Ensemble Methods
TD - Basics in Machine Learning

M2 Computer Science - MALIA

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Abstract

The aim of this tutorial is to determine your knowledge of Machine Learning, particularly in practical terms. Is the process of learning a model known/mastered? But also to see your different reflexes when faced with datasets that may have different characteristics.

Study of the datasets

You can find different datasets at the following link:

Download datasets.

You can use them throughout this first session. You willl also find Python code for formatting the various datasets so that they're ready to use.

Preparing the data

We can carry out the various steps:

- look at the previous code to identify the various stages of data loading and preparation
- detect missing values and perform imputation
- identify dataset characteristics (size, dimension, problem type, class ratio)

Experiments

Once the data is ready, you will try to implement the learning process using your current knowledge in the field.

- learning process
- choose the appropriate performance measure
- implementation of learning methods on data
- use different algorithms and try to make a **fair** comparison in terms of both performance and computation time
- choose the appropriate performance measure regarding the type data.
- ...

At the end of this section, you should have a table (maybe several) summarizing the results obtained using the different methods on the datasets used for this first task. This (.ese) table(s) shall have the following shape:

Dataset	Algo 1	Algo 2	Algo 3	Algo 4	Algo 5	Algo 6	...
Data 1							
Data 2							
Data 3							
Data 4							
Data 5							
...							
Mean value							

Some advices:

- Do not forget to perform a cross-validation at each step.
- Think if it is relevant to perform only one test.
- Try to measure the stability/variability of each method.
- Compare algorithms that are comparable! Do you think it is relevant to compare linear methods with linear ones?
- Try to study the set of dataset and see if it can be interesting to divide them into two groups...
- Do not hesitate to adapt the strategy if you are facing a particular type of problem
- Do not forget to normalize your data if you think it is relevant!