UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Astrostatistics

2122-1-F5802Q014

Aims

The use of statistics is ubiquitous in astronomy and astrophysics. Modern advances are made possible by the application of increasingly sophisticated tools, often dubbed as "data mining", "machine learning", and "artificial intelligence". This class provides an introduction to (some of) these statistical techniques in a very practical fashion, pairing formal derivations to hands-on computational applications. Although examples will be taken almost exclusively from the realm of astronomy, this class is appropriate to all Physics students interested in machine learning.

Contents

Detailed program

Statistics topic covered:

- Brief recap on probability and statistical inference.
- Bayesian inference (role of priors, difference and similarities with the frequentist approach, model selection, MCMC).
- Looking for structure in the data (cluster algorithms, parametric vs non-parametric estimators).
- Dimensionality reduction (e.g. Principal Component Analysis).
- Regression problems (overfitting, gaussian process regression).
- Classification problems (neural networks, ROC curves).
- Deep learning.
- Time-series analyses.

Some examples of astrophysical datasets we might use include (TBC):

- Data from the Sloan Digital Sky Survey. Large dataset with 357 million unique sources and 1.6 million follow-up spectra.
- The LINEAR database, containing time-domain observations of thousands of variable stars.
- Time series data from the Laser Interferometer Gravitational-Wave observatory (LIGO), which has detected more than 50 gravitational wave events.
- Parallaxes data from the GAIA satellite.

Prerequisites

No formal prerequisites. Some previous knowledge of the python programming language is highly recommended (see below for some catch-up resources).

Teaching form

Lessons, 6 credits.

Textbook and teaching resource

Main textbook:

- "Statistics, Data Mining, and Machine Learning in Astronomy", Željko, Andrew, Jacob, and Gray. Princeton University Press, 2012.
- Supporting software: https://www.astroml.org/

Other relevant resources:

- "Statistical Data Analysis", Cowan. Oxford Science Publications, 1997
- "Data Analysis: A Bayesian Tutorial", Sivia and Skilling. Oxford Science Publications, 2006
- "Bayesian Data Analysis", Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin. Chapman & Hall, 2013. Free: http://www.stat.columbia.edu/~gelman/book/
- "Python Data Science Handbook", VanderPlas. O'Reilly Media, 2016. Free: https://jakevdp.github.io/PythonDataScienceHandbook/
- "Practical Statistics for Astronomers", Wall and Jenkins. Cambridge University Press, 2003

- "Bayesian Logical Data Analysis for the Physical Sciences", Gregory. Cambridge University Press, 2005
- "Modern Statistical Methods For Astronomy" Feigelson and Babu. Cambridge University Press, 2012
- "Information theory, inference, and learning algorithms" MacKay. Cambridge University Press, 2003. Free: https://www.inference.org.uk/mackay/itila/book.html

- “Data analysis recipes", Hogg et al.:
  - https://arxiv.org/abs/0807.4820
  - https://arxiv.org/abs/1008.4686
  - https://arxiv.org/abs/1205.4446
  - https://arxiv.org/abs/1710.06068
  - https://arxiv.org/abs/2005.14199

Catch-up resources for students who need to refresh their Python skills:

- https://github.com/jrjohansson/scientific-python-lectures
- https://astrofrog.github.io/py4sci/
- "Learning Scientific Programming with Python", Hill, Cambridge University Press, 2020. https://scipython.com/

Semester

Second semester.

Assessment method

The class will be assessed with an oral exam. A computational problem will be assigned beforehand; students will need to complete it in their own time and discuss it during the exam.

All classes, excercises, and exams will be in English.

Office hours

Any time, please contact me by email.