Horizon 2020
European Union Funding
for Research & Innovation

Machine Learning in Astronomy

Begoña Ascaso

PhD in Astrophysics, Data Scientist

APC, Astroparticule et Cosmologie, Université Paris Diderot, CNRS/IN2P3 (France). E-mail: bego.ascaso.work@gmail.com

Many of the problems faced in Astronomy are not too different to those found in the world of Data Science. Often, Astronomers need to classify galaxies into different morphological types, predict the distance of a galaxy or its composition based on indirect data, detect emergent structures, etc. I present here three examples of astronomical problems and associated algorithms and results using Machine Learning (ML) techniques. Many other problems analyzed with ML can be found in the literature and a non-exhaustive list is also shown.

FINDING GALAXY CLUSTERS: CLUSTERING TECHNIQUES

Problem: Given a galaxy catalogue, find the galaxy clusters, the largest structures in the Universe containing tens to hundreds of galaxies.

Approach: A technique called the Bayesian Cluster Finder[1] (BCF, ***Ascaso et al. 2012, 2014, 2015***), has been developed to detect galaxy clusters. The technique first computes Bayesian probabilities based on the galaxy luminosity, density and photometric redshift combined with a galaxy cluster prior that accounts for colour-magnitude relations and brightest cluster galaxy-redshift relation. Then, a modification of the k-means clustering algorithm is applied to the positions and probabilities to find the centroids. The method is able to determine the position, redshift and richness of the cluster.

Results: The method has been successfully applied to a number of different surveys, obtaining large samples of galaxy clusters and groups down do different mass limits and redshift ranges (see Fig. 1). The algorithms is usually trained and tested with simulations.

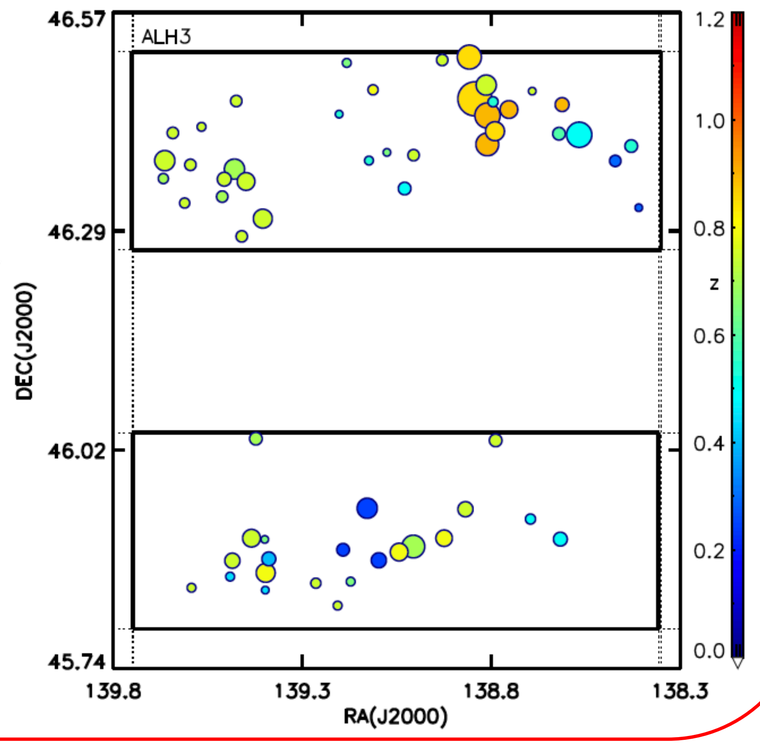


Fig 1. Galaxy cluster detections in 0.5 deg² in the ALHAMBRA survey (Ascaso et al. 2015). The size of the circle represents its mass, and the colour its redshift (~distance).

[1] http://bascaso.net46.net/research.html

CLASSIFYING GALAXY MORPHOLOGY: SVM TECHNIQUES

Problem: Given a particular galaxy, at a particular redshift, classify it into at least four morphological types.

Approach: A Support Vector Machines method, GalSVM[2] (*Huertas-Company 2008*) has been implemented to create a non-linear boundaries between different galaxy parameters, particularly the concentration and asymmetry.

Results: The method has been applied to different works (*Huertas-Company 2008, 2009, 2011*), obtaining <20% error in the completeness limits for galaxies brighter than K<22 (see Fig. 2).

Fig 2: Sample of galaxies in the SDSS DR7 survey, together with its morphological classification) as given by the GalSVM (Huertas-Company et al. 2009). The morphological types are E: Elipticals, S0 lenticulars, Sab: Early Spirals and Scd: Late Spirals.

[2] http://www.lesia.obspm.fr/perso/marc-huertas/galvsm.php

PREDICTING GALAXY DISTANCES: NEURAL NETWORKS

Problem: Given a particular galaxy with a several measurement of its photometry in different band passes, predict its redshift (~distance).

Approach: A variety of methods exist. In particular, ANNz[3] (*Collister & Lahav 2004*), uses artificial neural networks trained with a sample of galaxies with precise measurements.

Results: This method has been applied to many datasets. When the training set is large and representative, the results are much more accurate than traditional methods, as for the SDSS data (see Fig. 3).

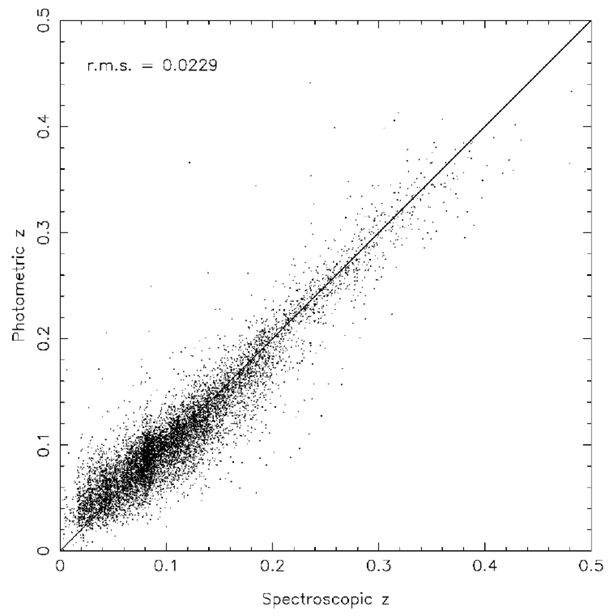
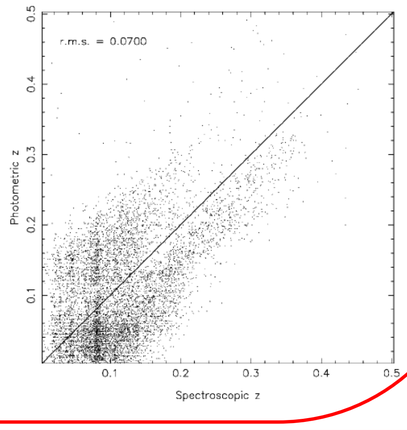


Fig 3: (Left top plot) Photometric redshift obtained using ANNz versus spectroscopic redshift for a sample of the SDSS data. The scatter obtained in the results is >3 times smaller than other techniques applied in the same dataset such as template fitting (right bottom plot).



[3] https://github.com/IftachSadeh/ANNZ

MANY MORE…

Many other problems in Astronomy has been analyzed using ML techniques with excellent results. Some other examples are:

1. **Classification of emission-lines** in galaxy spectra using **PCA, k-nearest neighbor, k-means clustering and SVM (**e.g. *Beck et al. 2016*)
2. **Supernovae classification** using **Kernel Principal Component Analysis** (e.g. *Ishida & de Souza 2013*)
3. **Prediction of dynamical mass measurements** using **regression and Support Distribution machine (SDM) methods** (e.g. *Ntampaka et al. 2015, 2016*)

REFERENCES

Ascaso, Wittman & Benítez 2012, MNRAS, 420, 1167
Ascaso, Wittman & Dawson 2014, MNRAS, 439, 1980
Ascaso et al., 2015, MNRAS, 452, 549
Beck et al. 2016, MNRAS, 457, 362
Collister & Lahav 2004, 2004, PASP, 116, 345
Huertas-Company et al. 2008, A&A, 478, 971
Huertas-Company et al. 2009, A&A, 497, 743
Huertas-Company et al. 2011, A&A, 525, 157
Ishida & de Souza 2013, MNRAS, 430, 509
Ntampaka et al. 2015, ApJ, 803, 50
Ntampaka et al. 2016, ApJ, 831, 135