MAP534 - Machine Learning I

Sylvain Le Corff

November 9, 2020

Dimension reduction

Principal component analysis

In Machine Learnning, an **input** data will be written $X \in R^p$: it is called an individual, or a sample. This input X is associated with an **output** Y. From there, machine learning can solve two types of problems:

1. **Classification**: $Y \in \{1, \ldots, M\}$. In this case, Y is the label of the group to which Xbelongs. Your objectif is to build a map $f : R^p \longrightarrow \{1, ..., M\}$ which associates any individual to the good category.

2. **Regression**: the output $Y \in R^p$. Usually we consider models of the form $Y = f(X) + \epsilon$ where f is the unknown function to be estimated and ϵ is the random noice, independent of X.

We will consider different models and different algorithms to solve regression and classification problems.

From there, two approaches are possible. In **supervised learning**, we have a set $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ of observations of the experiment. In this case, the "best" function f is built using this dataset.

Before going into the details of machine learning algorithms, we discuss some preprocessing of the data to deal with high dimensionality.