Machine Learning (for astronomers) (regressors only)

Christophe MORISSET

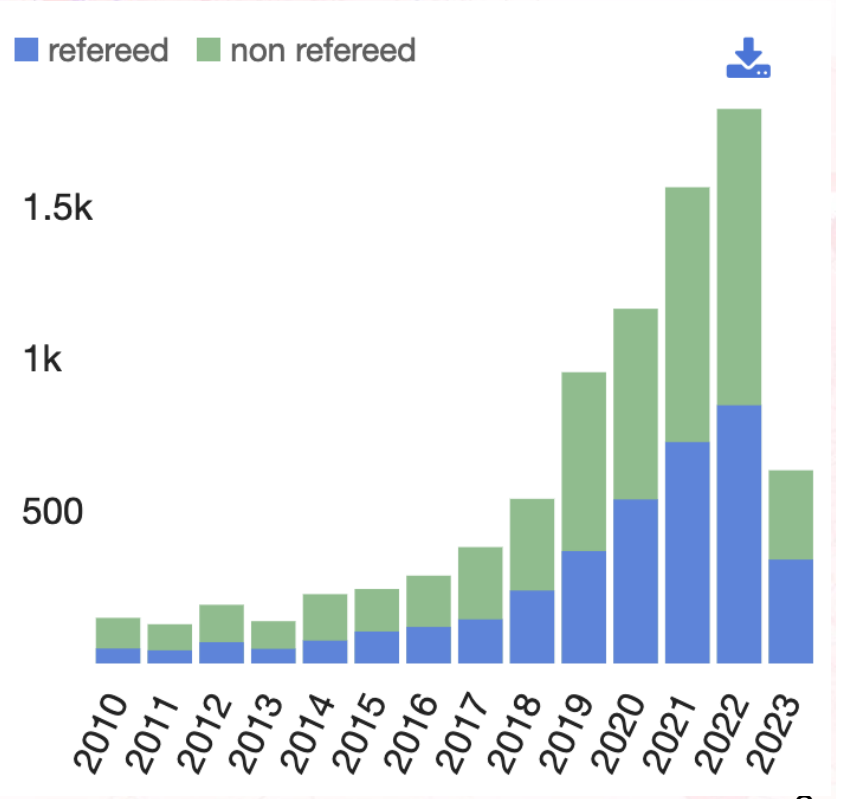
IA – UNAM Ensenada

Machine learning - Introduction

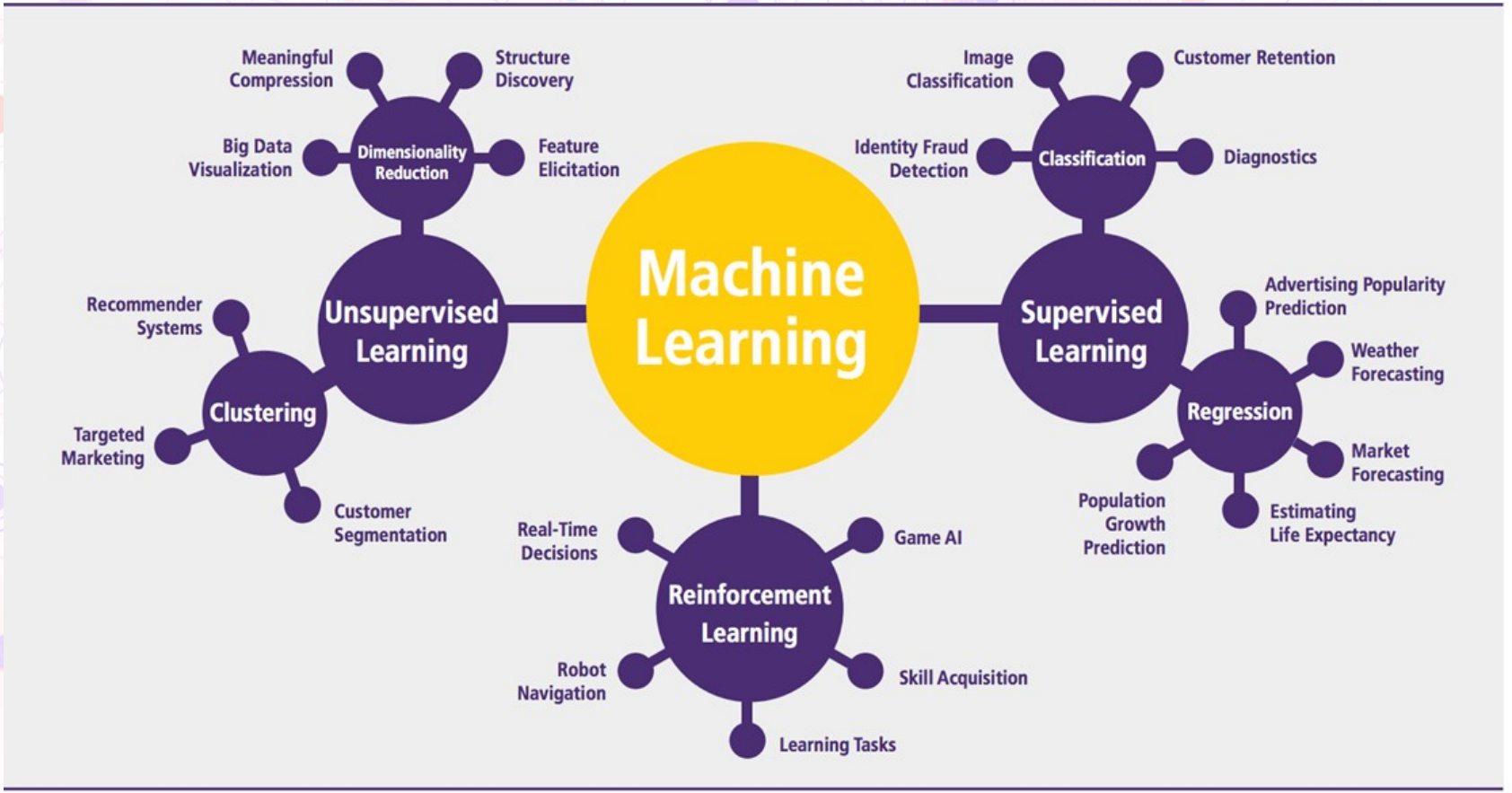
- Machine learning is a branch of **algorithmic** that manages models for a sample of data, based on a set of **examples**, to **predict** behavior of another set of data (training set and test set).
- It is part of “Artificial Intelligence”.
- Its performances recently increased due to improvements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



Machine Learning : a whole world



Machine Learning : a whole world

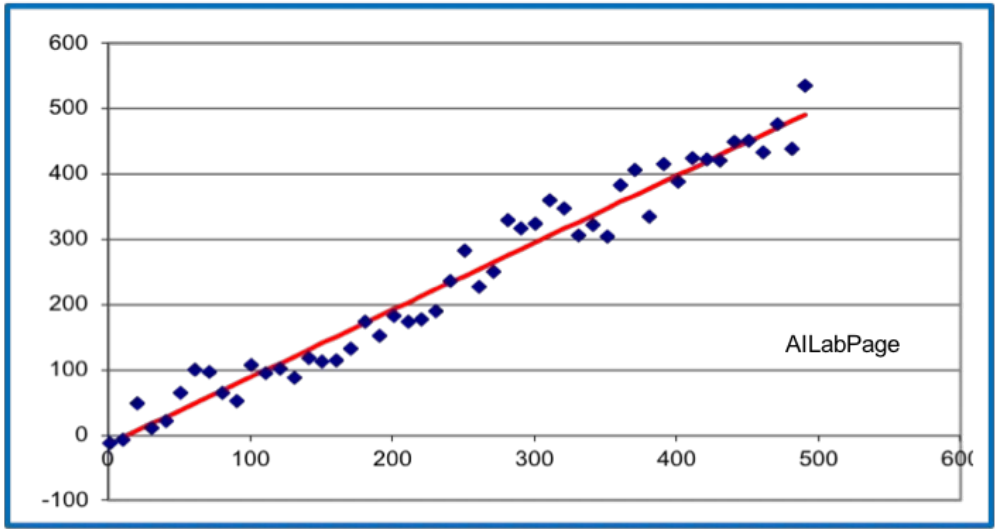
- Unsupervised Learning
  - Clustering: *Classifying a frequently repeating fast radio burst, FRB 20201124A, with unsupervised machine learning*, Chen+23
- Reinforcement Learning
  - *Application of Deep Reinforcement Learning to Major Solar Flare Forecasting* Yi+23
  - *Advances in model-based reinforcement learning for adaptive optics control* Nousiainen+22
- Supervised Learning
  - Classification: a lot e.g. *Comparison among different Clustering and Classification Techniques: Astronomical data-dependent study* Banarjee+23
  - Regression: a few + **THIS WORKSHOP**

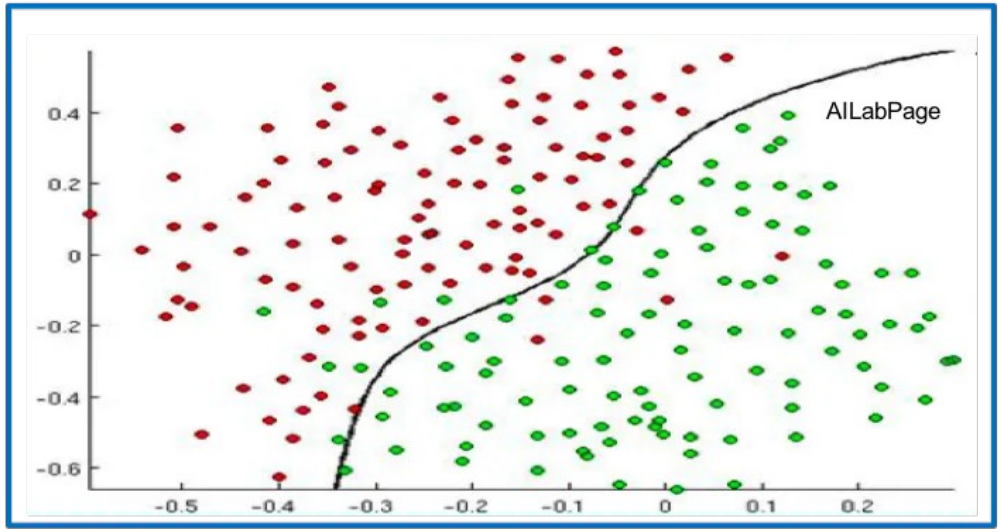
Reviews:

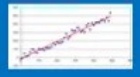
- *Surveying the reach and maturity of machine learning and artificial intelligence in astronomy,* Fluke & Jacobs 2020
- *Artificial Intelligence in Astrophysics,* book Zelinka+21

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

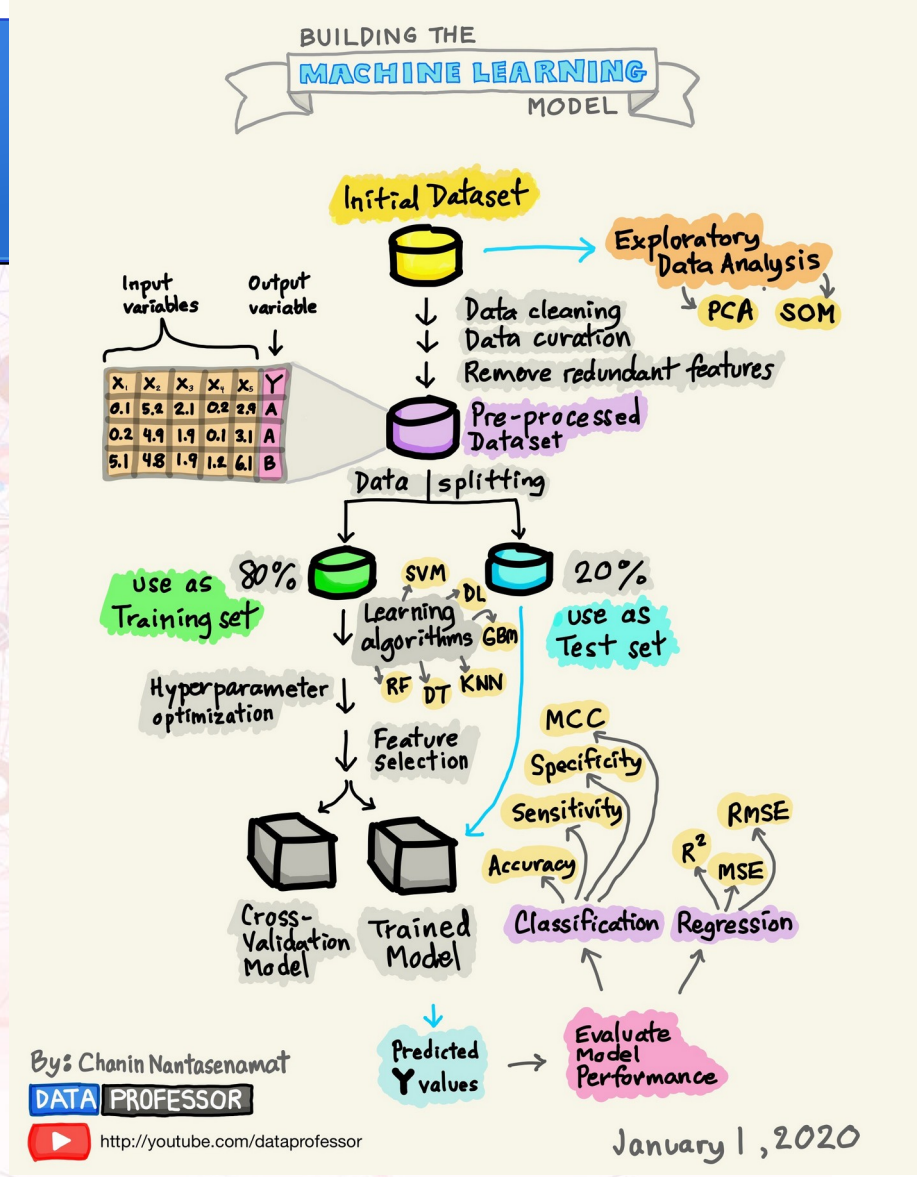
1. In classification, predictions are made by classifying them into different categories.
2. Discreate / categorical variable – Classification problem
3. Example – 1. Type of cancer 2. Cancer Y/N

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.

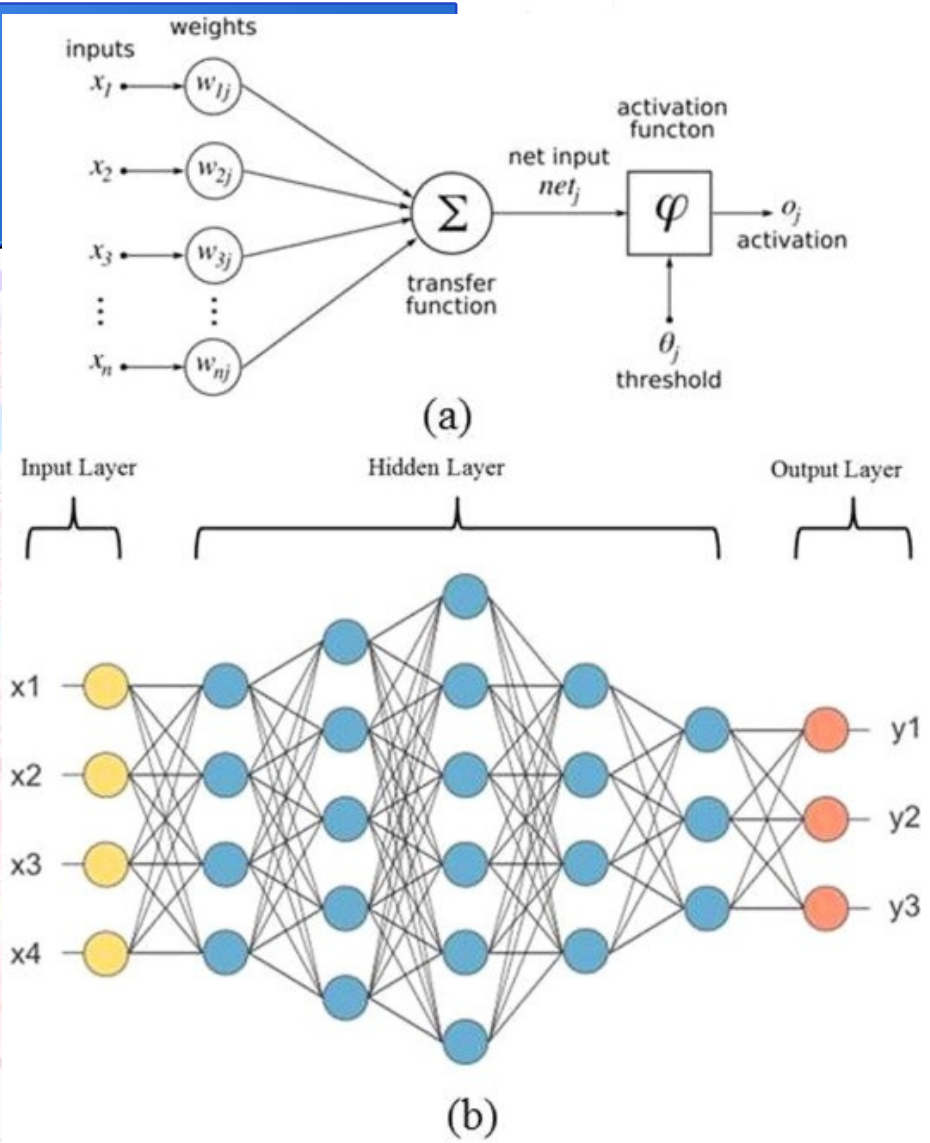


Python ML libraries

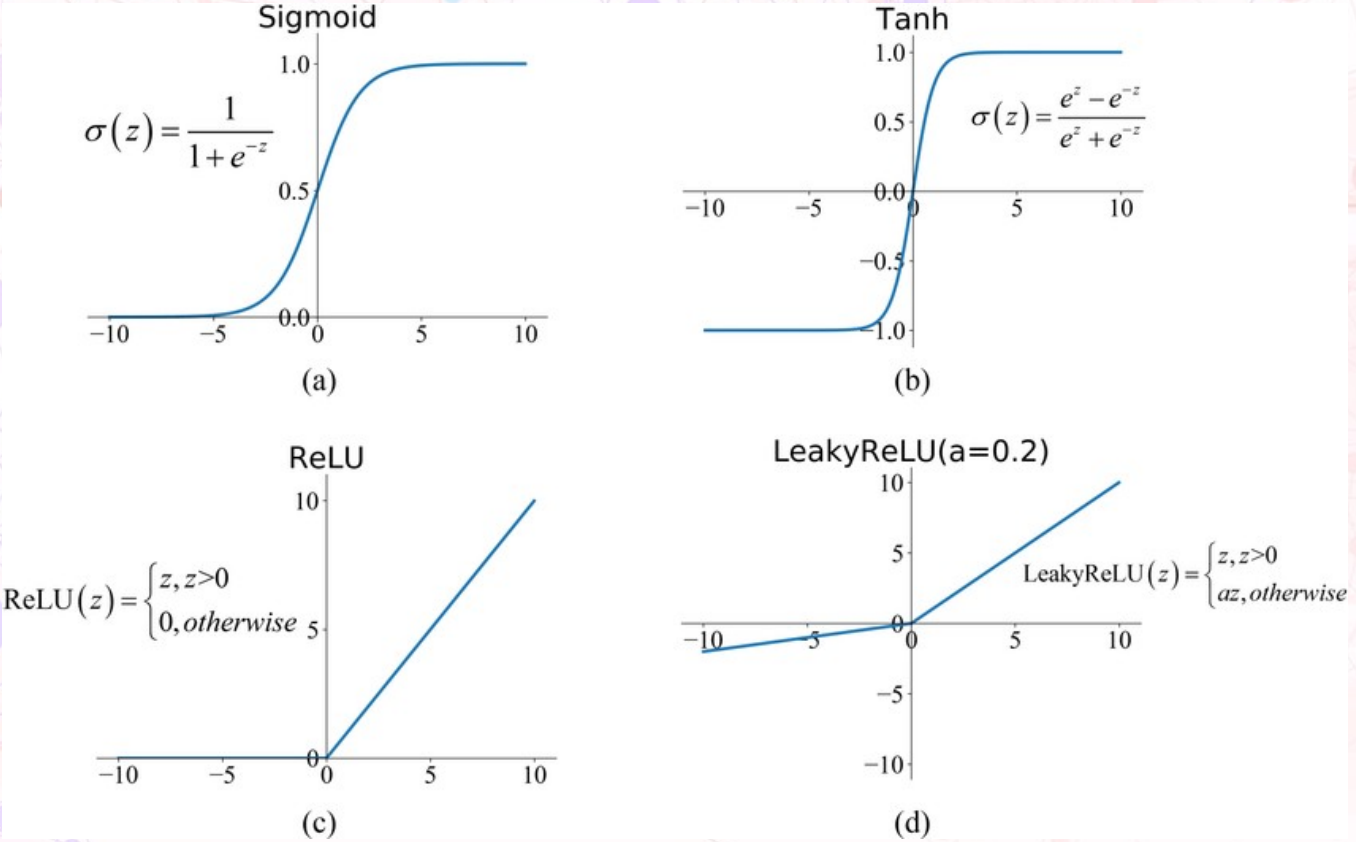
- Numpy, Scipy, Pandas: build the algorithm from scratch block by block.
- **Scikit-learn**: supports most of the ML algorithms (ANN, SVM,…). First developed by French Institute for Research in Computer Science and Automation (2010).
- **TensorFlow**: developed by Google.
- Theano: from Montreal Institute 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

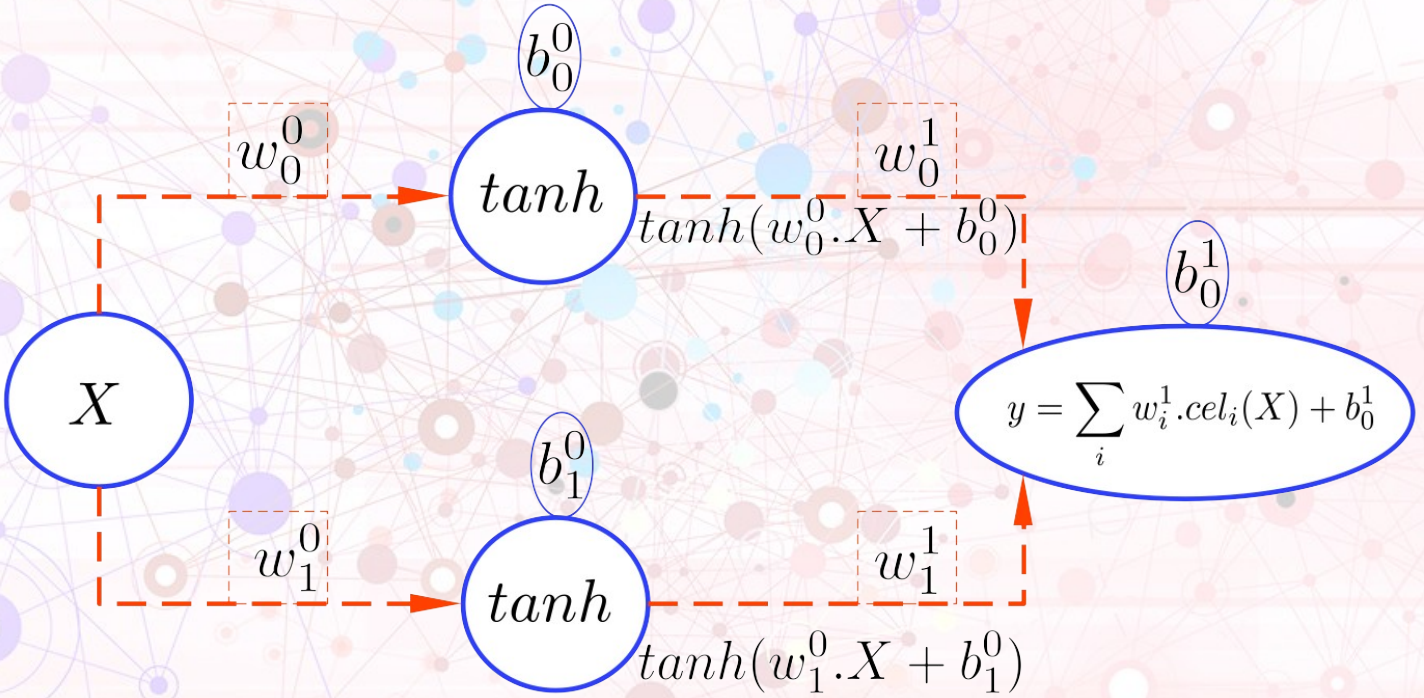
- Each neuron is a function that receives data (inputs) and produces a single output.
- The output is obtained by applying an simple activation function to the weighted sum of the inputs
- A constant term can also be added (bias or threshold).
- Neurons are grouped together by layers.



Activation functions



Simple example of a 2 neurons network



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

Example of ANN compared to polynomial fit

See 1.ComparePolynom.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 classification method rather than a regressor: it will output the probability to be in each bin of the output range.
- See 2.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 be 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)