Christophe Morisset
IA-UNAM

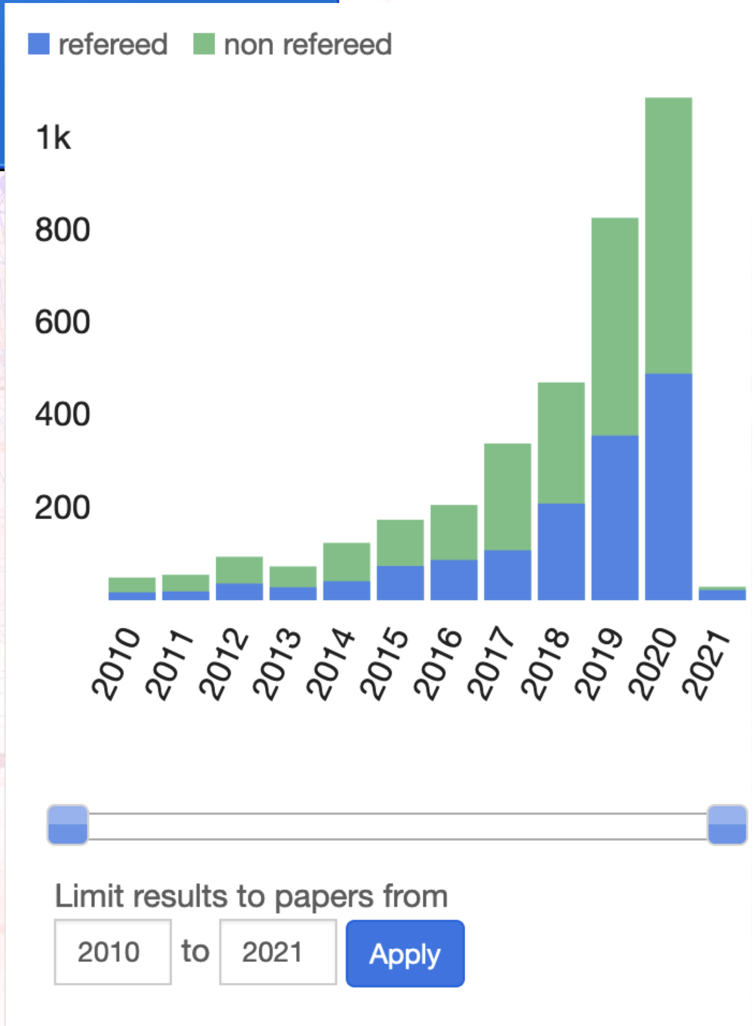
Machine Learning (for astronomers)

Machine learning - Introduction

- Machine learning is a branch of **algorithmic** that manages models for a sample of data, based on a set of **examples**, to **predict** behavior of another set of data (training set and test set).
- It is part of “Artificial Intelligence”.
- Its performances recently increased due to improvements in hardware (GPU) and software (libraries).
- Deep learning is machine learning ** 100 (in terms of data set size, and number of neurons).
- It is widely used (and developed) every day by e.g. GAFAS.

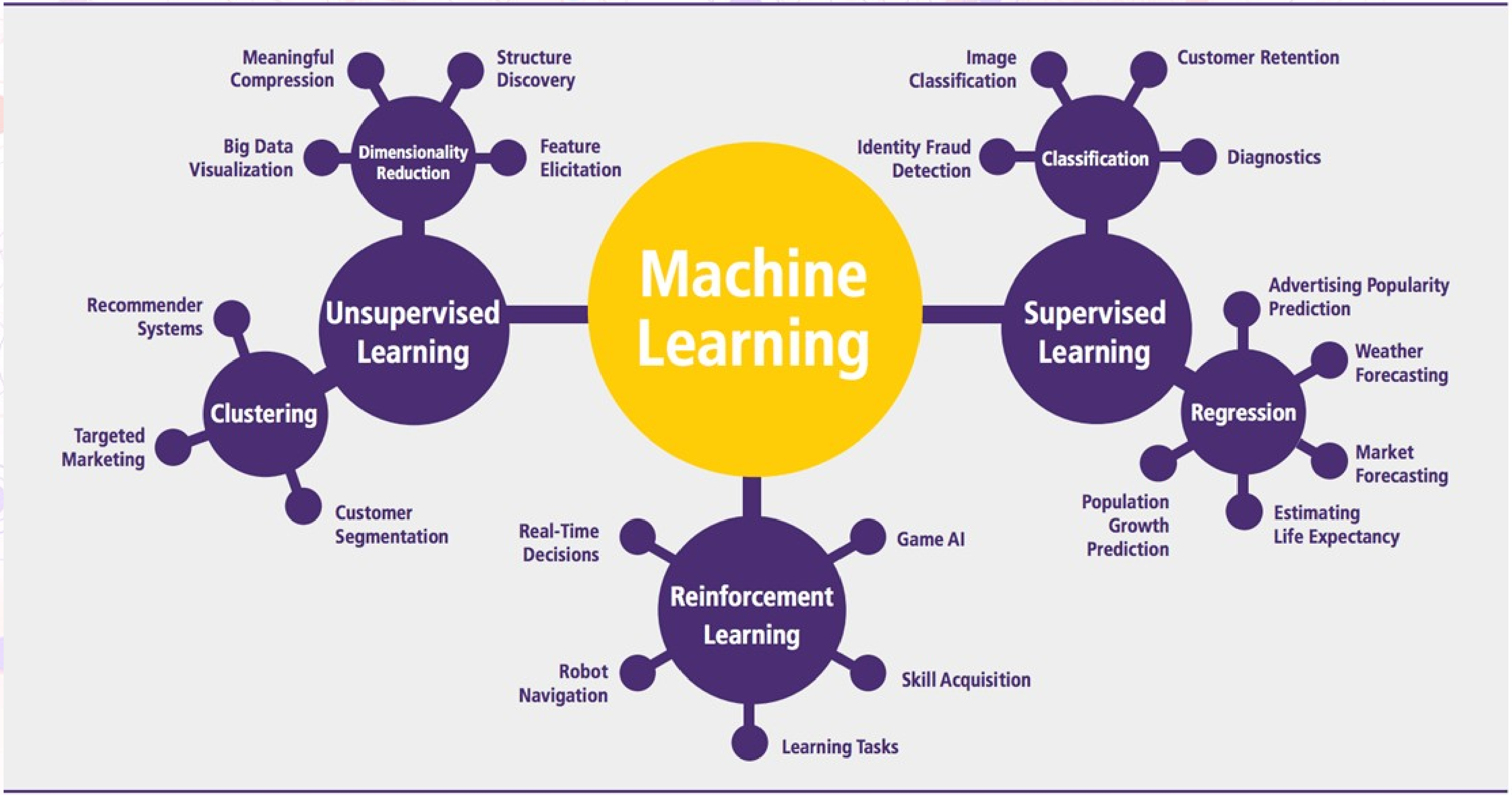
ML use in astronomy

Astronomy papers in ADS containing "Artificial Intelligence" or "Machine learning" or "Deep learning" in the abstract.



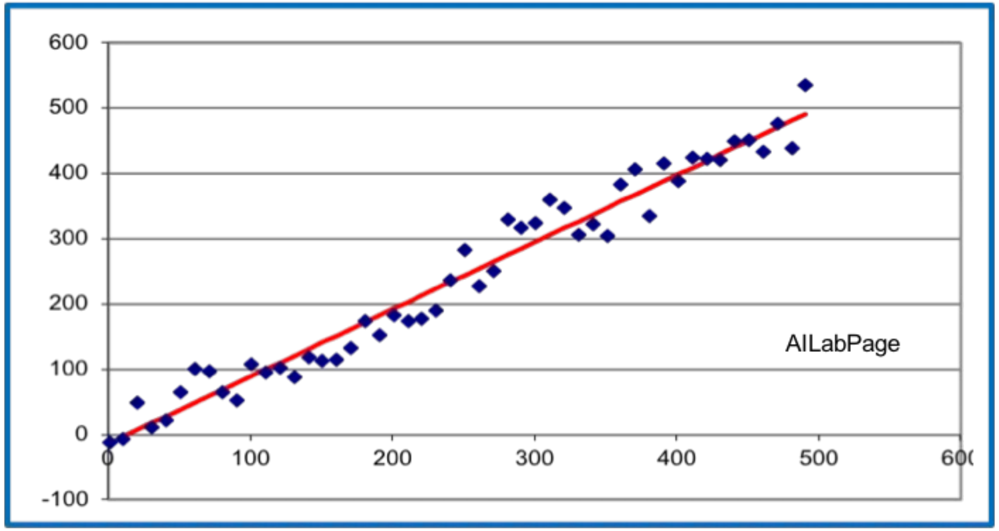
Limit results to papers from 2010 to 2021

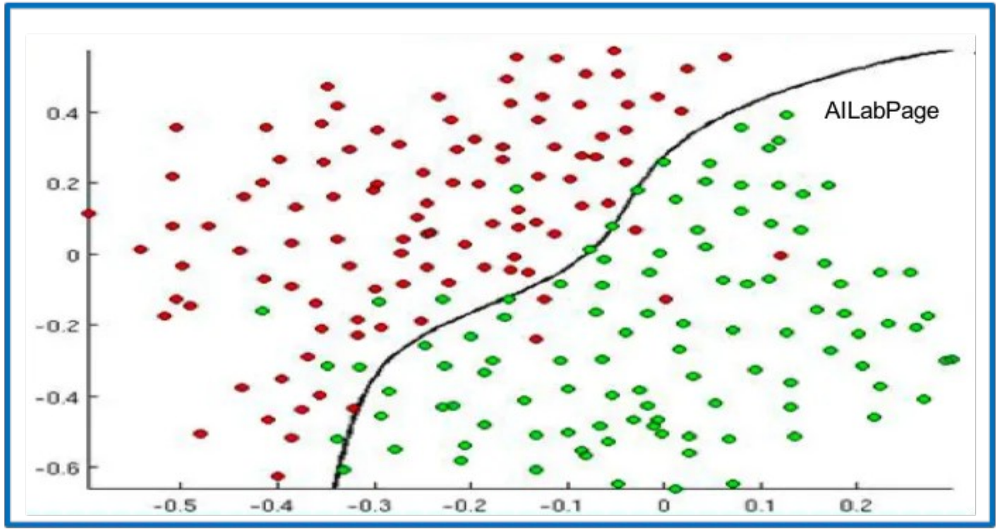
Machine Learning : a whole world

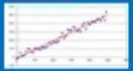


2 main kinds of predictions from **supervised learning**

- Classification: predict to which set of **categories** an elements belongs. Example: pictures of cats or dogs. Used for example in autonomous cars.
- Regression: predict a **value**. Example: price of an house given some properties (situation, number of rooms, pool, garden).







Regression

1. The system attempts to predict a value for an input based on past data.
2. Real number / Continuous numbers – Regression problem
3. Example – 1. Temperature for tomorrow

Classification

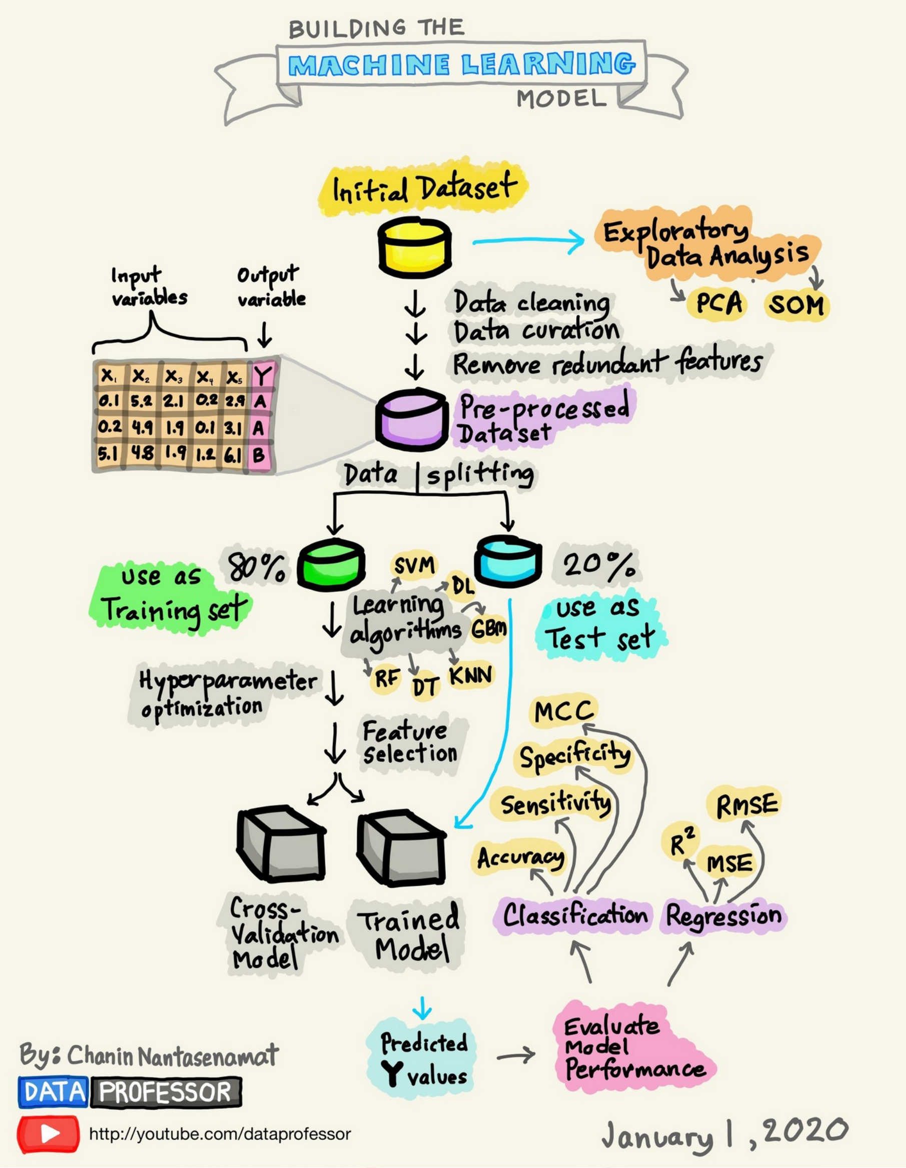
1. In classification, predictions are made by classifying them into different categories.
2. Discreate / categorical variable – Classification problem
3. Example – 1. Type of cancer 2. Cancer Y/N

Different ML models

- Artificial Neural Networks
- Decision Trees
- Support Vector Machines
- Regression Analysis
- Bayesian Network
- Genetic Algorithms
- Boosting
- … check https://scikit-learn.org/stable/supervised_learning.html

ML flow chart

Generic description of the steps needed to build a Machine Learning Model.



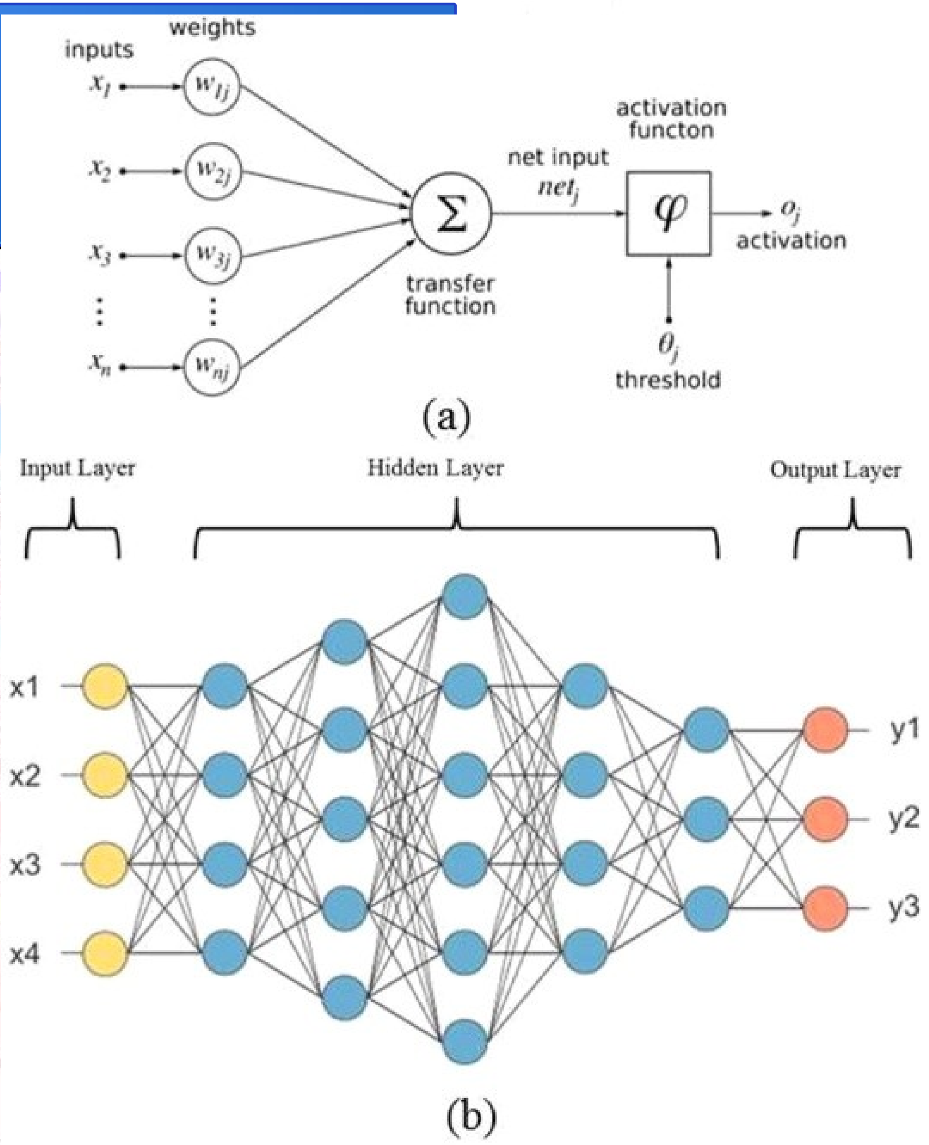
X_1	X_2	X_3	X_4	X_5	Y
0.1	5.2	2.1	0.2	2.9	A
0.2	4.9	1.9	0.1	3.1	A
5.1	4.8	1.9	1.2	6.1	B

Python ML libraries

- Numpy, Scipy, Pandas: build the algorithm from scratch block by block.
- **Scikit-learn**: supports most of the ML algorithms (ANN, SVM,…). First developed by French Institute for Research in Computer Science and Automation (2010).
- **TensorFlow**: developed by Google.
- Theano: from Montreal Institute for Learning Algorithms (Canada). Now developed by pyMC3 team.
- Keras: on top of other. Now included in TensorFlow.
- PyTorch: developed by Facebook.

Artificial Neural Network

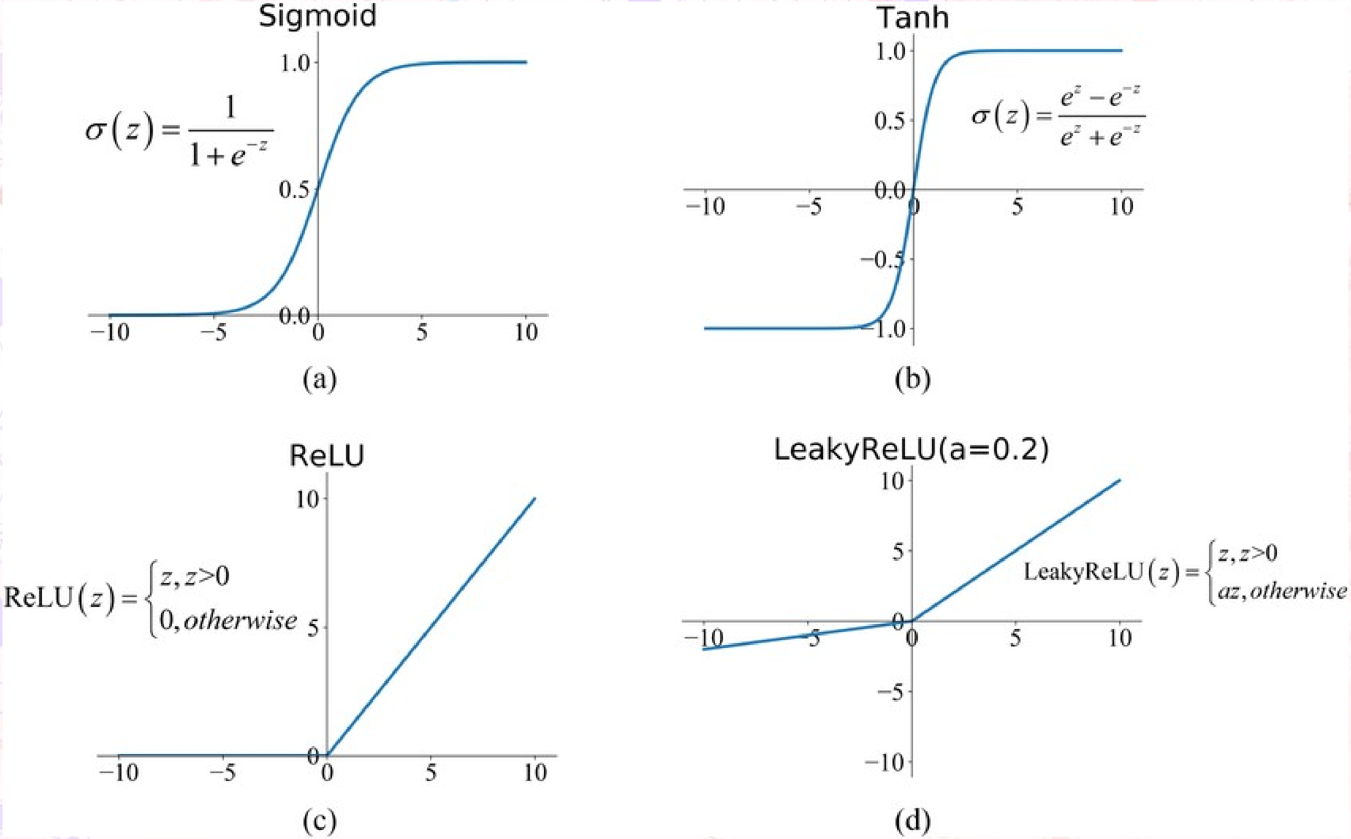
- Each neuron receives data (inputs) and produces a single output.
- The output is obtained by applying an activation function to the weighted sum of the inputs
- A constant term can also be added (bias).
- Neurons are grouped together by layers.



(a)

(b)

Activation functions



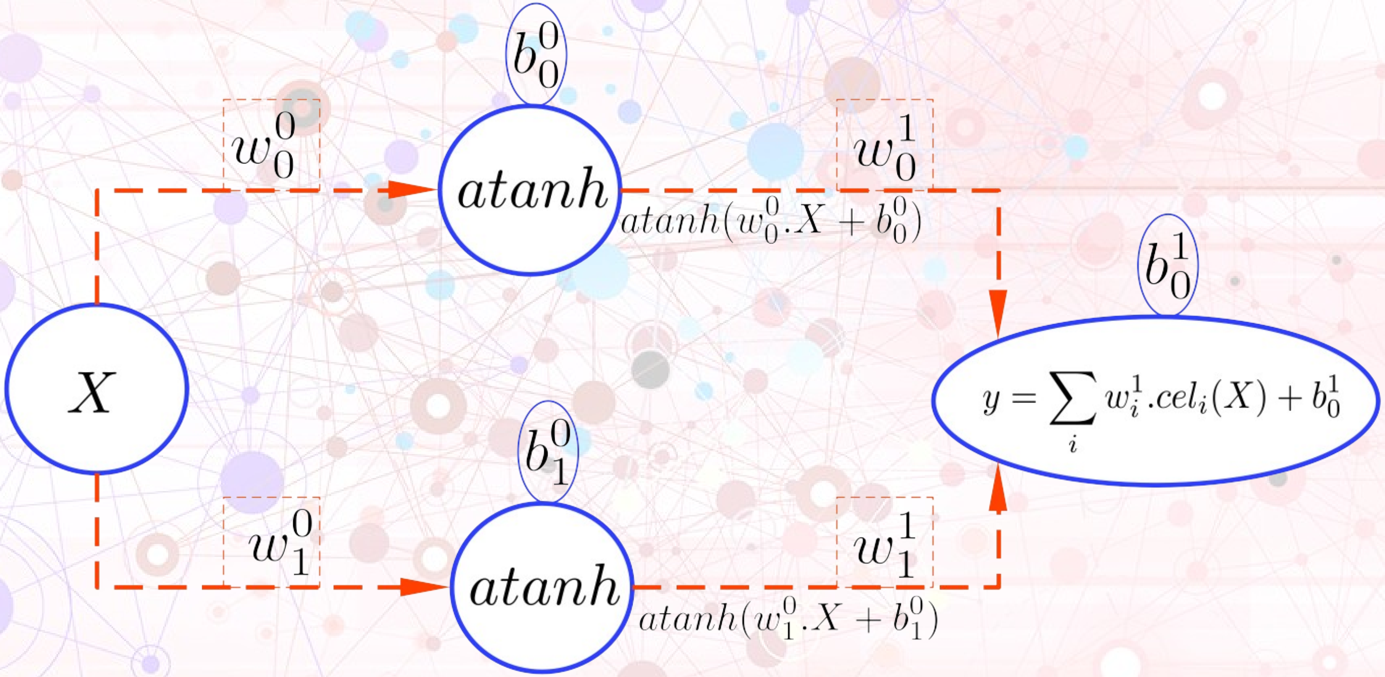
(a)

(b)

(c)

(d)

Simple example of a 2 neurons network



$$y(X) = w_0^1.atanh(w_0^0.X + b_0^0) + w_1^1.atanh(w_1^0.X + b_1^0) + b_0^1$$

Example of ANN compared to polynomial fit

See ANN.ipynb

Discretization of the output

- A regressor outputs a single value.
- One can discretize the output on an ad-hoc binning of the output, and call a classificador rather than a regressor: it will output the probability to be in each bin of the output range.
- Have a look at https://github.com/Morisset/AI4neb/blob/master/docs/SquareDiscret.ipynb

Backward method and forward method.

- A very common situation in physics is that some parameters have effects on some observables. Models can compute the observables, given the input parameters:

  O = M(P).

- Then one have some observations for which we want to determine the parameters: P’ = M^{-1}(O’).
- The problem is that most of the time M^{-1} is not well defined, and can even be degenerated.

Backward method and forward method.

- A solution may to train a ML algorithm to obtain P from O.
- Discretization of the output (here P) helps to detect degeneration
- One can also train the ML algorithm to predict O from P (what M does, but needing more CPU time), and then use another algorithm (Genetic method, MCMC) to find “all” the values of P that give O using the ML as model (fast).

Machine learning, conclusions

- Some limitations of ML methods
  - Biases, ethics.
  - Loose of connection with the original data: adjusting hyperparameters instead of physical parameters.
- Importance of preparing the data: help the process (using log values, make the ratio before, PCA dimension reduction)