9

Functional Nonparametric Unsupervised Classification

Most of the sections presented in this chapter are readable by a very large public. Methodological, practical and computational aspects take a large share whereas theoretical developments are presented in a self-contained section. In fact, the spirit of this chapter is close to the spirit of exploratory data analysis. More precisely, when an unsupervised classification is performed, the statistician or more generally the user does not know how to validate the obtained partition. Only some additional information collected after the analysis can confirm or refute the results. So, according to their experience, the statistician will try to propose more pertinent answers to the classification problem. This is exactly what we try to do here but in a new field which concerns functional data. Heuristics and theoretical aspects are developed in a complementary way which produces an original nonparametric classification method for functional data. Note that this chapter is quite different from the other ones. From a theoretical point of view, we propose a classification method which involves the mode of the distribution of a functional random variable. This leads us to deal with the density of functional random variables and new problems emerge as soon as we focus on the asymptotic behaviour of the estimator of the "functional" mode. First results allow us to point out difficulties linked with the infinite dimensional setting. From a practical point of view, it is much more simple to solve the prediction problem than the unsupervised classification one. For all these reasons, this chapter is certainly more open than the other ones and will certainly deserve future investigation.

9.1 Introduction and Problematic

Unsupervised classification is an important domain of statistics with many applications in various fields. The aim of this chapter is to propose a nonparametric way to classify a sample of functional data into homogeneous groups. The main difference with discrimination problems (see Chapter 8) is that the

group structure is unknown (we do not have any observations of some categorical response), and this makes such a statistical study more delicate. The general idea is to build a descending hierarchical method which combines functional features of the data with a nonparametric approach. More precisely, the proposed methodology performs iteratively splitting into less and less heterogeneous groups. This forces us to define what means heterogeneity for a class of functional objects. To this end, we measure the closeness between some centrality features of the distribution. The great interest of the nonparametric modelling consists in estimating such characteristics without specifying the probability distribution of the functional variable. This is a required point because the distribution generating the sample of functional data is supposedly unknown (free-distribution modelling) and even if one would specify the distribution, it would be impossible to check it. Concerning the way to split the functional data, we make a feedback between practical aspects and recent theoretical advances, which allows us to introduce a partitioning based on small ball probabilities considerations.

Concerning the organization of this chapter, we voluntary insist on methodological and practical aspects as in the discrimination chapter because most of the expectations in this domain are oriented towards the applications. However, some first asymptotical advances are given in a self-contained section in order to point out open problems in relation to the infinite dimensional feature, especially when uniform consistency is required. In addition, a feedback practice/theory will be emphasized in the proposed methodology. We start with Section 9.2 which presents functional versions of the usual notions like mean, median and mode. After that, Section 9.3 proposes a simple way to measure the heterogeneity of a sample of functional data by comparing the previous indices throughout a semi-metric. Once the heterogeneity index is defined, Section 9.4 describes a general descending hierarchical method based on a notion of gain or loss when a sample of functional data is partitioned. In particular, if we have to fix some smoothing parameters (which can appear in the estimator of the mode), an automatic selection procedure based on the maximization of the entropy is proposed. Section 9.5 presents a short case study which illustrates the easiness of both implementation and use of such a nonparametric functional classification whereas its good behaviour is emphasized. Section 9.6 studies the asymptotical behavior of the kernel estimator of the mode. These theoretical developments are quite different from those detailed in Chapter 6 because a uniform-type consistency of the density estimator is necessary. Finally, this chapter ends with a bibliographical overview which places this work in the recent literature.

Before going on, we recall that the same notation are used as before. Thus, $\boldsymbol{\mathcal{X}}$ denotes a generic functional variable taking its values in the infinite dimensional semi-metric space (E, d). In addition, let $\mathcal{S} = \{\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n\}$ be a sample of n variables identically and independently distributed as $\boldsymbol{\mathcal{X}}$, let χ be a fixed element of E and let $\chi_1, \ldots, \chi_n$ be the functional dataset associated with the functional sample $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_n$.