Warsaw University of Technology

ALICE

Introduction to Machine Learning Techniques for HEP

Kamil Deja

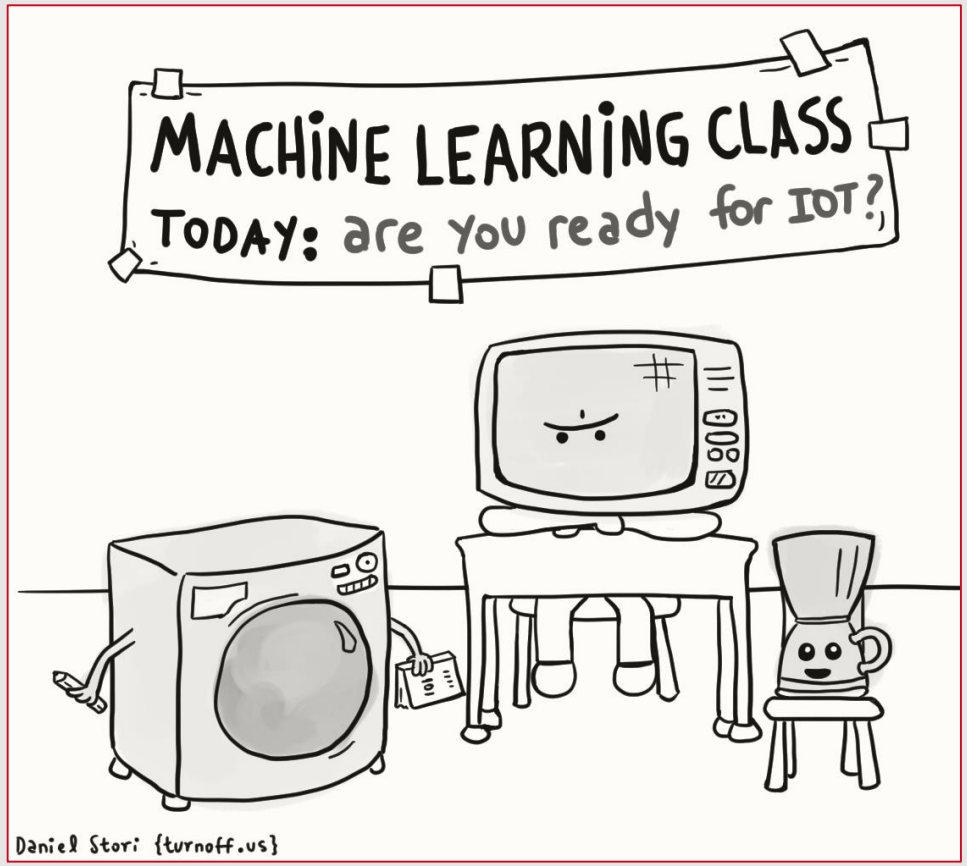
Agenda

- What is Machine Learning?
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
  - Frequent sets and Association Rules
  - Dimensionality reduction
- How to use ML for quality assurance?
- Additional Examples
- Most common tools and packages

What is machine learning?

- Algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task.
- Machine learning algorithms build a mathematical model of **training data**, in order to make predictions or decisions without being explicitly programmed to perform the task.

https://xkcd.com/1838/



Supervised learning

Supervised learning

- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.

Classification

Attributes			Target
X	**Y**	**Z**	**Class**
1	2	2	**A**
1	2	3	**B**
1	4	3	**B**
4	2	6	**A**

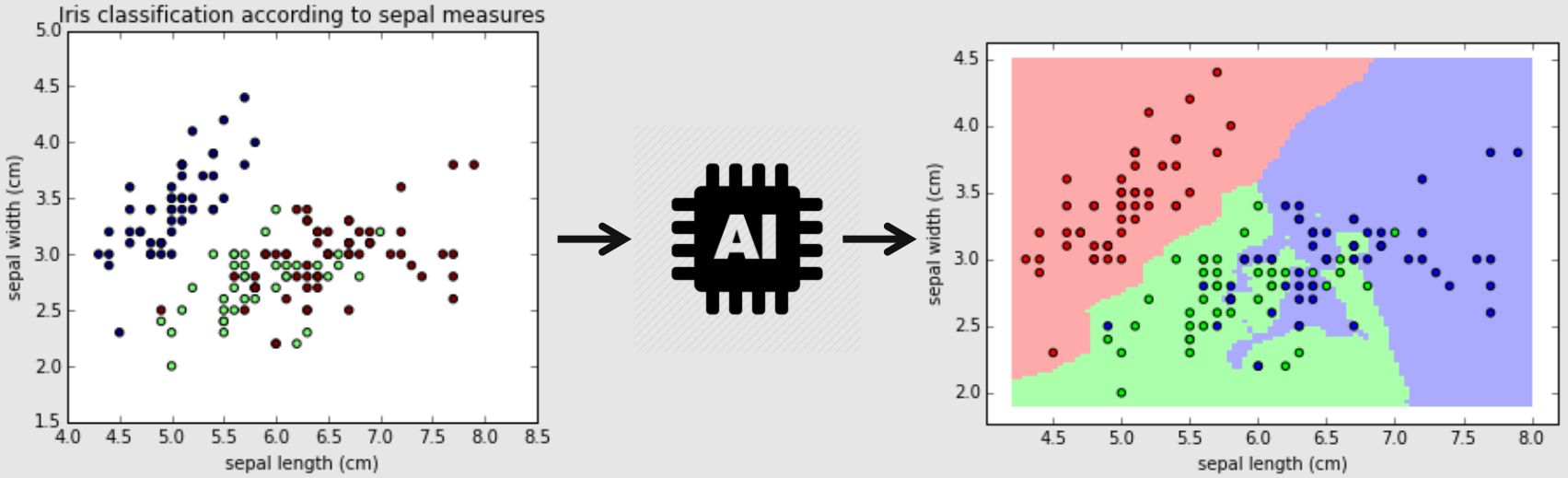
Regression

Attributes			Target
X	**Y**	**Z**	**Variable**
1	2	2	**1.3**
1	2	3	**2.1**
1	4	3	**4.93**
4	2	6	**5**

Classification

- Definition

Problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.



http://stephanie-w.github.io/brainscribble/classification-algorithms-on-iris-dataset.html

Examples

High Energy Physics:

- Particle Identification based on training set of full simulation
- Jet tagging based on simulation data
- Object tagger in ATLAS for pair production of heavy vector-like quarks with hadronic final states based on simulation data
- Run quality classification based on trending.root and RCT

Outside:

- Prediction of patients diseases based on their lab results and symptoms

Regression

- Definition

Regression is a set of statistical processes for estimating the relationships among variables.

- In Machine Learning it is usually referred as a method for predicting the value of one variable

Examples

High Energy Physics:

- Energy regression in calorimeters
- Calculate jet-underlying background in Pb-Pb

Outside:

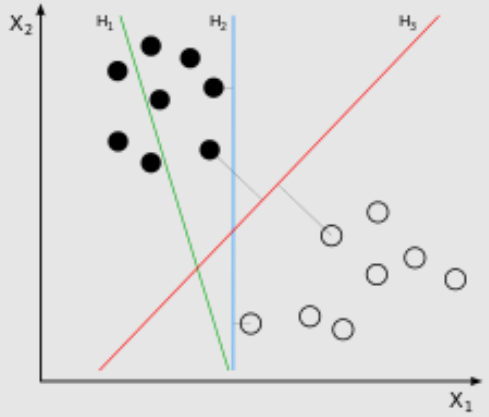
- Predicting the market revenue of the property

Classification and regression algorithms

Naïve Bayesian Classifier

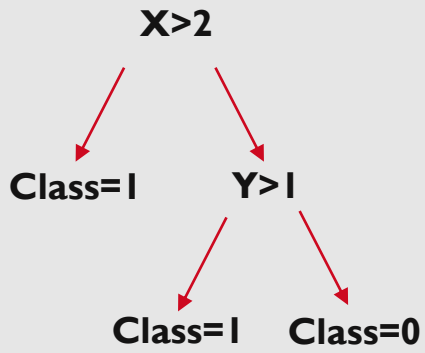
$$P(target|attributes) = \frac{P(attributes|target)P(target)}{P(attributes)}$$

Support Vector Machine (SVM)



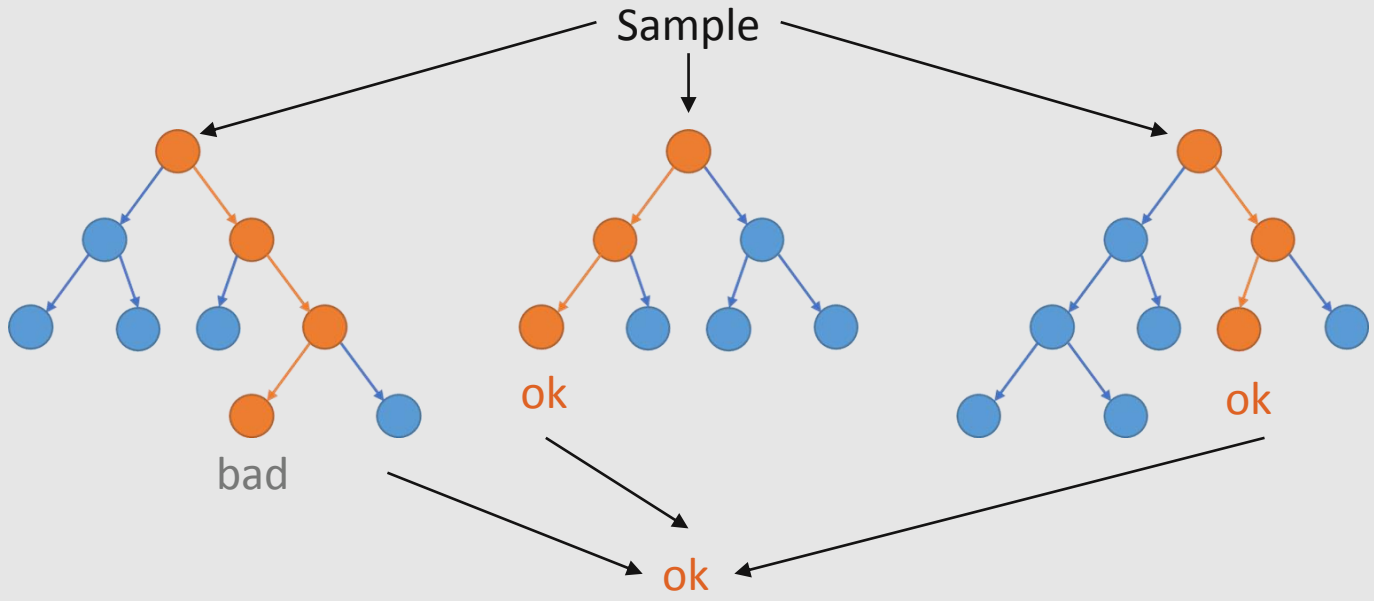
Decision Tree

X	Y	Class
3	1	1
4	2	0
2	4	1
3	3	0



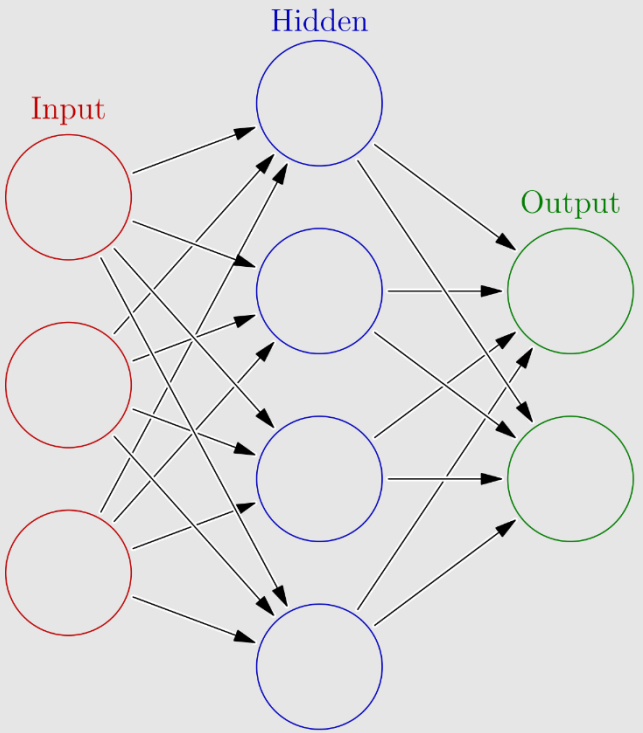
Random Forest /xgboost

- Ensemble learning methods for classification/regression, which operates on **multiple decision trees**, and outputs class based on **combined predictions** of individual trees.



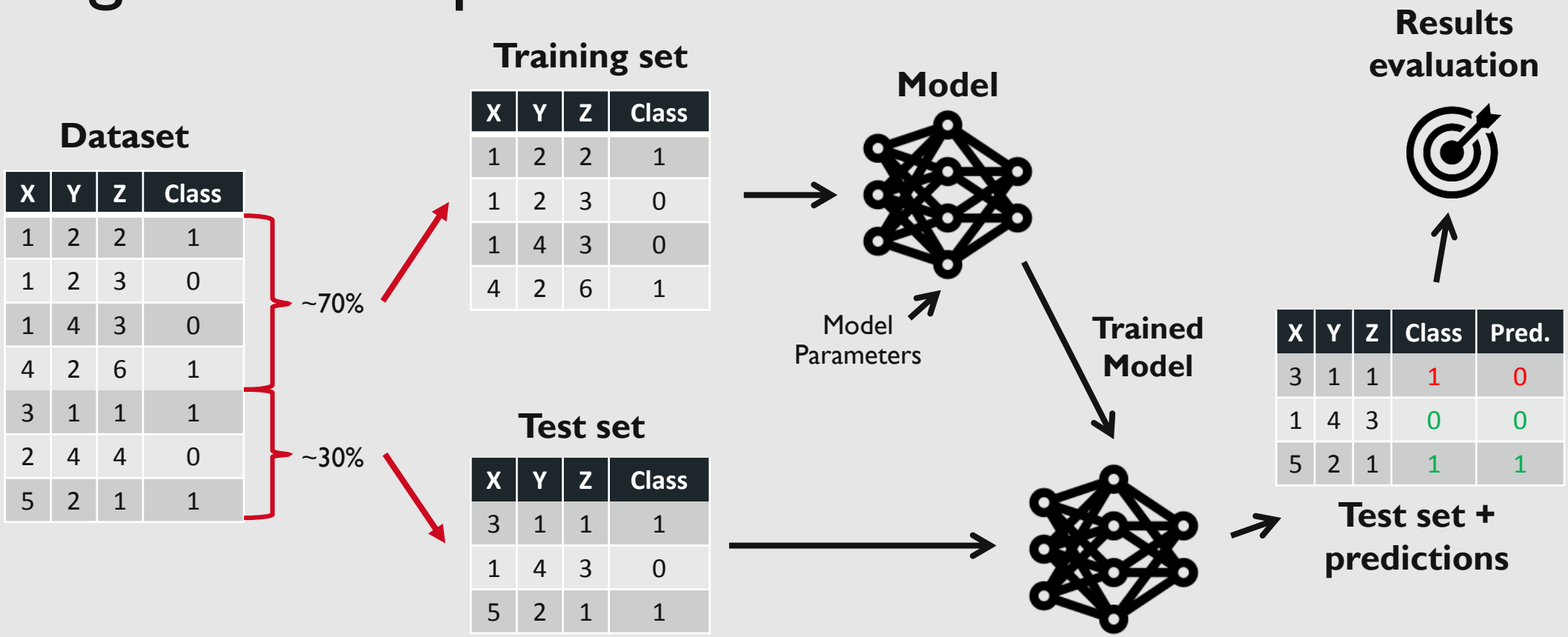
Neural Networks

- Directed graph where vertex (**neurons**) are grouped in **layers** which are connected by **weighted synapses**.
- Training is used to adjust the weights so that the relation between input and output fits the training data

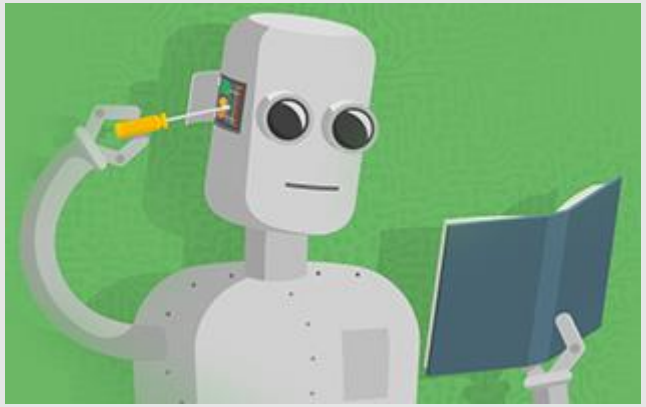


- „Machine learning framework”
- Inspired by human brain
- Commonly used for Classification/Regression, but also suitable for other tasks

Standard workflow for classification or regression experiment



Unsupervised learning



Clustering

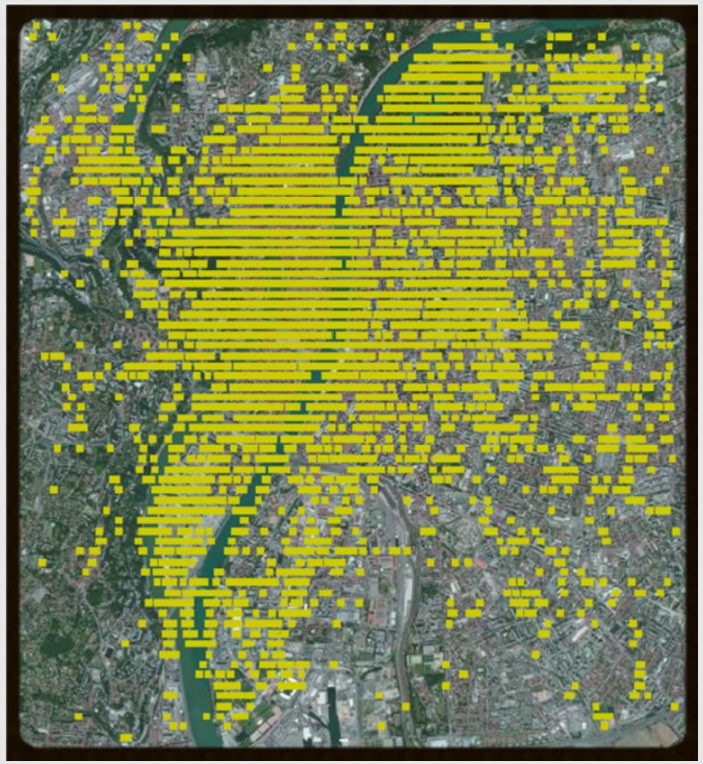
- Definition

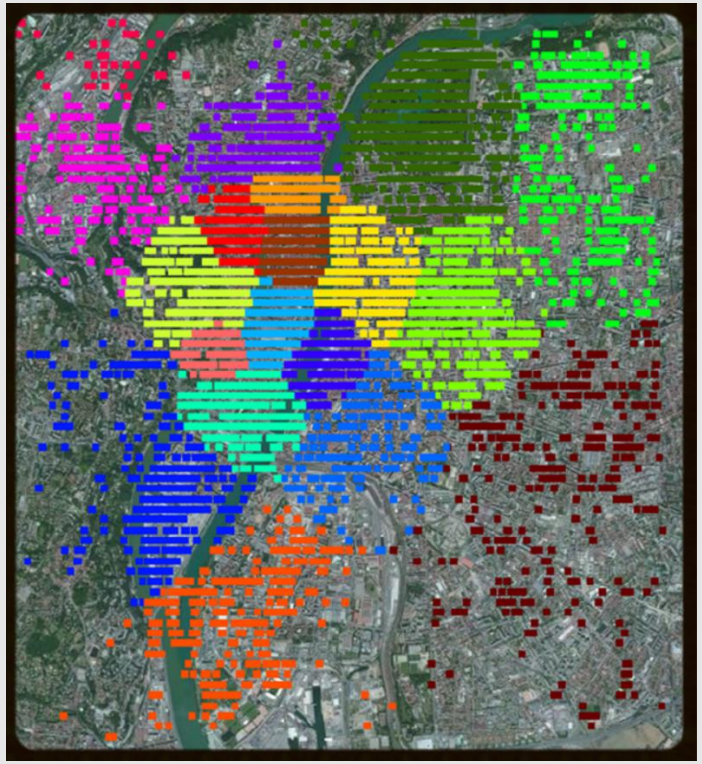
Machine Learning method of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (**clusters**).

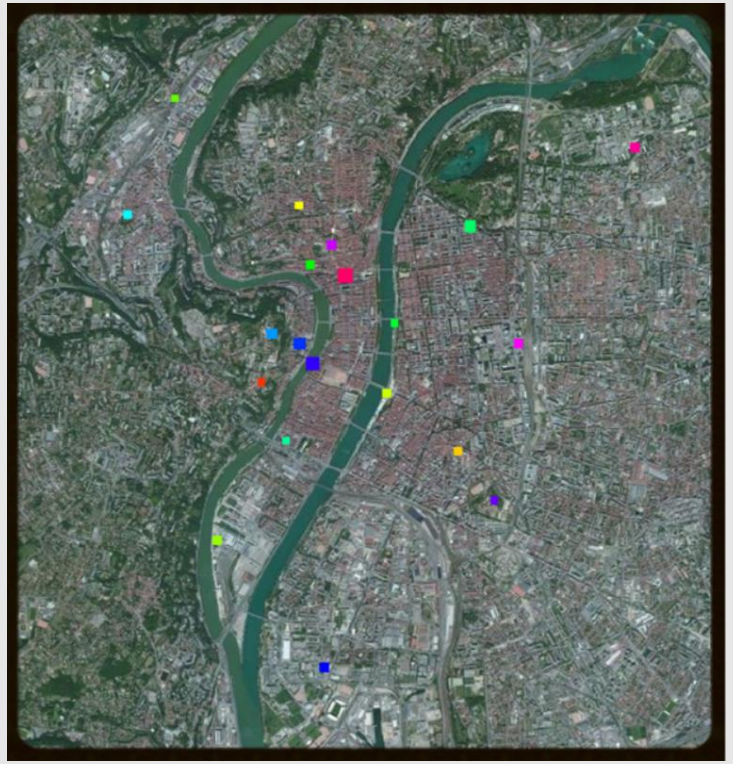
ID	X	Y	Z
1	1	2	2
2	4	5	6
3	1	1	2
4	3	4	5
5	1	2	3

ID	X	Y	Z
1	1	2	2
5	1	2	3
3	1	1	2
4	3	4	5
2	4	5	6

Example – clustering of photos in Lyon







Examples

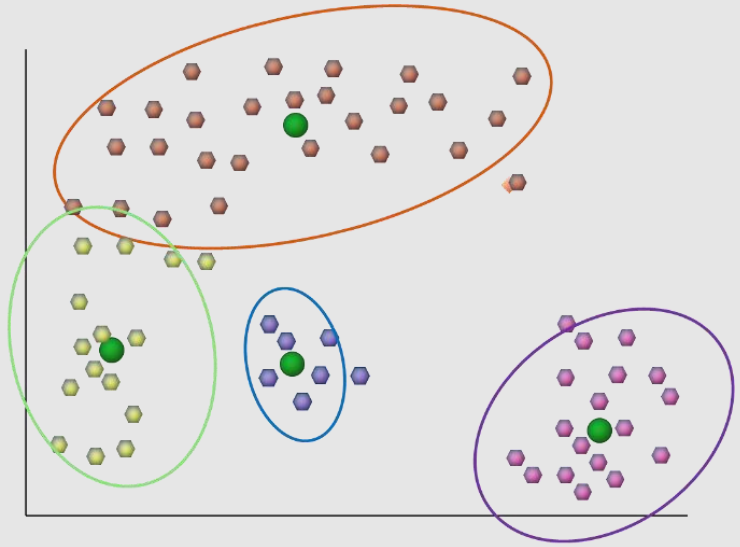
High Energy Physics:

- TrackML – Clustering of particle hits

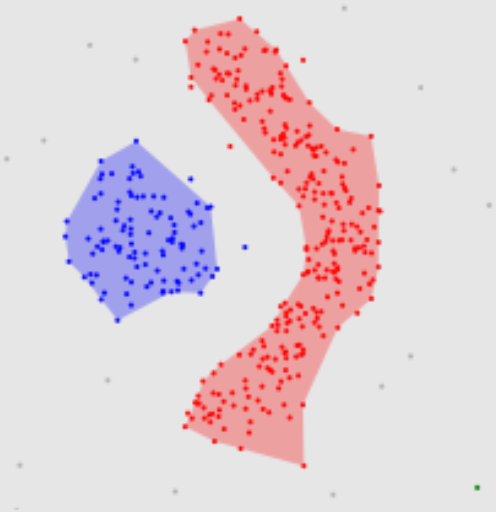
Others:

- Clustering of clients in banks/telecom/Netflix
- Phone users clustering

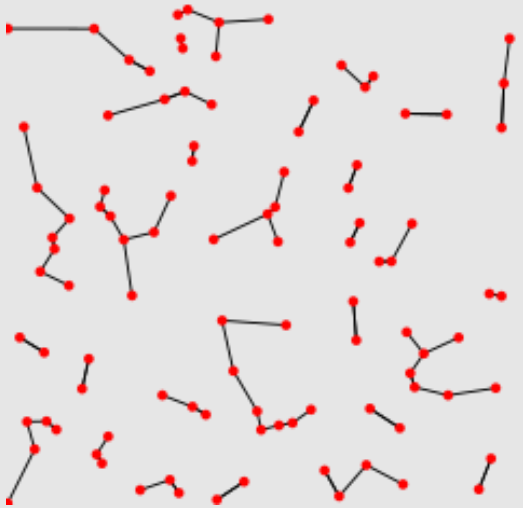
Clustering algorithms



K-means



DBSCAN

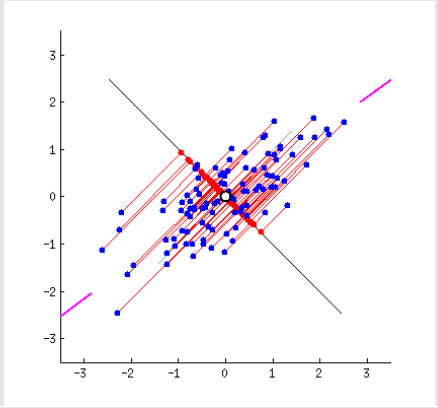


K-Nearest neighbours

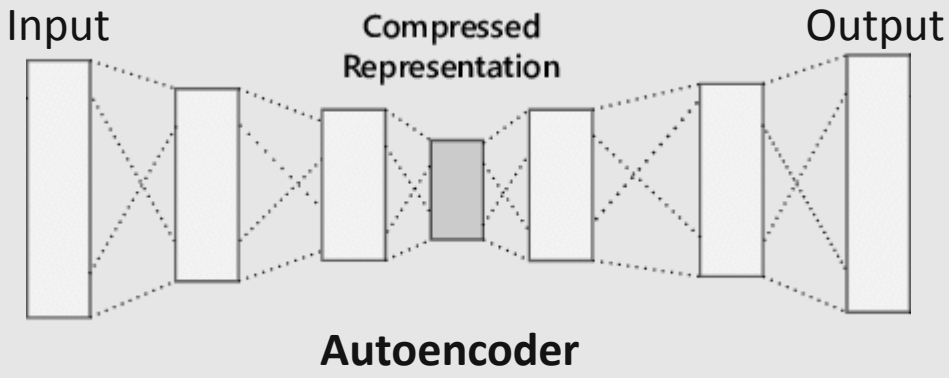
Dimensionality Reduction

- Definition

Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables.



Principal Component Analysis (PCA)



Autoencoder

Examples

High Energy Physics:

- Run classification – data preparation
- Fast simulation with Variational Autoencoders – data representation
- Search for similar events – data representation

Outside:

- Genetics – data representation and visualisation

Frequent sets and Association Rules

- Definition

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

ID	Items
1	Coke, Bread, Diaper, Milk
2	Beer, Bread, Diaper, Milk, Mustard
3	Coke, Bread, Diaper, Milk
4	Beer, Bread, Diaper, Milk, Chips
5	Beer, Coke, Diaper, Milk
6	Beer, Diaper, Bread, Eggs
7	Bread, Milk, Chips, Mustard

Rules mining →

Bread, Diaper -> Beer
(80% confidence, 60% support)
Milk, Diaper -> Coke
(60% confidence, 40% support)

Anomaly detection with ML

- Classification of anomalies (needed: labelled dataset)
- Regression of one value which may indicate anomalies (needed: dataset with known values)
- Clustering of unknown data and searching for outliers (needed: noisy data)
- Dimensionality reduction for sparse data representation and searching of outliers (needed: high dimensional data)

Model evaluation – binary classification

ID	Class	Prob.	Pred
1	0	0.1	0
2	0	0.1	0
3	1	0.2	0
4	0	0.3	0
5	0	0.4	0
6	1	0.4	0
7	1	0.6	1
8	0	0.7	1
9	1	0.7	1
10	1	0.8	1

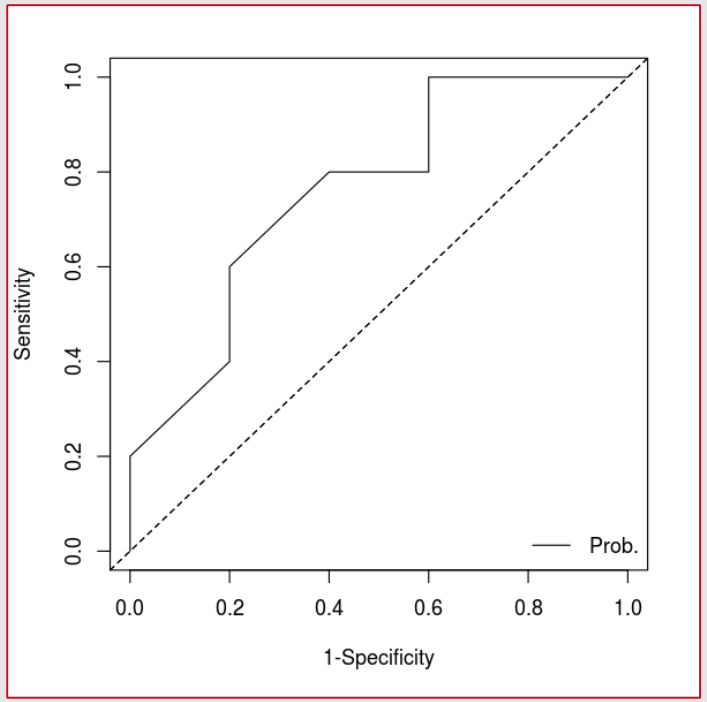
Test set + predictions

		Class	
		1	0
Prediction	1	TP	FP
	0	FN	TN

		Class	
		1	0
Prediction	1	3	1
	0	2	4

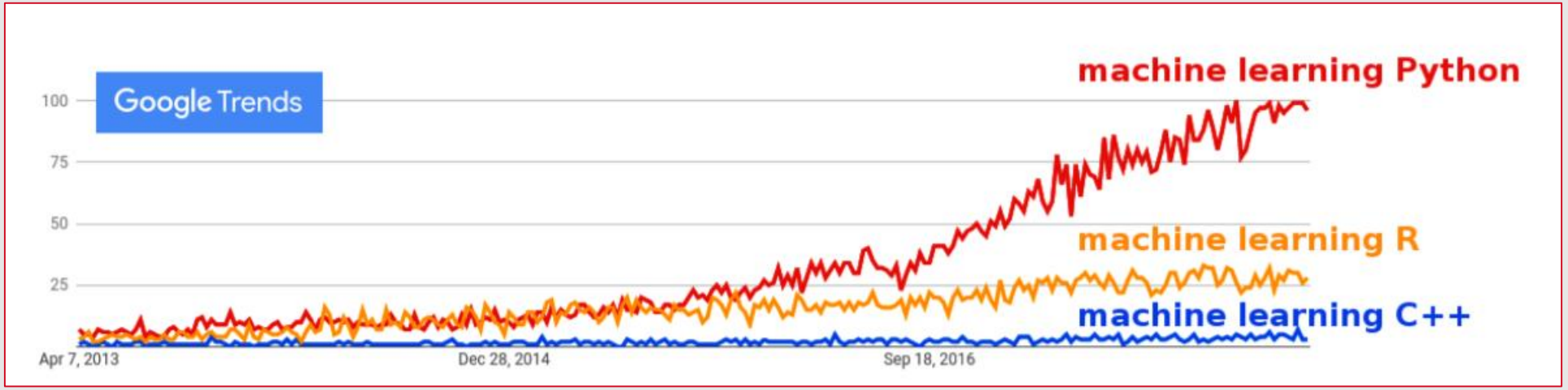
- Sensitivity, recall, True Positive Rate = $\frac{TP}{P} = 0.6$
- Specificity, True Negative Rate = $\frac{TN}{N} = 0.8$
- Precision = $\frac{TP}{TP+FP} = 0.75$
- Accuracy = $\frac{TP+TN}{TP+FP+FN+TN} = 0.7$

ROC Curve



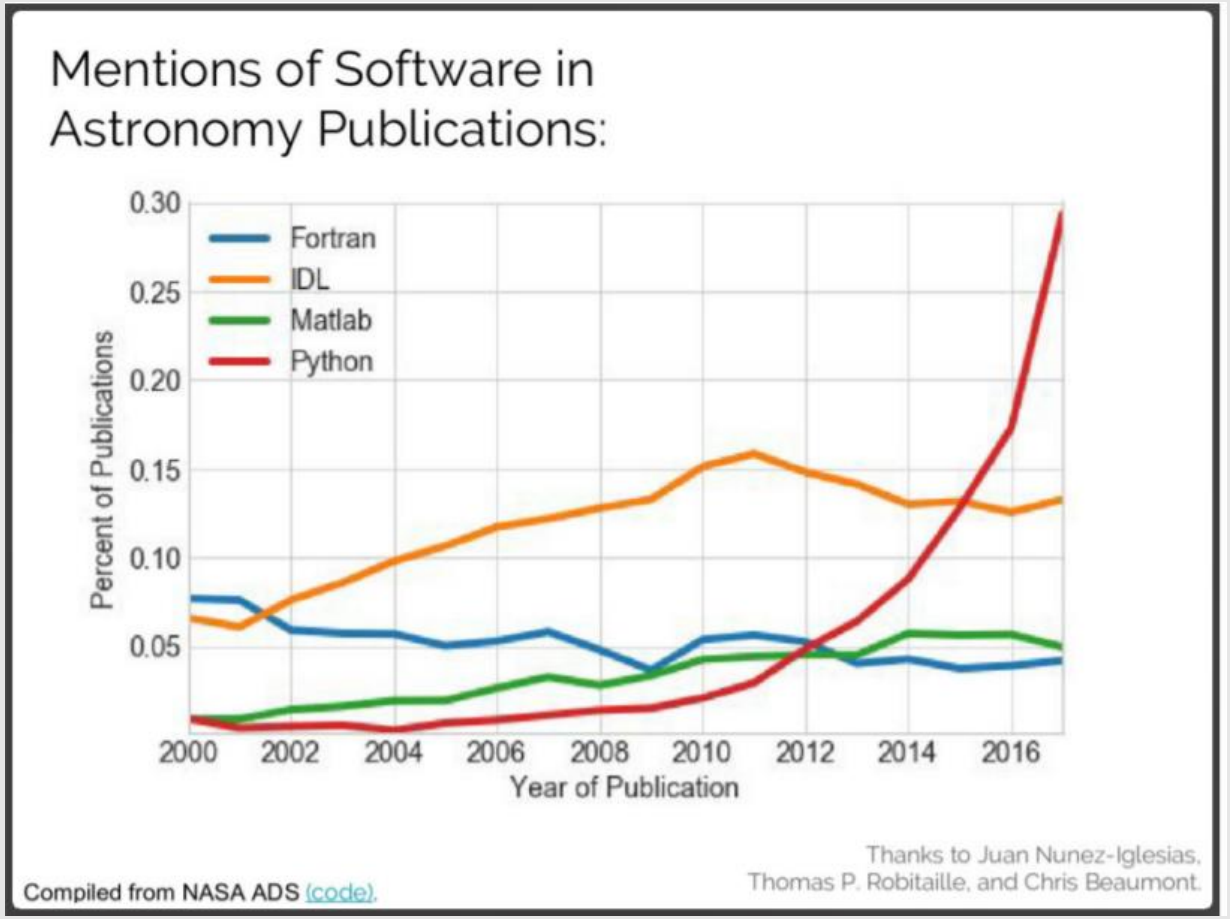
How to do ML?

What is trendy in ML?



pivarski-bigdata-software

It is possible!



Don't turn Your back to the community

AliROOT 70 000 commits 136 contributors

Most common ML libraries on github:

Python:

Sci-kit learn 23 000 commits 1200 contributors

Multilingual (Python/R/Java):

Tensorflow 45 000 commits 1750 contributors

Keras 5000 commits 750 contributors

PyTorch 15 000 commits 850 contributors

Spark 23 000 commits 1300 contributors

(Not so) common knowledge

- Up to 70% of the Data Scientist/ML researcher work is spent on data preparation
- „Garbage in garbage out”
- You will never have a perfect model
- Neural Networks are not the Holy Grall