Statistics for Astrophysics: Clustering and Classification
D. Fraix-Burnet and S. Girard (eds)
EAS Publications Series, **77** *(2016) 1–2*

FOREWORD

This second session of the School of Statistics for Astrophysics has been attended by an increased number of people coming from accross the world, mostly post-docs and students. This confirms the success of the first session in 2013 and of the specificity of this school: lecturers are statisticians and half of the courses are devoted to practice. Statistical tools are not blackboxes, it is essential to know some mathematics behind the techniques. The practice is a complementary necessity to understand the possibilities and limitations of a particular technique.

The subject of this 2015 School that took place at l'École de Physique des Houches near Chamonix-Mont-Blanc in France, was clustering and classification.

Grouping objects obtained from the astronomical observations into distinct categories has always been a necessity imposed by their vast diversity. This is the case for stars, galaxies, asteroids, supernova, active galactic nuclei, gamma-ray bursts and many others. This clustering (unsupervised classification) is a prerequisite to any physical modelisation. In this purpose, astronomers have always used heuristic, simple and subjective techniques, based on only one or two parameters, most often with the help of a visual examination. Consequently, for a given type of astrophysical objects, a large number of classifications may exist, sometimes overlapping each other.

However, this traditional practice cannot be transposed in the era of huge databases. The primary role of a classification is to describe the diversity with a simple scheme that also helps understand the relationships between the classes. Astronomers are now obliged to consider using objective and multivariate clustering methods that allow for automatic (supervised) classifications.

During the School and in this book, some general concepts of statistics, classification, clustering and an introduction to the environment `R` are provided in three chapters at the beginning to help follow the more in depth lectures. The first such lecture covers the mixture model approaches which are quite powerful both in unsupervised and supervised learning. This chapter also introduces many other more classical methods like discriminant analyses or the Expectation-Maximization algorithm.

The following chapter is devoted to the problem of the big data sets. High-dimensional data means high number of objects (observations) and/or high

DOI: 10.1051/eas/1677001

number of variables. Popular model-based techniques for clustering, renowned for their probabilistic foundations and their flexibility, suffer from the well-known curse of dimensionality in these cases. This chapter presents a comprehensive review of the recent approaches that overcome these drawbacks.

The next chapter is devoted to the clustering of variables. Astronomers are well acquainted with the Principal Component Analysis which aims at reducing the number of variables by grouping together those that are well correlated. This chapter presents other possibilities to cluster variables into less numerous synthetic variables.

Kernel methods with an emphasis on the Support Vector Machine techniques, for supervised classification, are then described. Originally, the Support Vector Machine techniques are designed to solve binary problems where the class labels can only take two values. Fortunately, various approaches have been proposed to cope with several classes.

Finally, the last two lectures cover a difficult topic rather unfamiliar to astronomers: this is classification using graphical approaches. Indeed, one such technique has been used for some thirty years to cluster the galaxies onto the cosmic web using two spatial coordinates and the redshift: this is the Minimum Spanning Tree. This is a very simple and intuitive technique, though very efficient in this case. The first lecture provides some feeling of the fascinating graph theory, by explaining how the probability theory can be represented graphically. While its use for classification in astronomy has to our knowledge never been attempted, it is widespread in many other disciplines, and the given illustrations on image segmentation should convince of its potential power.

Tree-like graphs, heavily used in bioinformatics, are presented in the last chapter. The branches depict the relationships between the classes, and can be hypothesized or interpreted as evolutionary relationships. The associated methods are called phylogenetic approaches and are mathematically not limited to living organisms at all. Several astrophysical applications of the Maximum Parsimony (or cladistics) have been published in the last ten years, and the chapter presents other such techniques.

The reader can find on the school website `http://stat4astro2015.sciencesconf.org/` the `R` codes and the necessary data sets corresponding to the illustrations and exercices described in this book.

Some mathematics given in the chapters may frighten astronomers, but statistics being an active research field, our aim is to raise the awareness of the modern techniques. They solve many problems encountered with the more classical methods that are easier to understand with the mathematical background of astronomers. Statistical clustering and classification is a huge domain, we hope that this book will provide the keys for astronomers to collaborate with statisticians.

Didier Fraix-Burnet
Stéphane Girard