23 – 25 July 2018

Lviv Data Science Summer School 2018

Machine Learning for Medical Applications:

Data Exploration

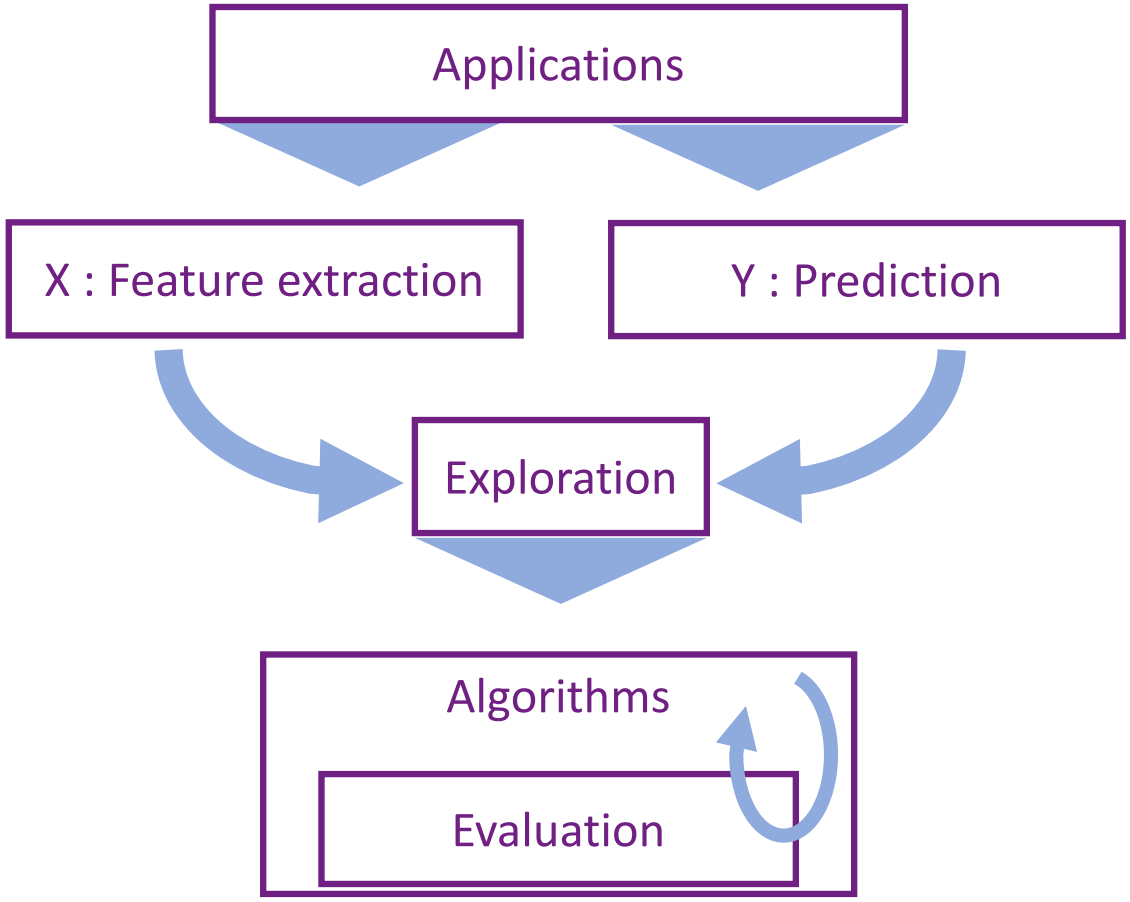
Igor Koval

PhD Student in Applied Mathematics

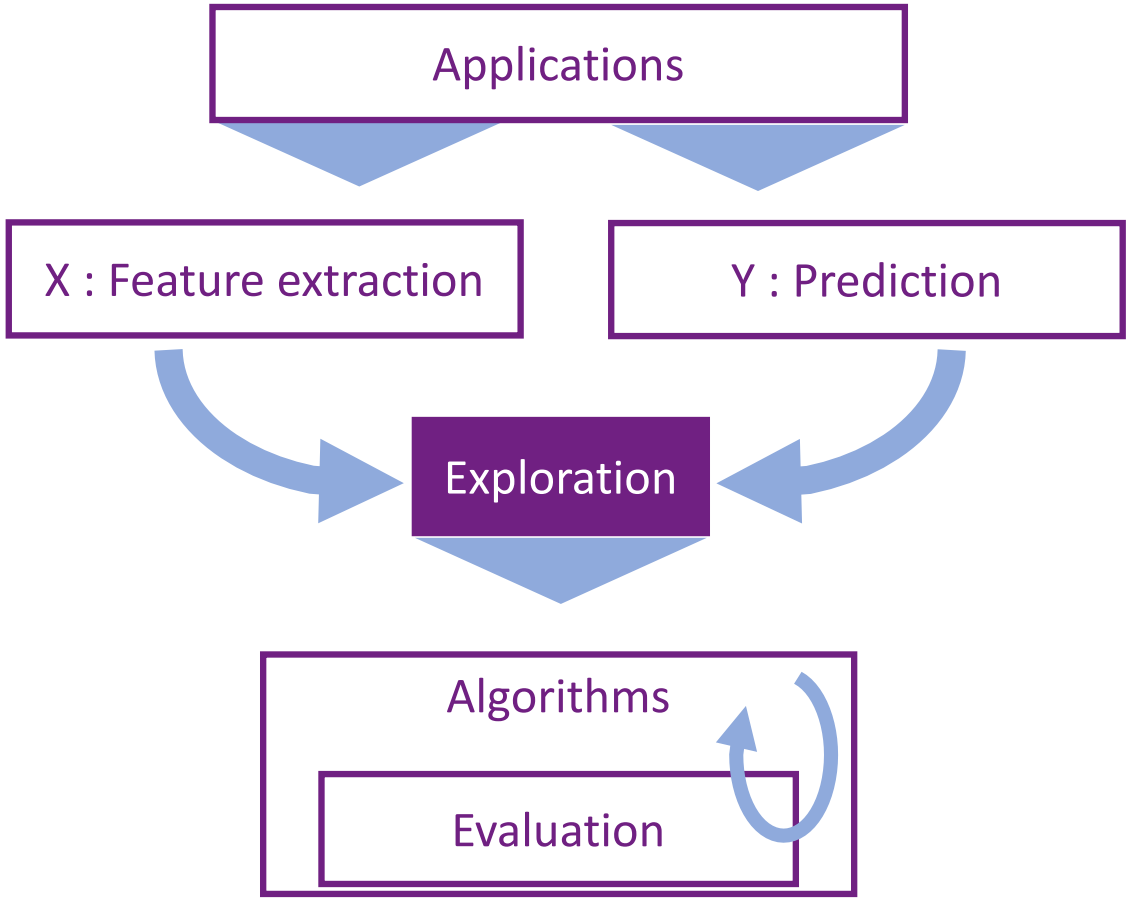
Brain and Spine Institute, Pitié Salpétrière Hospital, Paris, France
& Mathematical Laboratory of Ecole Polytechnique

igor.koval@icm-institute.org

Plan


Applications
X : Feature extraction
Y : Prediction
Exploration
Algorithms
Evaluation

Plan


Applications
X : Feature extraction
Y : Prediction
Exploration
Algorithms
Evaluation

Practical session 2

ARAMIS LAB
BRAIN DATA SCIENCE

UKRAINIAN CATHOLIC UNIVERSITY
FACULTY OF APPLIED SCIENCES

Database:

ADNI, preprocessed by UCL for an international challenge (TADPOLE) of Alzheimer's Disease prediction

Prediction:

Alzheimer's Disease

Features:

Biomarkers
Structural imaging
Functional imaging
Others

Objectives:

Process a raw database into a ML-like database

- Part 1 : Global overview
- Part 2 : Types of data
- Part 3 : Normalization
- Part 4 : Categorical data
- Part 5 : Dates
- Part 6 : Useless features
- Part 7 : Missing values & Outliers
- Part 8 : Text
- Part 9 : Unbalanced classes
- Part 10 : Dimensionality reduction
- Part 11 : Further exploration

Categorical data, Text, Dates, ...

ARAMIS LAB BRAIN DATA SCIENCE

UKRAINIAN CATHOLIC UNIVERSITY | FACULTY OF APPLIED SCIENCES

- Some categorical data are **ORDERED** or **UNORDERED**

ORDERED	UNORDERED
- Educational level - Level of pollution - Allele number	- Socio-demographic features - Gender - Allele number
Ordered labels	One hot encoders

- Sometimes, it is useful to turn a continuous feature into a categorical ordered feature : for instance, Huntington Disease
- Int/Float values does not mean they are "numbers" : it can be an index (country list, phone code, ...)
- The distance between 0 and 1 is not necessarily the distance between 1 and 2
- Dates : In general, the learning algorithm cannot deal with them
  -> Transform into a continuous feature
- Text : If not categories but free text : Fully dedicated topic -> NLP

Normalization, Useless features, …

ARAMIS LAB BRAIN DATA SCIENCE

UKRAINIAN CATHOLIC UNIVERSITY | FACULTY OF APPLIED SCIENCES

- Are the variables always comparable?

  For instance, what does it mean to have a large ventricule or a small hippocampus?

  It needs to be normalized (by the heart volume or the brain volume)

- Some features present only one instantiation

  -> They don't provide any information : erase them

- It can be interesting to try "co-features" : +, -, *, /, f(x1, x2), …

Missing values & Outliers

ARAMIS LAB BRAIN DATA SCIENCE

UKRAINIAN CATHOLIC UNIVERSITY | FACULTY OF APPLIED SCIENCES

Outliers

- Fake or not : they may change the algorithm consistency
- Typo
- Person that do not fill correctly the data

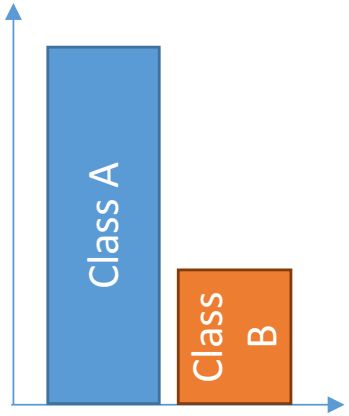
There are specific algorithms to spot them

A important characteristic of each algorithm is its robustness to outliers

Missing Values

- Given an upper ratio of missing values, it is possible to remove the feature
- Usually : Mean, average, ...
- kNN to find the closest datapoint(s)
- Regress the variable with the other one
- Sometimes (Particularly in medecine), the exam is not passed because its useless : default value

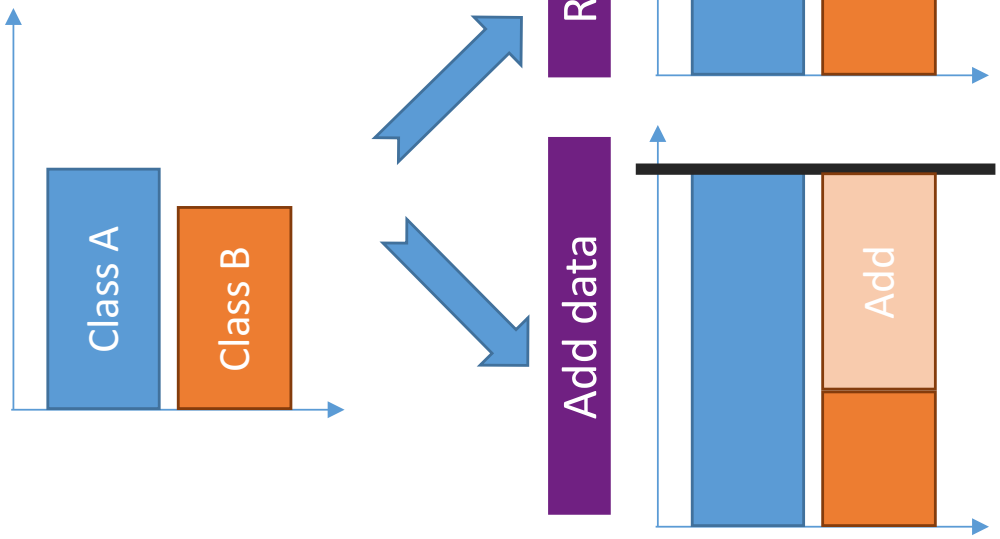
Unbalanced classes



The train set should have the same distribution as the test set

BUT

The algorithm learns only how to evaluate well the predominant class

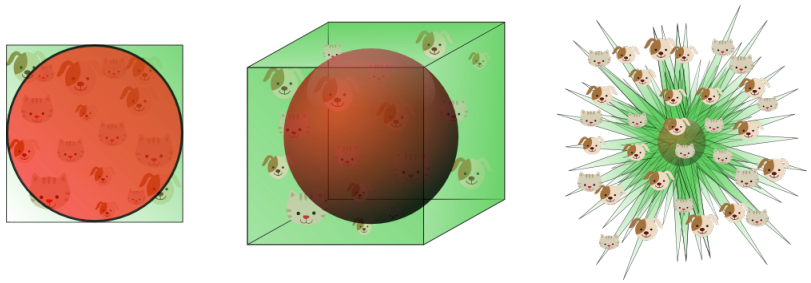


	Advantages	Drawbacks
Remove data	Easy to remove	Less data in the train set
Add data	All the data are used	How to add some data? - New dataset - "Bootstraping" methods - Data augmentation

Dimensionality reduction

Curse of dimensionality : $p \gg n$

Why is it important?



1. Minimize the reconstruction error

$$\Phi: X \mapsto X'$$
$$x \mapsto \Phi(X)$$

Minimize $|x - \Phi^{-1}(\Phi(x))|$

2. Distance preservation

$$\Phi: X \mapsto X'$$
$$x \mapsto \Phi(X) = x'$$

Minimize $\left|d(x_i, x_j) - d'(\mathrm{x}'_\mathrm{i}, \mathrm{x}'_\mathrm{j})\right|$

Dimensionality reduction

ARAMIS LAB BRAIN DATA SCIENCE

UKRAINIAN CATHOLIC UNIVERSITY | FACULTY OF APPLIED SCIENCES

Minimize the reconstruction error

- PCA - Principal Component Analysis
- Kernel PCA
- Dictionary learning
- ICA – Independent Component Analysis
- LDA – Latent Dirichlet Allocation
- NMF – Non-negative Matrix Factorization
- ...

Distance Preservation

- Locally Linear Embedding
- Multidimensional Scaling
- ISOMAP
- t-SNE (StochasticNeightbor Embedding)
- ...

Feature Importance

- Random Forest (Decision trees in general)
- Regressions (with normalized features)

Manifold Learning

The data belong to a parametric subspace that can be learnt