Mixture models in RKHS

Supervisor: Alain Celisse, Assistant Professor
Université Lille 1 – Modal INRIA
celisse@math.univ-lille1.fr

Context and motivation

In industry, medecine, and biology, an object of interest (the patient for instance) is often described by descriptors (variables) of several different natures such as phenotypical data (eye color, sex, height,...), and also experimental outcomes (curves observed along time, mRNA expression data,...). An important challenge in these numerous areas is to enable dealing with all these data together. Indeed successively considering only part of them (for instance quantitative data) as it is the case nowadays induces a dramatic loss of information (and money).

Kernel methods (and associated RKHSs) are a convenient means to overcome this challenge through combining several kernels. A bunch of approaches have been developed in the context of kernels. However most of them are predictive ones, that is they only predict a status for instance without providing any helpful insight on the phenomenon of interest. Conversely very few generative approaches have been proposed in that setting, essentially in the machine learning community. By definition, generative approaches are based on a (parametrized) probabilistic model from which data are assumed to be drawn. The statistical question is then estimating the model parameters, which provides experimenters with a more informative understanding of the mecanism they are infering.

The purpose of this intern is to develop such a flexible and meaningful generative model to make inference in RKHS structures.

Outline

At the beginning (depending on the student skills), the student will spend some time on kernel methods and RKHS structures. The first step of the study will consist in defining a Gaussian process on the given RKHS. The question of the parameters estimation for this process on true data will be of interest. Inference will be made to assess how reliable Gaussian processes are to modelize the general behaviour of data in RKHS.

The second step will be devoted to extend the previous strategy to the case of a mixture of Gaussian processes, which is more flexible but more difficult to deal with at least in terms of parameter estimation.

The last step is to develop a strategy to perform clustering, that is unsupervised classification, of true data.

Prerequisites

Basic concepts in Statistics are required. Preliminary strong knowledge on kernels is not required even though it could help. Skills on programming with for instance Matlab or R will be helpful.