# Practical-Sessions-LELEC2870

August 2020

## Notations and vocabulary

- a scalar: A number represented with an upper-case letter when defining bounds on a series, or a lower-case one otherwise e.g. $x$ in $1..X$
- a vector: A 1-dimensional array that will always be written with a **bold lower-case letter**
- a matrix: A 2-dimensional array that will always be written with a **bold upper-case letter**
- a tensor: A N-dimensional array that will always be written with a **bold upper-case letter**. While it uses the same notation as the matrix, it should always be clear which one is meant.
- an iteration: A time-measure for *iterative* algorithms. It denotes a pass over a part of the training dataset.
- an epoch: A time-measure for *iterative* algorithms. It denotes *one* pass over the complete training dataset. An epoch is composed of at least 1 iteration, but most of the time of multiple iterations especially when large datasets are involved.

# 1 Session 1

## Objectives

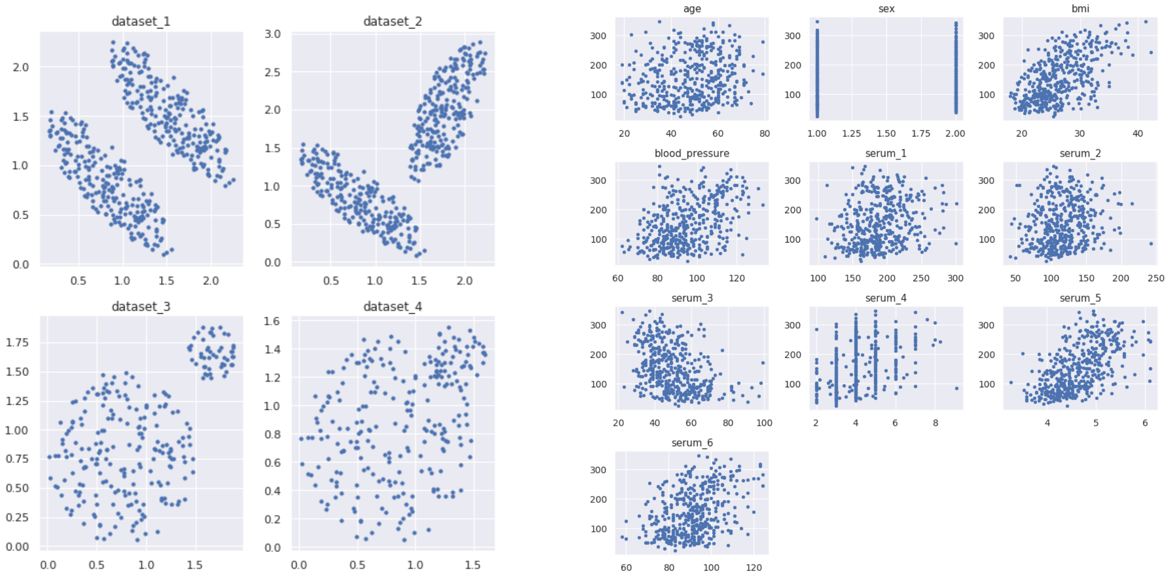
In this first exercise session, you will implement 2 variations on vector quantization techniques (K-means and frequency sensitive learning) as well as linear regression.
We provide you two datasets for the two topics covered in this session. Those datasets can be found on the Moodle page of this course.

### BEFORE the session

**Load** the first two datasets and **plot** them.

1. For dataset 1 (**VQ**), you will see that groups of points are clearly visible, and your objective in this session will be to learn these groups automatically. There are 4 (synthetic) datasets with 2 features every time.

2. For dataset 2 (**LR**), your objective will be to learn the linear relation between the target and the inputs. The target variable ('t' in the dataframe) represent the blood sugar levels. The features ('X' in the dataframe) represent: the age (feature 1), the sex(feature 2), the body mass index (feature 3) and the blood pressure (feature 4) of the patient, as well as the result of several serum measurements (features 5 to 10)



(a) data1

(b) data2

Figure 1: (a) The scatterplots of the datasets for the clustering exercise (b) The blood sugar levels in function of the different features for the linear regression exercise

- **Vetor Quantization (VQ)**. It consists in reducing the size of a dataset while minimizing the loss of information, by summarising it with a set of centroids. An archive containing all necessary data and and a python script for visualization are on the Moodle website.

- **Linear Regression (LR)**. Linear methods often allow to obtain good results in regression, when the relationship between the features (or inputs) and the target (or output) is not too non-linear. You will implement linear regression for a set of features that you will select using simple linear methods.