Review of the thesis

Apprentissage statistique non-supervisé : grande dimension et courbes principales

by Aurélie Fischer

The thesis presents the author's results on the theory and practice of non-supervised statistical learning. The main focus is on clustering of high-dimensional data and principal curves. The tools of modern statistical learning theory and model selection are used to develop numerous novel theoretical results. On the other hand, the methods are successfully implemented for simulation studies and for analyzing data in various applications.

Non-supervised learning, especially clustering and principal component analysis, are among the most widely used methods in scientific data analysis and it has had an enormous importance in a wide variety of applications. At the same time, analysis of high-dimensional data has become one of the most important challenges of modern statistics, and therefore the topic of the thesis is highly relevant.

The thesis presents an impressive wealth of interesting results in this important area. The basic mathematical tools are based on the modern theory of empirical processes and model selection. These are successfully applied in various problems of high-dimensional non-supervised learning.

Chapter 1 deals with the problem of clustering (a.k.a. vector quantization) of functional data. It is assumed that the data points take their values in a Banach space and distortion is measured by a Bregman divergence. This is a novel model that generalizes many previous approaches. General conditions are given for the existence of an optimal quantizer (i.e., one that minimizes the expected distortion measure) and consistency of empirical distortion minimization is established under very general conditions. Upper bounds for the rate of convergence are also offered. Many of the results previously known from the literature are recovered but thanks to deep functional-analytic techniques the author managed to take an important step forward.

The material of Chapter 2 is motivated by a concrete application of clustering of functional data in nuclear technology. A novel projection-based clustering technique is introduced and analyzed. The main theoretical result (Theorem 2.2.1) estimates the effect of projection. The effect of basis selection is also studied in several important cases. Finally, empirical results are presented for main motivating application. This chapter is a very nice

example of how deep mathematical tools can be used fruitfully in concrete complex statistical problems.

In Chapter 3 the choice of the number of groups in clustering is studied. This is one of the most fundamental and intriguing problems of the theory and practice of clustering. The author introduces a new technique, based on the "slope heuristics" of model selection of Birgé and Massart. The theoretical result of this section is a more-or-less straightforward application of general results of model-selection but the application is novel and the experimental results convincingly demonstrate that the new method works very well. There is a huge body of literature on this subject and the author does a very good job of discussing the relevant methods but perhaps the (in)stability results of Ben-David, Pál, von Luxburg, and others might also be mentioned in this context.

The second part of the thesis investigates principal curves. This very interesting generalization of principal component analysis has gained a lot of popularity in the recent years and various algorithms have been introduced for their construction. Previous theoretical results by Kégl, Krzyzak, Linder, and Zeger and Sandiliya and Kulkarni proved consistency of curves of bounded length and bounded turn, respectively. However, all previous methods contained two ad hoc parameters (the number of linear pieces of the constructed curve and its total length or turn) whose data-based calibration is essential for the performance. Chapter 2 of Part II of the present thesis offers a principled automatic way of selecting these parameters. Once again, the theory of model selection (in particular, results of Birgé and Massart) is a key ingredient of the theoretical analysis, together with careful bounding of appropriate Rademacher complexities and entropy integrals. The main theoretical results are oracle inequalities that guarantee a near optimal choice of the free parameters. Experimental results are also presented both for simulated data and for a very interesting application based on locations earthquakes. The unknown constants are specified by using the "slope heuristics". This chapter synthesizes model selection and empirical process theory with the theory of principal curves in a concise and convincing way and makes an important step in the direction of the practical applicability of the theory of principal curves. A minor remark about Theorem 2.2.2 is that the term proportional to $\sqrt{\ell}$ can be omitted as $\sqrt{\ell} \leq (1/2)(\sqrt{k}+\ell/\sqrt{k})$.

In summary, this is a very strong thesis containing a wealth of novel results that make important advance in both the theory and practice of unsupervised statistical learning. The material is presented in a remarkably coherent way and the entire thesis is written in a smooth and pleasant style which was a pleasure to read. The depth of the underlying mathematics,

the clarity of the ideas, the impressively thorough literature review, and the high quality of the presentation show the work of a mature researcher.

I warmly recommend that the Ph.D. degree be given to Aurélie Fischer.

Gábor Lugosi
ICREA and Pompeu Fabra University
Barcelona.