Robust statistics, Data Mining & Machine Learning in Astronomy

SUMMARY.

This METEOR provides an introduction to advanced concepts and methods for the analysis of data in Astrophysics. It covers a set of robust statistical analyses, data mining and machine learning methods that are used by astronomers for analyzing data from ground or space instruments. These tools are useful in any topics of Astrophysics and illustrations examples will range from cosmological studies to exoplanets, with data from MUSE and SPHERE instruments @ VLT, KEPLER mission and others. The METEOR intends to bridge astronomers' and statisticians' vocabulary and practice, with emphasis on robust (versatile and general) data analysis methods.

OBJECTIVES

- One main objective of this METEOR is to train the students to learn in autonomy, to identify what can help them to progress, to identify and correct their errors, to define a project related to a problem of their interest and to solve it.
- The students will understand and practice general methodological tools in robust statistics, data mining and machine learning.
- The students will learn to identify specific problems posed by the analysis of astrophysical data sets and to pose the problems in statistical terms.
- The students will learn to find solutions to specific statistical problems from the literature.
- The students will learn to implement the solutions in the form of Python programs by using relevant packages or by developping their own code.

PREREQUISITES

Preparatory MAUCA courses (not mandatory):

- Statistical methods
- General Astrophysics
- Numerical methods
- Signal & Image processing

Interested students are encouraged to contact the supervisor and previous students who have followed this METEOR (or the METEOR exoplanet detection, with the same supervisor) to have an idea of how the METEOR works.

THEORY

by D. Mary

This METEOR contains four main parts (whose order and contents can change depending on when the METEOR takes places, M1 or M2):

1. The first part deals with classical statistical inference. It focuses on differences between frequentist and Bayesian inference, on Maximum Likelihood estimation, goodness-of-fit and model selection techniques, the Expectation-Maximization algorithm, how to estimate confidence intervals from data with the jackknife and the bootstrap. It also provides an introduction to hypothesis testing.

2. The second part provides a training to Bayesian inference. It deals with Bayesian priors, the quantification of uncertainty in Bayesian

approaches, Bayesian model selection, non uniform priors, numerical methods for complex problems (Markov Chain Monte Carlo) and hierarchical Bayesian modelling.

3. The third part deals with Data mining and Machine learning, including density estimation techniques, finding clusters in data and the study of correlation functions for analysing the underlying structure and scale of data sets.

4. The last part deals with dimensionality reduction techniques, which are increasingly useful owing the large size of current data sets. This part introduces the Principal Component Analysis (PCA) and also focuses on regression problems, including linear regression, PCA regression and Gaussian process.

APPLICATIONS

by D. Mary

- During the METEOR, the students will select and present one recent paper of the literature that deals with one Machine Learning techniques seen during the theoretical part.

- The students will (with the help of the supervisor) define their application project according to their interest. They may choose to study deeply one particular technique and to apply it to astrophysical data. Alternatively, they can choose to focus on astrophysical problems and data sets that they encountered during a previous METEOR. A third possibility is to be proposed a project related to the current research of the supervisor (e.g. exoplanets studies, with data from ground based instrument like SPHERE or space mission like Kepler, or galaxies studies with data from the MUSE instrument).

- The application examples of the theoretical part and the project part of the METEOR provide an intensive training to Python and to dedicated data mining & machine learning packages.

MAIN PROGRESSION STEPS

- During the whole duration of the METEOR, each student has a personal channel on Discord allowing easy connection with the supervisor outside the scheduled meeting slots. A general channel serves also as a forum for general infos/questions/hints.

- First half of the period (possibly more): the students learn theory. They are requested to work on the lecture notes on their own, with regular discussions planned with the supervisors to answer their questions. They do the theoretical and numerical exercises proposed in the lecture notes document and they post them on the fly on their personal channel.

- @ mid METEOR, the students identify a topic of the lecture they are mostly interest in, select and present one paper in Astrophysics that uses this technique. The students choose the topics of their project and the astrophysical application and data set. They also define interesting scientific questions related to the selected topics and data, and the main objectives of their projects.The supervisor helps the student to ensure that the project's objectives are reachable and the questions relevant.

- Second half of the period : the students work on their research project.

- Last week : last results and preparation of the final oral presentation.

EVALUATION

- Regular quizz, and two written exams (2h, typically during weeks 2 and 4) on the theoretical part. The average of the three marks provides the mark "Theory" (30% of the total mark).

- Permanent evaluation of the numerical project (mark called "P") according to the first objective (autonomy !).

- Permanent evaluation of the numerical exercises posted on line (mark "E").

- Oral presentation of statistical / data mining / machine learning techniques from the literature (mark "O"). Cross evaluation by the other students according to the same criteria grid that will be used by the jury for the presentations in the end of the METEOR.

- The final mark for the "Project part" is $0.25 \times O + 0.25 \times E + 0.5 \times P$.

- Final evaluation during the global oral presentation(40% of the total mark).

BIBLIOGRAPHY & RESSOURCES

- On-line lecture notes, slides,, homeworks, criteria evaluation grid, data, solution codes.

- *Statistics, Data Mining & Machine Learining in Astronomy*, Princeton Series in Modern Observational Astronomy, Second Edition, 2020

- *Modern Statistical Methods for Astronomy*, Cambridge University Press, 2012

- *Computer age large scale inference*, Cambridge University Press, 2019

CONTACT

☎ +33492076384 (D. Mary)
✉ david.mary@oca.eu