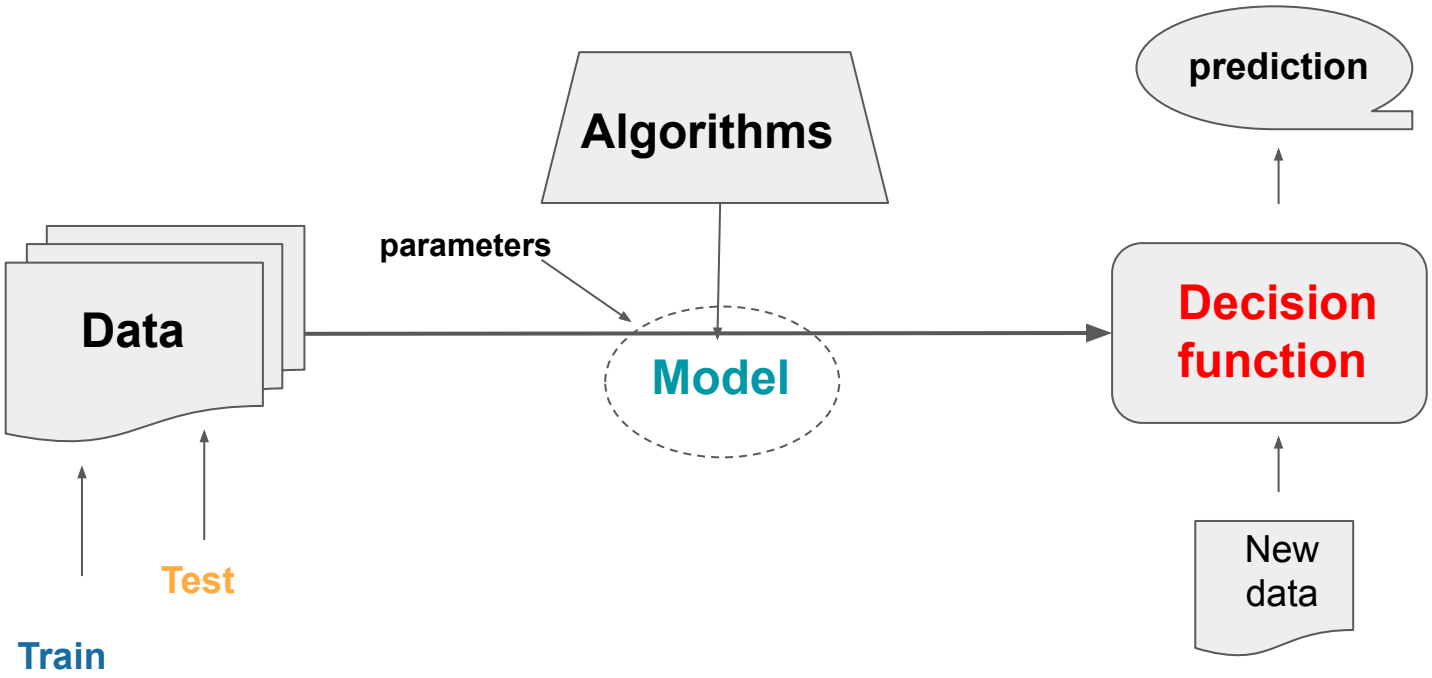
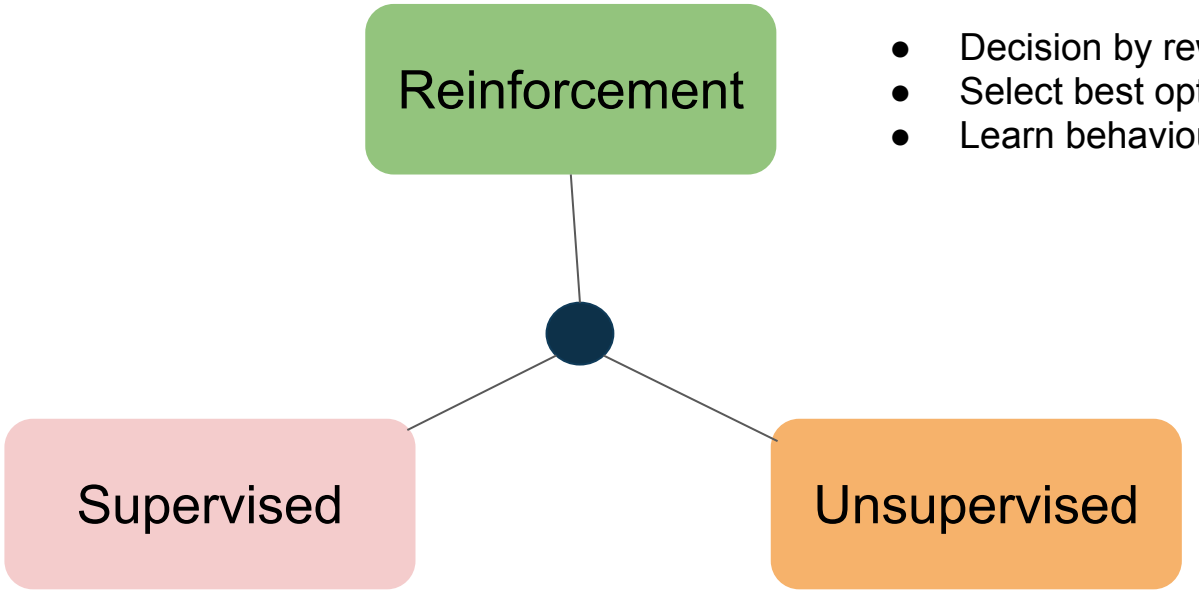
Intro Python, ML & file structure formats

Machine learning in a nutshell



Machine learning in a nutshell

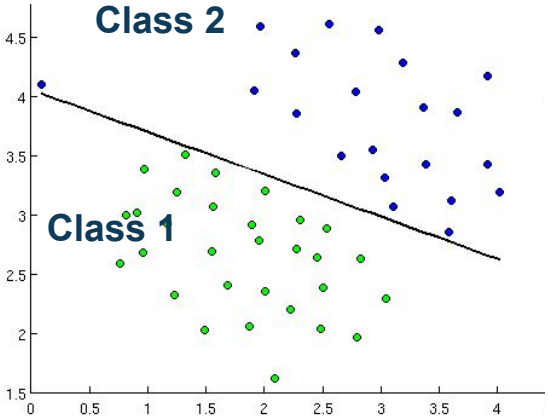


- Decision by reward
- Select best option
- Learn behaviour

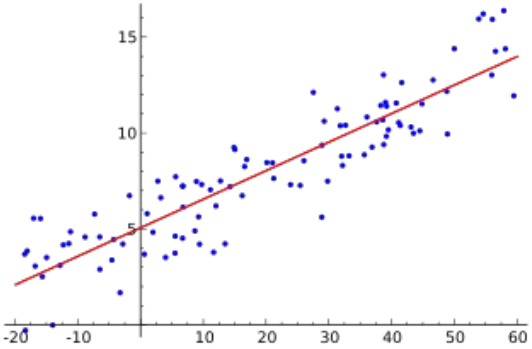
- Labeled data
- Classification (discrete values)
- Regression (real values)

- No data labels
- Clustering

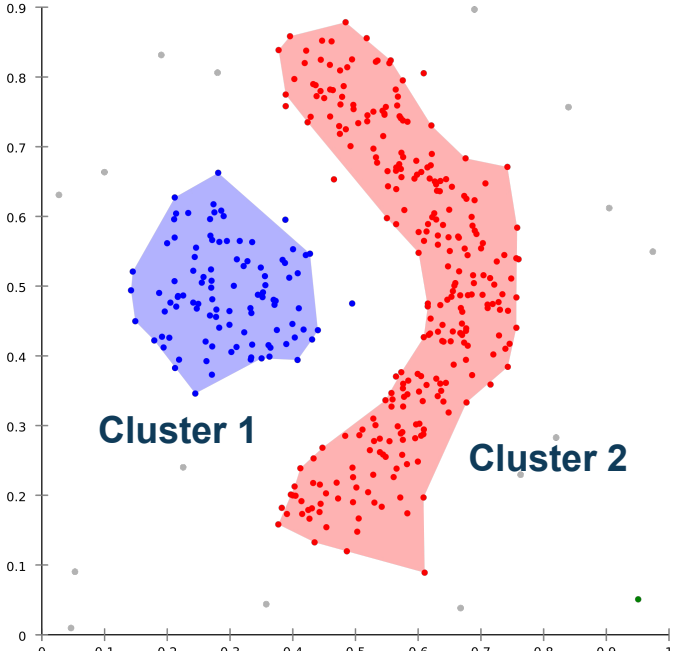
Classification / Regression / Clustering



- Classify future observations
- Known classes



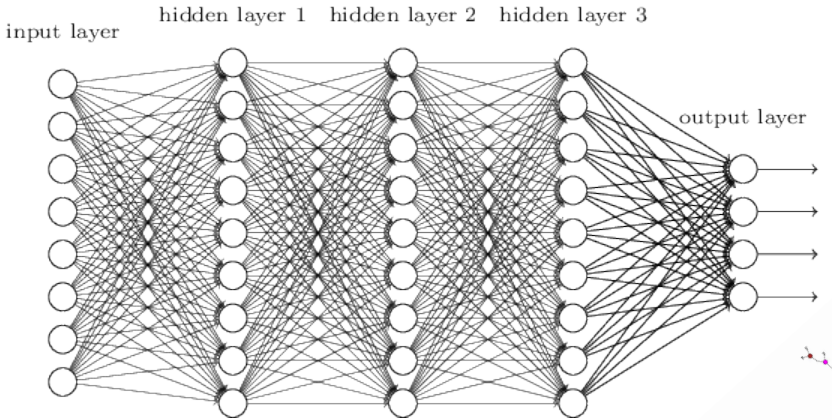
- Predict continuous attribute



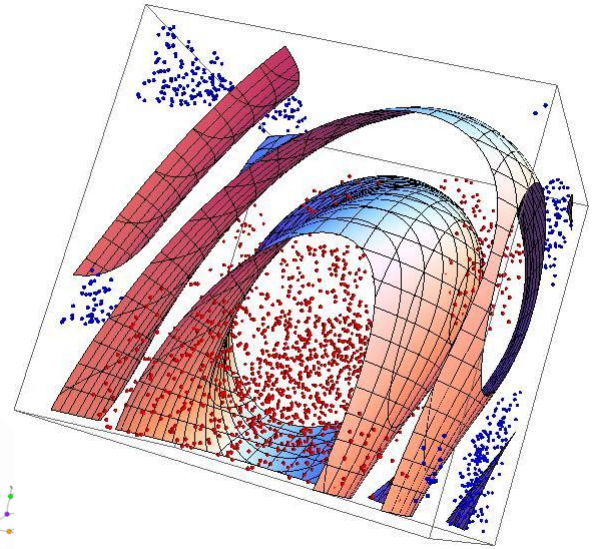
- No prior knowledge
- Discover patterns

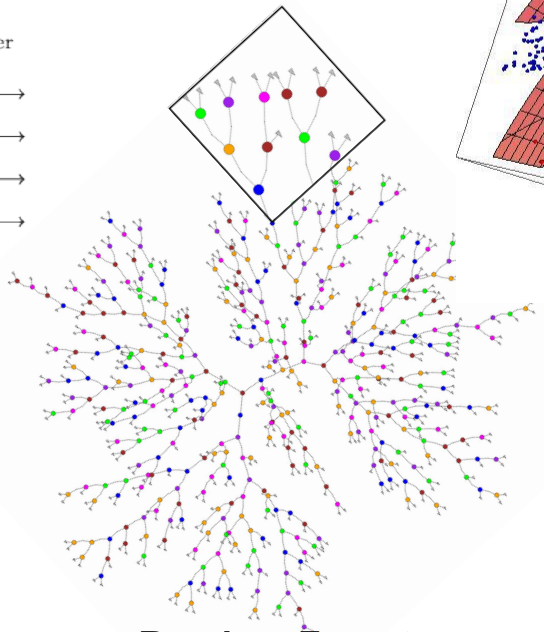
Popular ML models

Neural Networks



Support Vector Machines

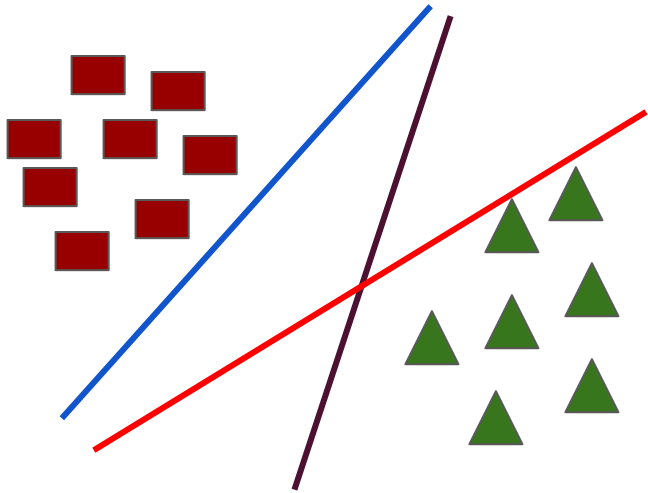




Random Forests

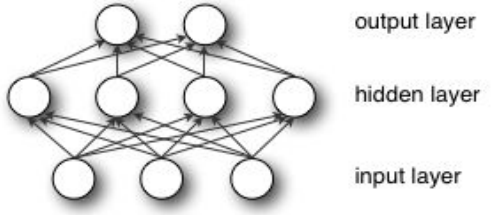
Support Vector Machines (SVM)

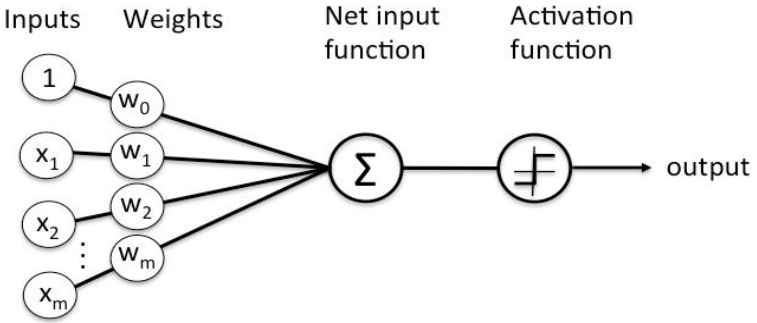
- In case of linearly separable data, an ML algorithm tries to find a decision boundary (by minimizing a classification error)
- In this case, how to choose best boundary ?
- SVM
  - Finds the most optimal decision boundary by maximizing the distance from the nearest data points of all the classes: finds a **hyperplane in N-dimensional space** (N: number of features)



Neural networks

- Inspired by biological neurons
- Designed to recognise patterns
- Key components
  - **Node**: represents an artificial neuron
  - **Weight**: importance of the node in the learning process
  - **Layer**: a set of nodes.
    - Input
    - Hidden: learns different aspects about the data by minimizing an error/cost function
    - Output
  - **Activation function**:
- Learning from sample observations by adjusting the weights to improve the accuracy of the result.
- Learning is done by minimizing the observed errors.





Images from [Ref]

Why Python ?

- General purpose language
- Easy to use
- Popular in data science community
- Integrated packages : data processing, ML, data structure saving/loading

Jupyter notebook

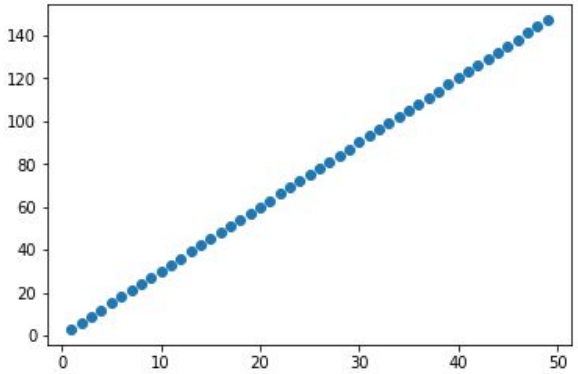
Open-source web application to create live code, equations, visualizations and text [Ref]

In [4]:
```
"""
Cirta Challenge 2019
"""
```
Out[4]: '\nCirta Challenge 2019\n'

In [3]:
```
import matplotlib.pyplot as plt

X = range(1, 50)
Y = [value * 3 for value in X]

plt.scatter(X,Y)
plt.show()
```



The Lorenz Equations ¶

```
\begin{align}
\dot{x} & = \sigma(y-x) \\
\dot{y} & = \rho x - y - xz \\
\dot{z} & = -\beta z + xy
\end{align}
```

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

Numpy

Package for scientific computing with Python [Ref]

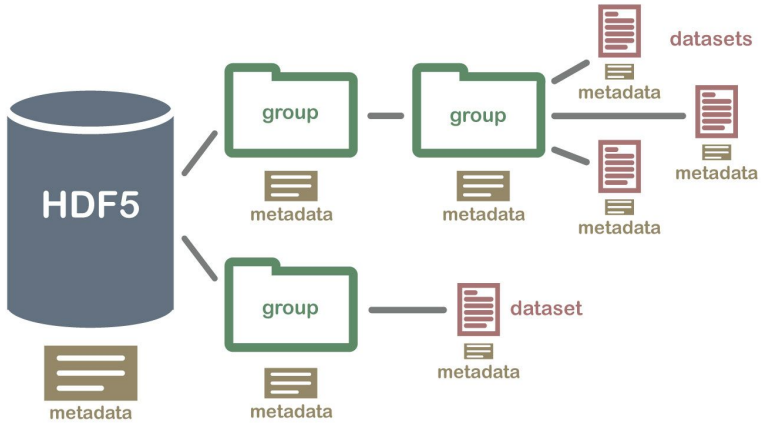
- a N-dimensional array manipulation
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

uproot

- Reader/Writer of the ROOT file format using **only Python and Numpy** [Ref]
- No dependence on C++ ROOT
- Uses Numpy to cast blocks of data from the ROOT file as Numpy arrays.
- Designed to stream data into machine learning libraries in Python

File structures: Pickle, HDF5

- Pickle
  - Used to serialize[1]/de-serialize python object structure (list, dict, etc..)
  - Converts python object into character stream
- HDF5
  - Hierarchical Data Format (HDF)
  - Supports large, complex, heterogeneous data

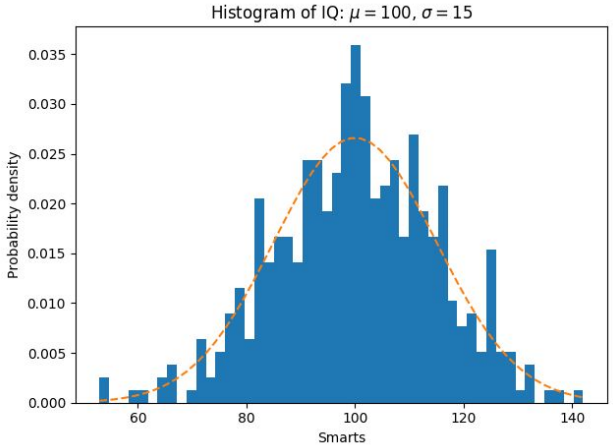


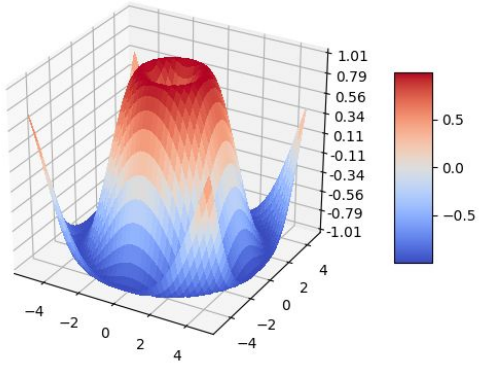
[1] Serialization is the process of translating data structures or object state into a format that can be stored or transmitted and reconstructed later

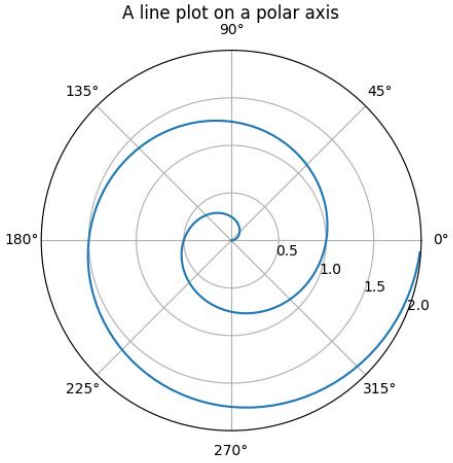
Image from Ref

Matplotlib

Plotting library [Ref]





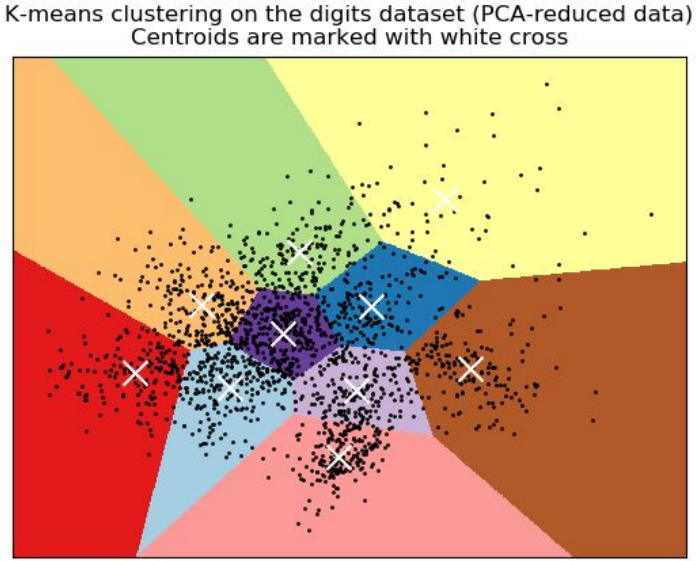


Scikit learn (sklearn) : ML

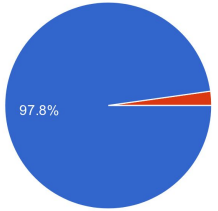
- Library for data analysis and data mining

  [Ref]

- What can we do with sklearn ?
  - Data preprocessing
  - Classification
  - Regression
  - Clustering
  - ..



Installing packages : all in one **Conda**



Windows

1. Download installer : https://docs.conda.io/en/latest/miniconda.html
2. Double-click the .exe file.

Linux

1. Download installer : https://conda.io/miniconda.html
2. In your terminal : bash Miniconda3-latest-Linux-x86_64.sh

Questions ?