Advanced Data Analysis Techniques

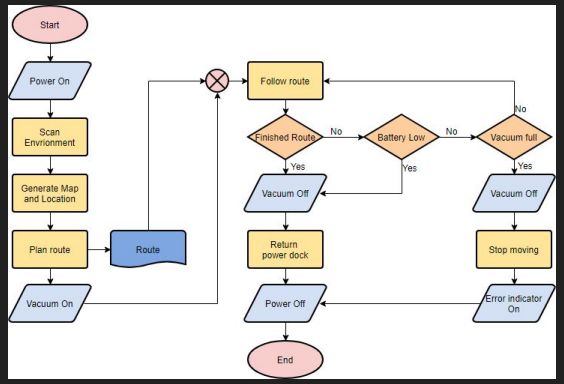
A.Ederoclite

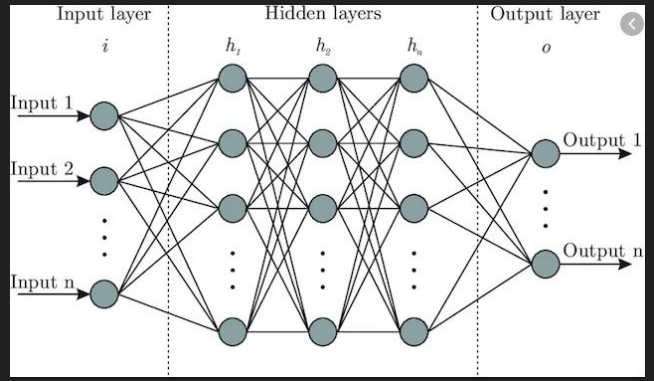
FONDO DE INVERSIONES DE TERUEL
GOBIERNO DE ARAGON
FEDER
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
AGENCIA ESTATAL DE INVESTIGACIÓN
CEFCA
Centro de Estudios de Física del Cosmos de Aragón
Ayuda ICTS-MRR-2021-03-CEFCA, financiada por:
MINISTERIO DE CIENCIA E INNOVACIÓN
Plan de Recuperación, Transformación y Resiliencia
ESPAÑA PUEDE.
Financiado por la Unión Europea
NextGenerationEU

I will want to talk to you because we may do something “hands on” in the afternoon

Machine Learning

Machine Learning and Artificial Intelligence have the inconvenient of “anthropomorphizing” computing science.





Resources

https://ned.ipac.caltech.edu/level5/March19/Baron/Baron_contents.html

Viviana Acquaviva's book

The classic Ivezic (et al.) book

This field is moving very fast and any resource you find is likely obsolete.

Let’s create our catalogue for today’s exercise.

Use TOPCAT to get the Gaia EDR3 Nearby Stars catalogue.

Then cross match it with Simbad and with Gaia DR3 Atmospheric Parameters.

Let’s look at the three “galaxies”.

There are mostly two flavours of ML:

- Supervised:
  - In supervised ML, the data are tagged (e.g. you know what is a galaxy or a star) and you use a training sample of tagged data to learn about untagged data
- Unsupervised:
  - In unsupervised ML, you try to guess “similarity” of untagged data

Advantage of ML:

Computers do not care if you have 1,2 or 150 dimensions

Disadvantage of ML:

If you do not know what you are putting in your algorithm, you will not understand what the algorithm is returning

Supervised ML

A couple of examples:

- Random forest
  - Random forest algorithm for classification of multiwavelength data (Gao et al. 2009) https://iopscience.iop.org/article/10.1088/1674-4527/9/2/011
- Neural Networks
  - J-PLUS: Identification of low-metallicity stars with artificial neural networks using SPHINX, Whitten et al. 2019 https://ui.adsabs.harvard.edu/abs/2019A%26A...622A.182W/abstract

Deep Learning

A “glorified” neural network

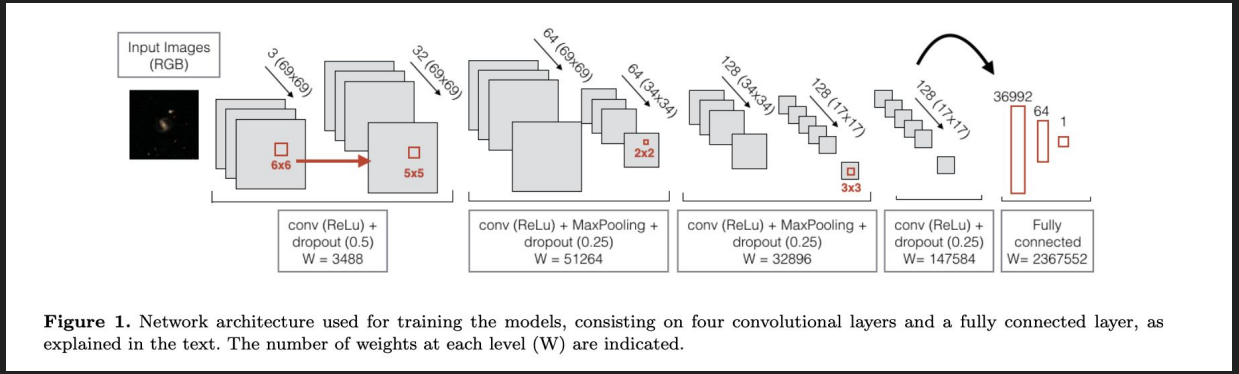


Figure 1. Network architecture used for training the models, consisting on four convolutional layers and a fully connected layer, as explained in the text. The number of weights at each level (W) are indicated.

Example: https://arxiv.org/abs/1711.05744

Unsupervised ML

Examples:

- Principal Component Analysis
  - A principal component analysis of quasar UV spectra at z ~ 3 (Paris et al. 2011) https://ui.adsabs.harvard.edu/abs/2011A%26A...530A..50P/abstract
- K-means
  - Automated Unsupervised Classification of the Sloan Digital Sky Survey Stellar Spectra using k-means Clustering (Almeida and Allende-Prieto 2013) https://ui.adsabs.harvard.edu/abs/2013ApJ...763...50S/abstract

The Bayesian Framework

The Bayes Theorem

The Bayesian framework is now obiquitous in astronomy.

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

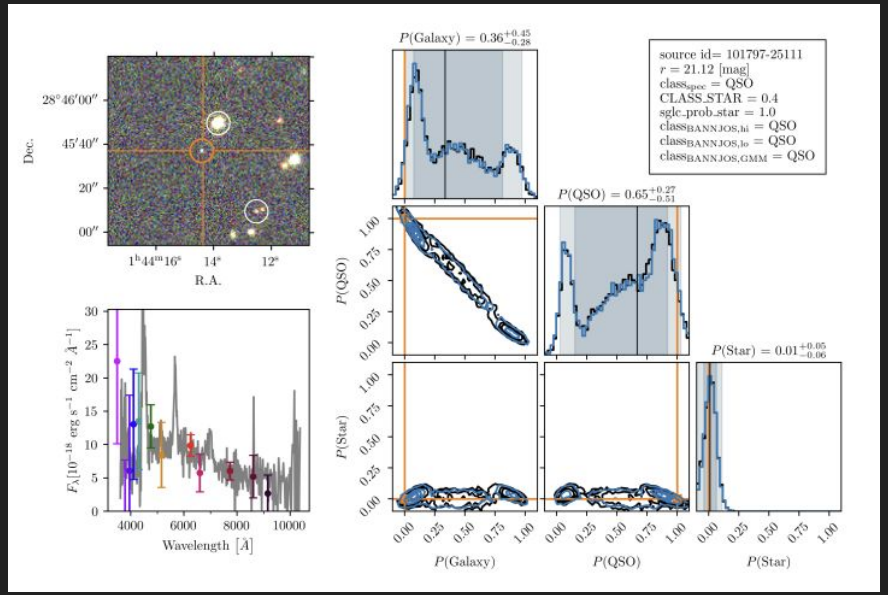
The Gaia parallaxes are probability distribution functions!

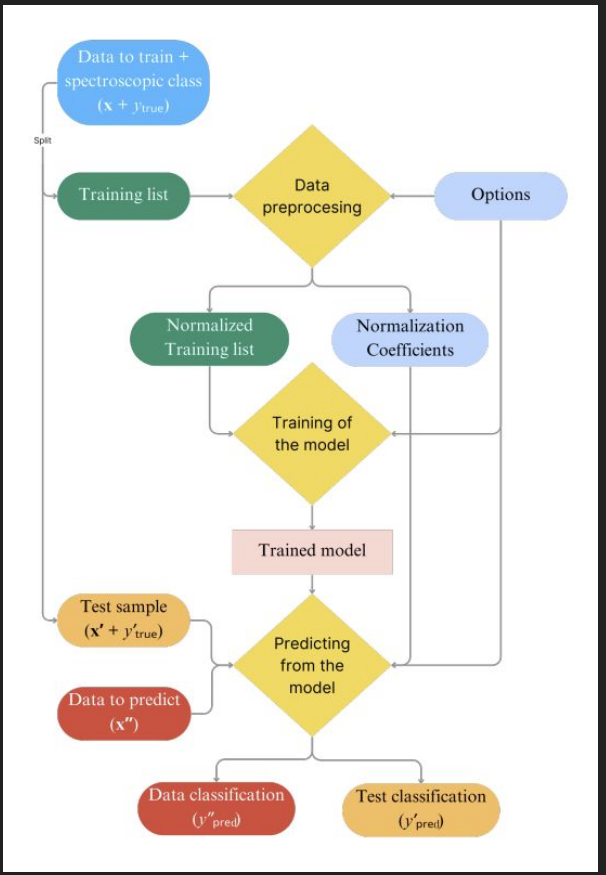
This is an excellent review:
https://arxiv.org/abs/2302.04703

$$p(d|y) \propto p(y|d)p(d).$$

Bayesian Neural Networks

Best of both worlds! https://arxiv.org/abs/2404.16567





Conclusions

In modern data analysis, we often run the risk of running algorithms as black boxes.

Even worse: we assume that the machine will answer questions we are not even asking.

Machine learning gives great tools for different tasks.

The Bayesian framework is a little explored dimension.