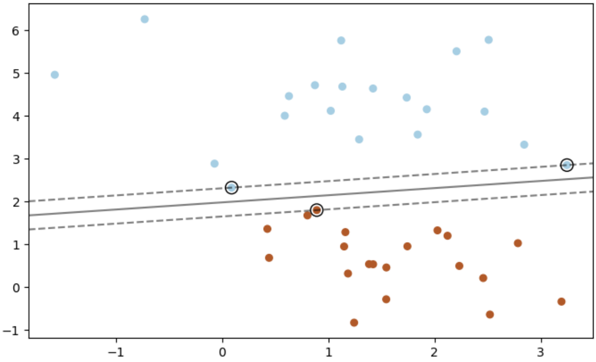
Pros: native prediction confidence estimator

Cons: kernel-dependant, computationally costly training

LABORATOIRE DE CHIMIE ENS DE LYON

Support Vector Machine (classification)

Input: m features, 1 categorical target

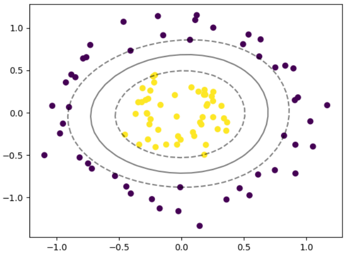


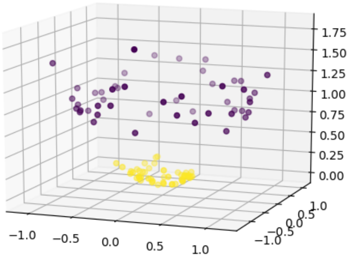
Intuition: Find best splitting hyperplane

LABORATOIRE
DE CHIMIE
ENS DE LYON

Support Vector Machine (classification)

Change the metric through a non-linear kernel → embedding into higher-dimensional space where separation could be possible.





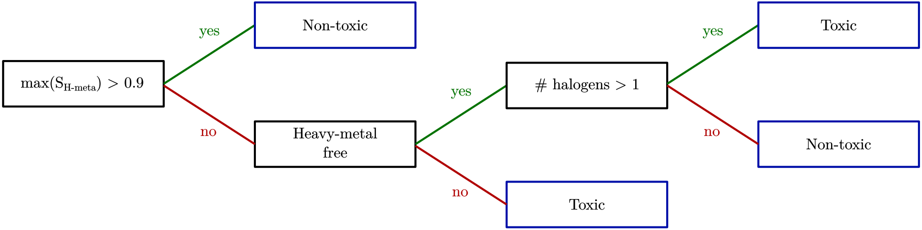
Pros: adapted for high-dimensionality

Cons: kernel-dependant, computationally costly training

LABORATOIRE DE CHIMIE ENS DE LYON

Decision tree (classification)

Input: *m* features, 1 categorical/continuous target

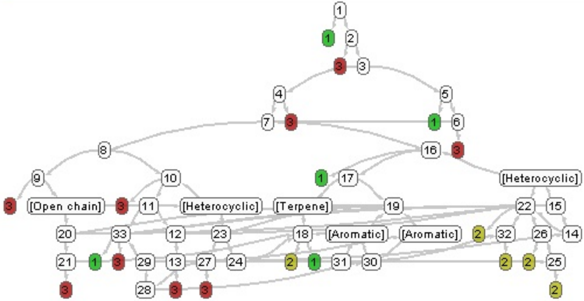


Intuition: Find best (feature+threshold) splitting on each node
Pros: easy interpretation (relevant features, ...), very fast prediction
Cons: requires balanced classes, optimal solution is NP-hard

LABORATOIRE DE CHIMIE ENS DE LYON

Decision tree (classification)

Input: m features, 1 categorical/continuous target



Intuition: Find best (feature+threshold) splitting on each node

Pros: easy interpretation (relevant features, ...), very fast prediction

Cons: requires balanced classes, optimal solution is NP-hard

LABORATOIRE DE CHIMIE ENS DE LYON

Families of Unsupervised learning

Aim: Without consulting the target/output (sometimes absent), finding the structure in the input data itself.

Useful in certain cases:

* No prior knowledge of how many/what classes is the data divided into.

* Find out the most important features of the input data before feeding it to a machine.

Typical methods:

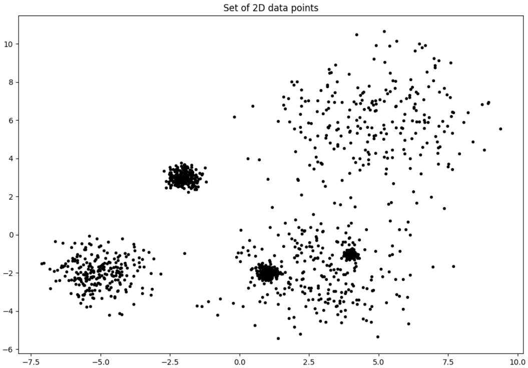
Clustering: k-means clustering, hierarchical clustering

Dimensionality reduction: PCA, Autoencoder

LABORATOIRE DE CHIMIE ENS DE LYON

Clustering illustration

Clustering is partitioning into groups of close points
Non trivial task: How many clusters do you identify?

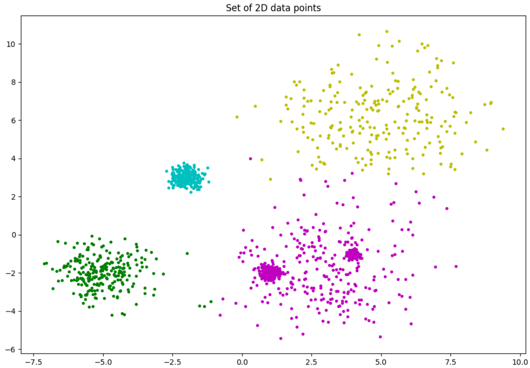


LABORATOIRE DE CHIMIE ENS DE LYON

Clustering illustration

Clustering is partitioning into groups of close points

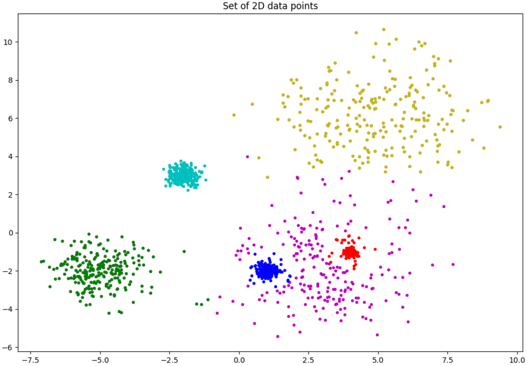
Non trivial task: How many clusters do you identify?



LABORATOIRE DE CHIMIE ENS DE LYON

Clustering illustration

Clustering is partitioning into groups of close points
Non trivial task: How many clusters do you identify?

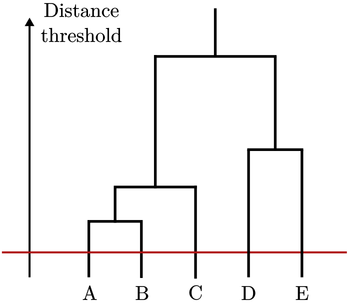


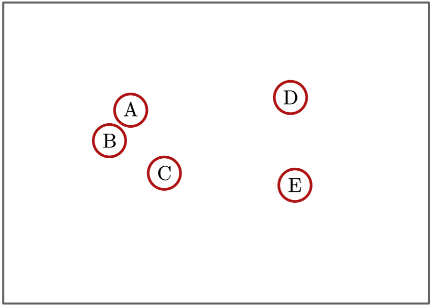
LABORATOIRE DE CHIMIE ENS DE LYON

Hierarchical clustering (clustering)

Input: metric between clusters, data points

Explore clusters generated for every threshold, produce dendrogram





Pros: easy to interpret, good overview

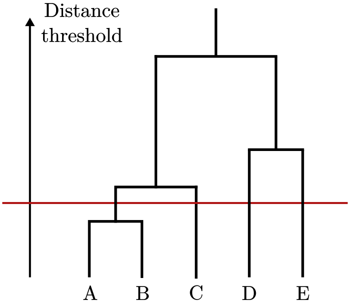
Cons: need user-defined metric between clusters, threshold selection

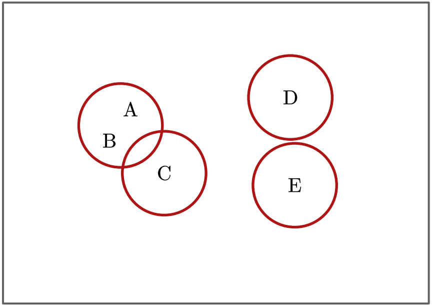
LABORATOIRE DE CHIMIE ENS DE LYON

Hierarchical clustering (clustering)

Input: metric between clusters, data points

Explore clusters generated for every threshold, produce dendrogram





Pros: easy to interpret, good overview

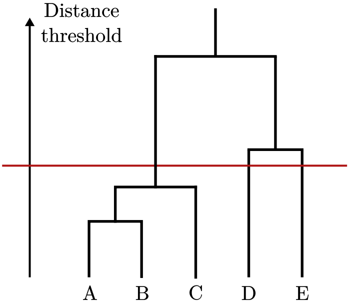
Cons: need user-defined metric between clusters, threshold selection

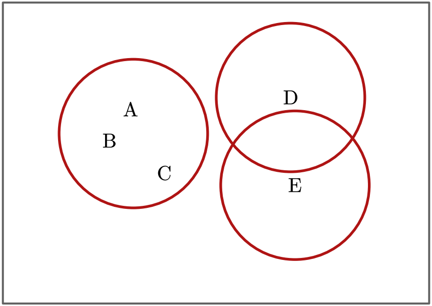
LABORATOIRE DE CHIMIE ENS DE LYON

Hierarchical clustering (clustering)

Input: metric between clusters, data points

Explore clusters generated for every threshold, produce dendrogram





Pros: easy to interpret, good overview

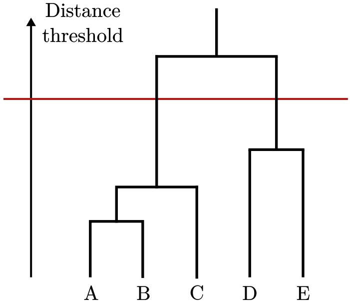
Cons: need user-defined metric between clusters, threshold selection

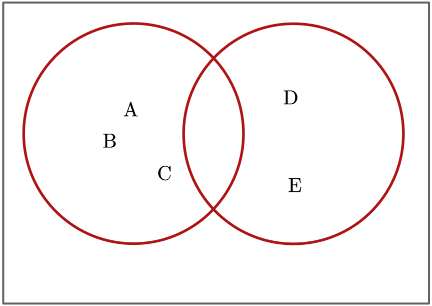
LABORATOIRE DE CHIMIE ENS DE LYON

Hierarchical clustering (clustering)

Input: metric between clusters, data points

Explore clusters generated for every threshold, produce dendrogram





Pros: easy to interpret, good overview

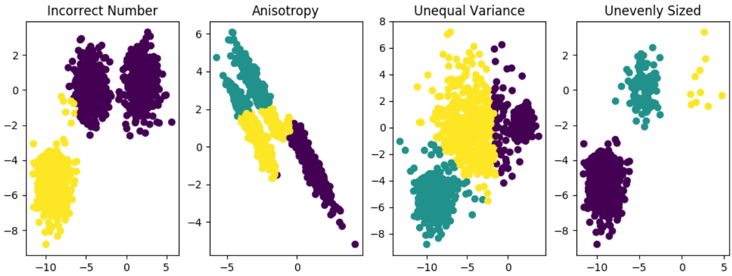
Cons: need user-defined metric between clusters, threshold selection

LABORATOIRE DE CHIMIE ENS DE LYON

K-means clustering (clustering)

Input: number of clusters, distance between data points

Find centroids that minimize the within-cluster sum-of-squares

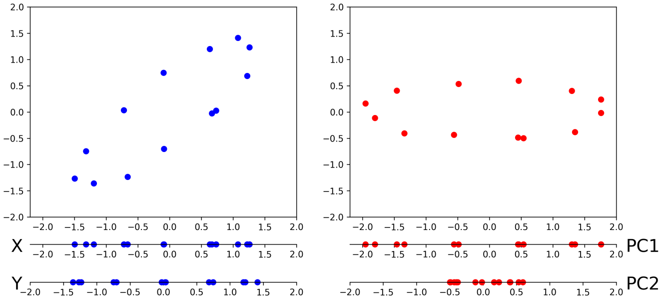


Pros: easy to understand

Cons: stochastic, assume convex and isotropic, poor high-dimensionality support

LABORATOIRE DE CHIMIE ENS DE LYON

PCA (dimensionality reduction)



Emphasize variation and bring out strong patterns in a dataset.

Useful for finding important features in high dimensional dataset.

LABORATOIRE DE CHIMIE ENS DE LYON

PCA (Wine chemistry)

Feature No.	Composition
1	Alcohol
2	Malic acid
3	Ash
4	Alcalinity of ash
5	Magnesium
6	Total phenols
7	Flavanoids
8	Nonavanoid phenols
9	Proanthocyanins
10	Color intensity
11	Hue
12	OD280/OD315
13	Proline

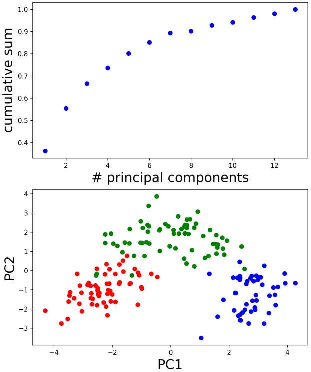
Class	Number of wines
1	59
2	71
3	48

https://archive.ics.uci.edu/ml/datasets/Wine, Accessed: 2019-11-29

LABORATOIRE DE CHIMIE ENS DE LYON

PCA (Wine chemistry)

Feature No.	Composition
1	Alcohol
2	Malic acid
3	Ash
4	Alcalinity of ash
5	Magnesium
6	Total phenols
7	Flavanoids
8	Nonavanoid phenols
9	Proanthocyanins
10	Color intensity
11	Hue
12	OD280/OD315
13	Proline



LABORATOIRE DE CHIMIE ENS DE LYON

Manifold (dimensionality reduction)

Non-linear dimensionality reduction:

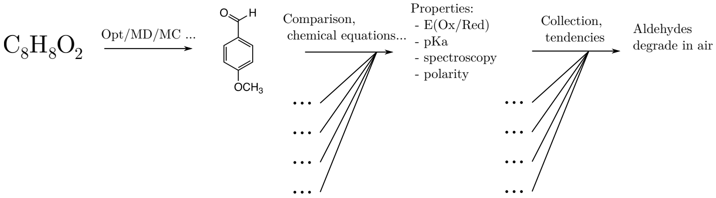
- Kernel PCA (use kernel instead of covariance)
- MDS (preserve distances)
- Isomap (preserve geodesic graph-based distances)
- Self-organizing maps (preserve topology)
- ...

Application

LABORATOIRE DE CHIMIE ENS DE LYON

Categories of applications in chemistry

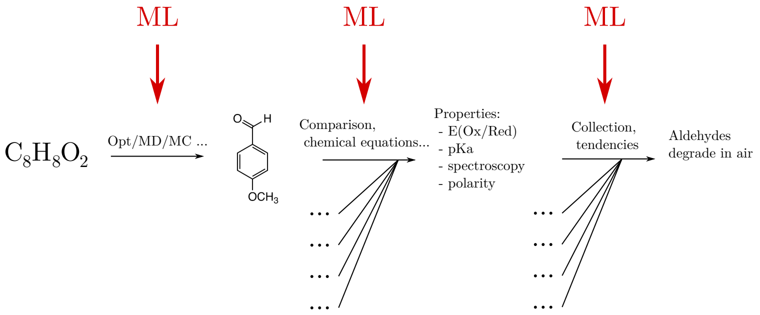
Substitute to computation algorithm, Direct property prediction, Data analysis



LABORATOIRE DE CHIMIE ENS DE LYON

Categories of applications in chemistry

Substitute to computation algorithm, Direct property prediction, Data analysis



LABORATOIRE DE CHIMIE ENS DE LYON

Application: Water/Platinum potential: Goal

DFT

500 ps of 20$\AA$ thick water layer on 3*3*4 Pt (111) slab
$\rightarrow$ 8 yrs on 100 processors

MM

$$E = E_{slab} + \sum_{wat}(E_{wat} + E_{slab/wat}) + \sum_{wat}\sum_{wat}(E_{wat/wat} + E_{slab/wat/wat}) + \sum_{wat}\sum_{wat}\sum_{wat}(\ldots) + \ldots$$

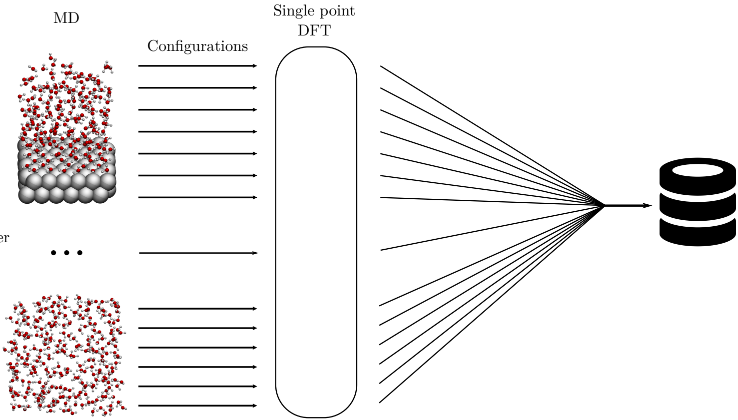
Neural Network

$$E = \sum_{atoms} E_{NN}(environement) \rightarrow \text{Might just work !}$$

LABORATOIRE
DE CHIMIE
ENS DE LYON

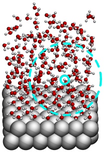
Step 1 Gathering data

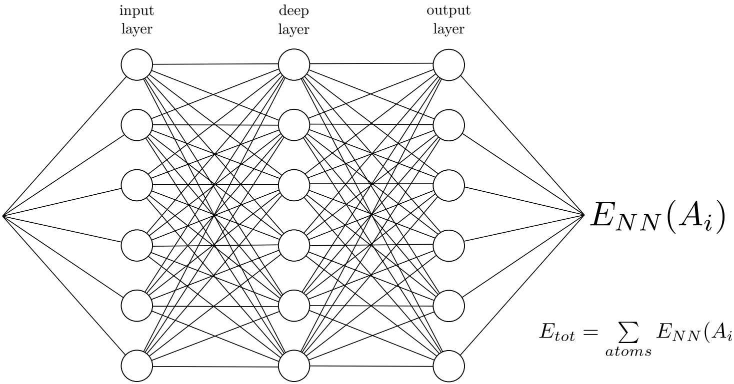
- large/small slabs
- hot/cold dynamics
- thick/thin water layer
- with/without void
- with/without slab
...



LABORATOIRE DE CHIMIE ENS DE LYON

Step 2 Input and Training the Neural Network

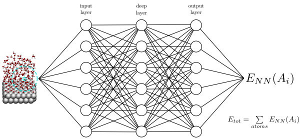


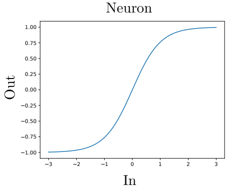


$$E_{tot} = \sum_{atoms} E_{NN}(A_i)$$

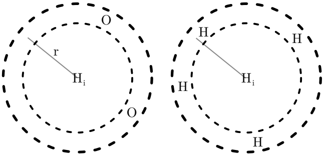
LABORATOIRE DE CHIMIE ENS DE LYON

Step 2 Input and Training the Neural Network





input layer <=>
ensemble of symetry function



$$F_{radial}(H_i, r, O) = 2$$

$$F_{radial}(H_i, r, H) = 4$$

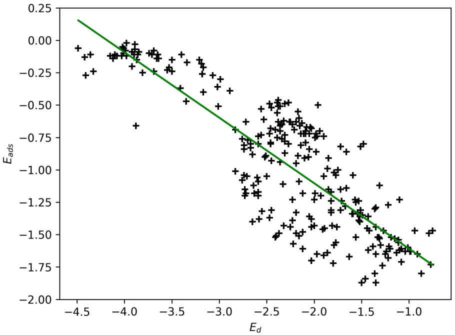
$N_{coefficient}$

$= (N_{input\_function} * N_{neuron,layer1} + N_{neuron,layer1} * N_{neuron,layer2} + N_{neuron,layer2}) * 3$ (O,H,Pt)

Artrith, N.; Behler, J. *Physical Review B* **2012**, *85*, DOI: 10.1103/PhysRevB.85.045439

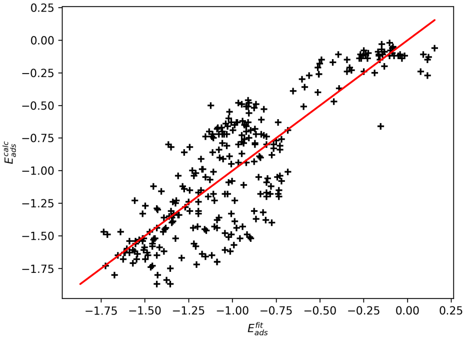
LABORATOIRE
DE CHIMIE
ENS DE LYON

The adsorption energy example



LABORATOIRE
DE CHIMIE
ENS DE LYON

Linear fit vs calculation



Only one descriptor (E_d) is not good enough!
Even if it is based on a physical model (tight binding model)

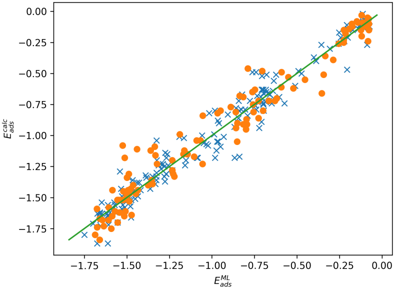
LABORATOIRE
DE CHIMIE
ENS DE LYON

What if we consider more descriptors?

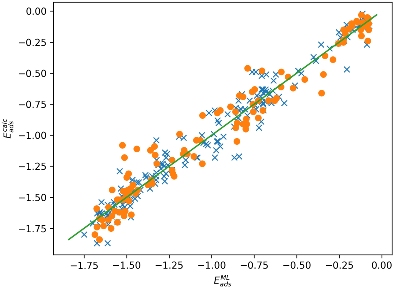
f	:	Filling of a d-band
E_d	:	Center of a d-band
W_d	:	Width of a d-band
γ_1	:	Skewness of a d-band
γ_2	:	Kurtosis of a d-band
W	:	Work function
r_0	:	Atomic radius
r_d	:	Spatial extent of d-orbitals
IE	:	Ionization potential
EA	:	Electron affinity
χ_0	:	Pauling electronegativity
χ	:	Local Pauling electronegativity
V_{ad}^2	:	Interatomic d coupling matrix element

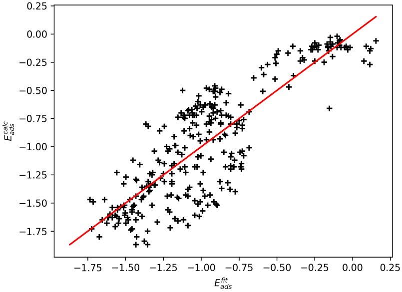
LABORATOIRE

ML vs calculation



Ma, X. et al. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3528–3533





ML significantly improves the fit by utilizing many descriptors.

ML needs input with chemistry insight in it.

ML is a tool, not magic.

Ma, X. et al. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3528–3533

LABORATOIRE DE CHIMIE ENS DE LYON

Domain of applicability

Models must be treated with care

Beware of **overfitting**:

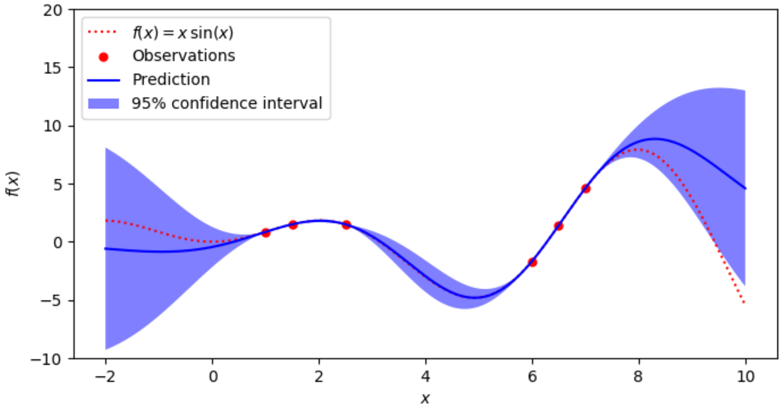
- Especially with high-dimensionality
- Quality estimator: **cross-validation** is a good starting point

Learned models are meant for **interpolation**, not extrapolation:

- Would require physico-chemical justification
- Training set should cover your subsequent usage

LABORATOIRE DE CHIMIE ENS DE LYON

Extrapolation illustration



LABORATOIRE DE CHIMIE ENS DE LYON

Checking assumptions

Theorem (No free lunch)

Any two optimization algorithms are equivalent when their performance is averaged across all possible problems

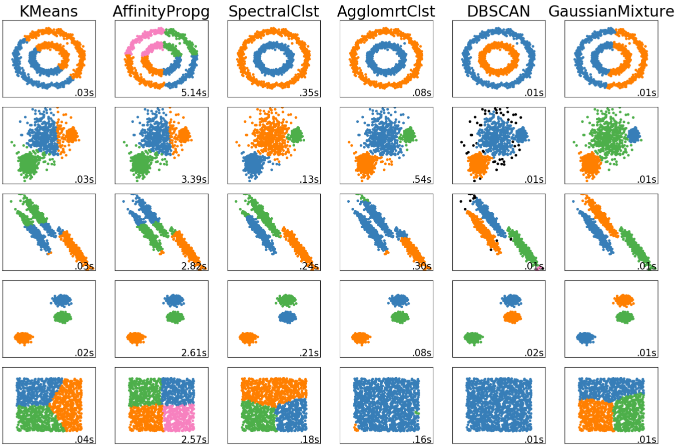
$\Rightarrow$ There cannot exist a machine learning algorithm that outperforms all other algorithms on every problem

$\Rightarrow$ A machine learning algorithm can be better than an other **only** under specific assumptions

The moral being: compare and select the algorithm that fits the best your problem

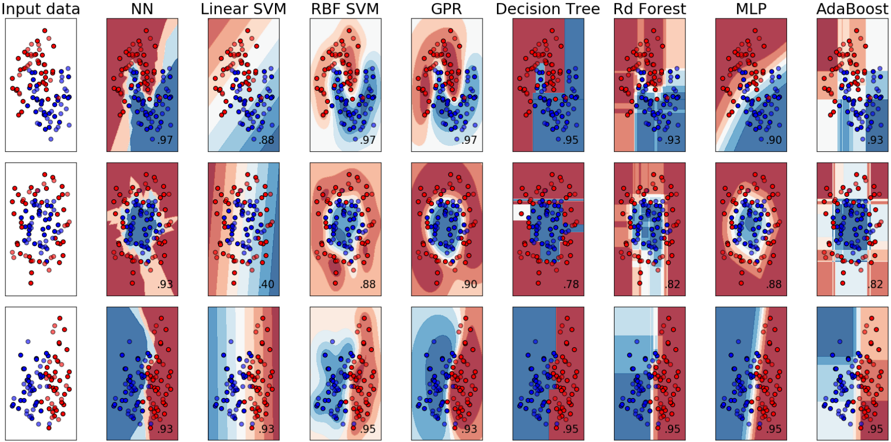
LABORATOIRE DE CHIMIE ENS DE LYON

Clustering comparison



LABORATOIRE
DE CHIMIE
ENS DE LYON

Classifiers comparison



LABORATOIRE DE CHIMIE ENS DE LYON

Another paradigm?

Third basic paradigm of machine learning: **reinforcement learning**

Aim: apply best policy to minimize regrets, without initial expertise (learn policies on the fly)

Trade-off between exploration and exploitation

Applications: decision making, global minimization, ...

Algorithms: MCTS, genetic algorithms, ...

Grandmaster level in StarCraft II using multi-agent reinforcement learning.png

nature > articles > article

MENU

nature

Article | Published: 30 October 2019

Grandmaster level in StarCraft II using multi-agent reinforcement learning

Oriol Vinyals, Igor Babuschkin, [...] David Silver

Nature **575**, 350–354(2019) | Cite this article

LABORATOIRE
DE CHIMIE
ENS DE LYON

The Singularity is Near