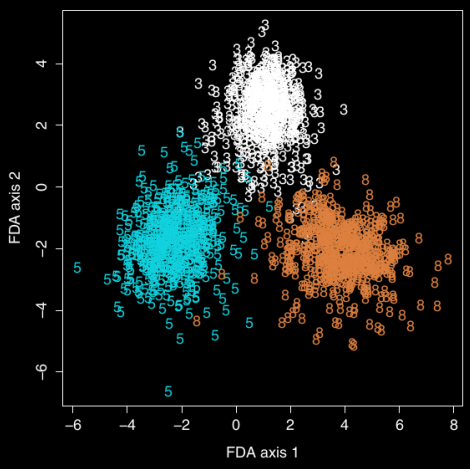
Large-scale classification of high-dimensional data



SUMMARY.
Most fields of astrophysics have to deal with very large data sets, with millions to billions of sources, each characterized by tens to hundreds of parameters. The ever increasing number of large sky surveys (Gaia, LSST, WEAVE, SKA) is pushing astronomy toward Big Data.
To extract relevant information from this wealth of data, modern data-science methods must be applied. The present METEOR focuses on the methods of clustering and classification, applicable to any research field. It combines theoretical knowledge on dimensionality reduction, normalization, clusterization, with practical experience with real data (visible and near-infrared spectra of asteroids).

OBJECTIVES

- Acquire fundamental and practical knowledge on supervised and unsupervised learning.
- Extract relevant information from articles. Develop codes in Python and R using open resources such as scikit-learn or TensorFlow. Classify large samples into coherent groups.

PREREQUISITES

Numerical methods, Signal/image processing, Maths/Stat

THEORY

by P.-A. Mattei, C. Bouveyron & B. Carry

The theoretical part of the METEOR covers both fundamental knowledge on machine learning and asteroid orbital and spectral properties.

- Supervised learning: logistic regression, linear discriminant analysis, deep discriminative models.
- Unsupervised machine learning: Probabilistic PCA (and other dimensionality reduction methods), clustering (from k-means to Gaussian mixture models).
- Dealing with dirty data: missing values, normalisation.
- Distribution of asteroids. Compositions and classification.

APPLICATIONS

by P.-A. Mattei & B. Carry

The project reproduces all the steps used nowadays to extract relevant information from a large corpus of data. The students will

- supervised and unsupervised classification of classical machine-learning data sets as well as orbital/spectral asteroids data
- missing data imputation and dimensionality reduction of asteroids spectra

MAIN PROGRESSION STEPS

- Once a week: theoretical courses (exam at end term).
- First half: supervised and unsupervised machine learning on low-dimensional data (classical machine-learning data sets and orbital data set)
- Second half: dealing with high-dimensional dirty data (asteroids spectra)
- Last week: preparation of the final oral presentation.

EVALUATION

- Written examination (40%), project (40%), and commentary on an article (20%).

BIBLIOGRAPHY & RESSOURCES

- Review on the history of our solar system from asteroids composition: DeMeo & Carry (2013, 2014).
- Chapters on spectroscopy, space weathering from Asteroids IV book.
- Chapters 1-4 and 8 of Bouveyron et al. (2019).

- Chapters 1,2,4,9, and 12 of Bishop (2006).

CONTACT

☎ +33492387575 (P.-A. Mattei)
✉ pierre-alexandre.mattei@inria.fr

☎ +33492003964 (B. Carry)
✉ benoit.carry@oca.eu