

Advanced Astroinformatics - Student Project

Machine Learning: Intro to Scikit-Learn

Dr. Nina Hernitschek
June 27, 2022

shifting use cases:

As data become more plentiful, finding the *right* data has become more important.

Motivation

Machine
Learning

scikit-learn

Data Mining

shifting use cases:

As data become more plentiful, finding the *right* data has become more important.



Data Mining

Motivation

Machine
Learning

scikit-learn

shifting use cases:

As data become more plentiful, finding the *right* data has become more important.



Data Mining

Individual measurements giving way to **statistics, clustering, patterns** in the data.

Data processing needs to be **highly automatized**.

Analysis growing more exploratory rather than pre-defined/
scripted.

Examples:

Finding and classifying variable stars in PS1 3π required processing of 10^9 sparse light curves \Rightarrow 44,000 RAB stars.

Transient science (gravitational wave follow-up, GRBs, unknowns from LSST) requires rapid access to data sets of what is already known, anywhere on sky.

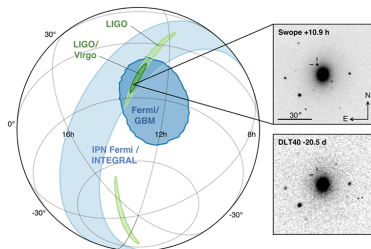


image credit: LIGO

Machine Learning

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

Motivation

Machine
Learning

scikit-learn

Machine Learning

Motivation

Machine
Learning

scikit-learn

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

⇒ allows to **uncover hidden correlation patterns** through iterative learning by sample data

Machine Learning

Motivation

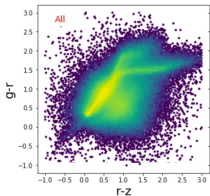
Machine Learning

scikit-learn

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959)

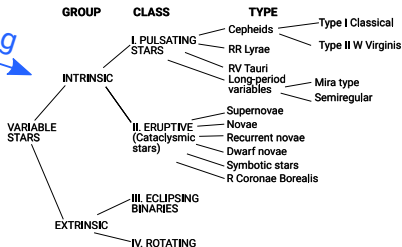
⇒ allows to **uncover hidden correlation patterns** through iterative learning by sample data

parameter space of measurements



machine learning

parameter space of astrophysical objects



Machine Learning

Motivation

Machine
Learning

scikit-learn

... is the sub-field of computer science that gives computers the ability to learn without being explicitly programmed
(Arthur Samuel, 1959)

⇒ allows to **uncover hidden correlation patterns** through iterative learning by sample data

⇒ allows **to model a survey**:

- describing data quality → outlier
- describing light curve characteristics → “features”
- classifying sources → catalogs
- finding substructure → clumps, overdensities, ...

Unsupervised vs. Supervised Learning

Motivation

Machine
Learning

scikit-learn

unsupervised learning or “learning without labels”

Clustering:

Find subtypes or groups that are not defined a priori based on measurements

⇒ members of the same cluster are “close” in some sense

vs.

supervised learning or “learning without labels”

Classification:

Use a priori group labels in analysis to assign new observations to a particular group or class

Regression:

instead of having training data with discrete labels, the “truth” is a continuous property and we are trying to predict the values of that property for the test data

Unsupervised vs. Supervised Learning

supervised learning or “learning without labels”

Classification:

Use a priori group labels in analysis to assign new observations to a particular group or class

Regression:

instead of having training data with discrete labels, the “truth” is a continuous property and we are trying to predict the values of that property for the test data

example:

The task of determining whether an object is a star, a galaxy, or a quasar is a classification problem: the label is from three distinct categories. On the other hand, we might wish to estimate the age of an object based on such observations: this would be a regression problem, because the label (age) is a continuous quantity.

Clustering Methods

- Abell clustering richness class (Abell 1958)



Motivation

Machine
Learning

scikit-learn

Clustering Methods

- Abell clustering richness class (Abell 1958)



- Gamma Ray Bursts: use properties of GRBs (e.g. location on the sky, arrival time, duration) to find classes of events

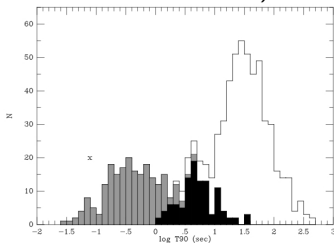
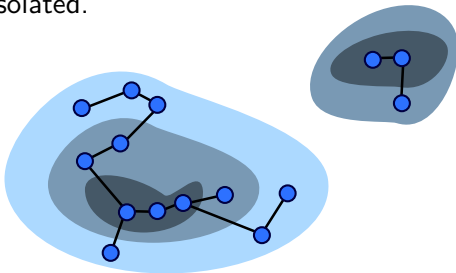


image credit: (Mukherjee+1998)

Clustering Methods

Percolation or 'Friends of Friends (FoF)' algorithm

1. Plot data points in a 2-dimensional diagram (or: calculate distances using a metric).
2. Find the closest pair, and call the merged object a *cluster*.
3. Repeat step 2 until some chosen threshold is reached. Some objects will lie in rich clusters, others have one companion, and others are isolated.

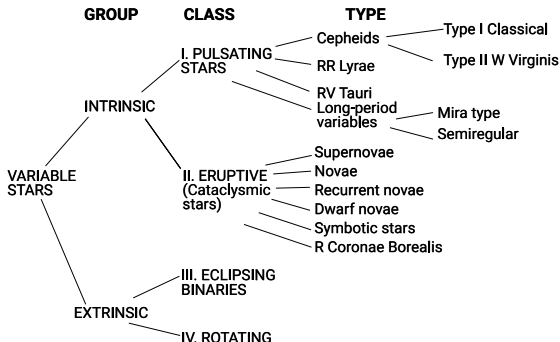


Classification Methods

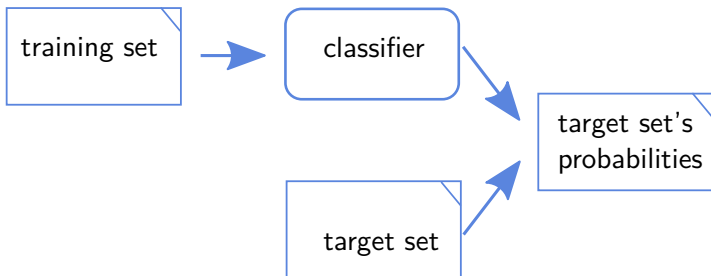
Classification

Use a priori group labels in analysis to assign new observations to a particular known group or class.

⇒ *supervised learning* or “learning with labels”.



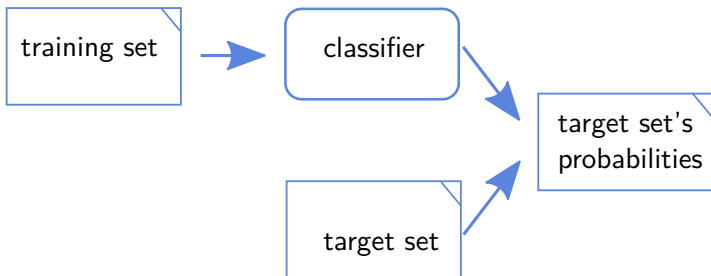
Concepts of Supervised Classification



training set:

- set of sources inside/outside the category we are looking for
- same data quality as found in target set

Concepts of Supervised Classification



training set:

- set of sources inside/outside the category we are looking for
- same data quality as found in target set

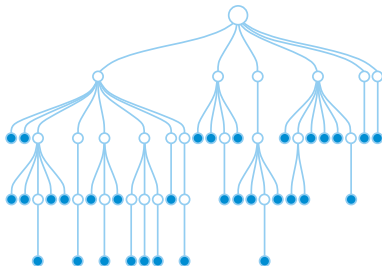
What's happening internally?

Concepts of Supervised Classification

The learning process (“training”):

To build a decision tree, the set is divided into smaller and smaller subsets by **splitting** w.r.t. a single **feature** at a time.

Split criteria: select feature and split point to produce the smallest impurity in the two resultant nodes based on the **training set**.



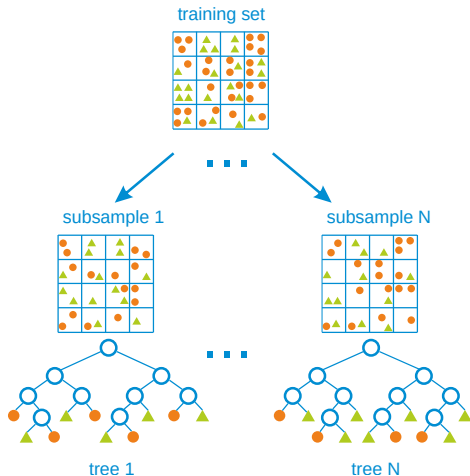
Supervised Classification - Ensemble Methods

Random Forest Classifier as ensemble method: many trees are grown from subsets of the training set

Motivation

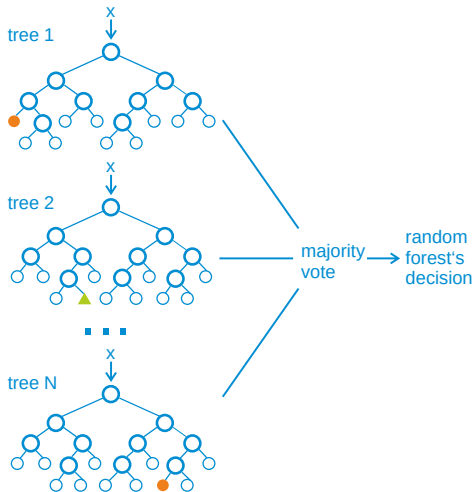
Machine
Learning

scikit-learn



Supervised Classification - Ensemble Methods

Random Forest Classifier as ensemble method: ... and are “voting” for classification



Motivation

Machine Learning

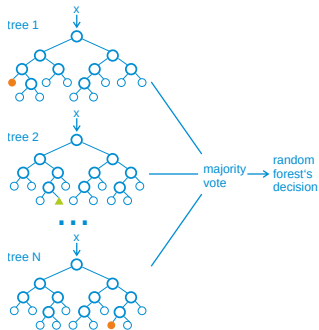
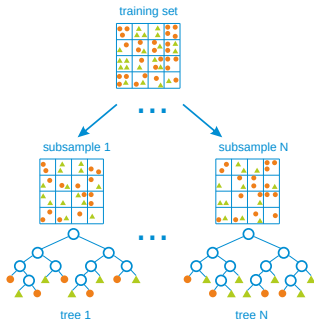
scikit-learn

Supervised Classification - Ensemble Methods

Motivation

Machine
Learning

scikit-learn



divide-and-conquer approach improves classification performance

- less sensitive to training set variances
- robust to outliers
- training and classification can be parallelized

Supervised Classification - Verification

don't apply a classifier as a “black box”!

several concepts for **verification**

Motivation

Machine
Learning

scikit-learn

Supervised Classification - Verification

don't apply a classifier as a "black box"!

several concepts for **verification**

make usage of the training set \Rightarrow **10-fold cross-validation**

10 % held out \Rightarrow train on 90%, apply to 10% in turn

Supervised Classification - Verification

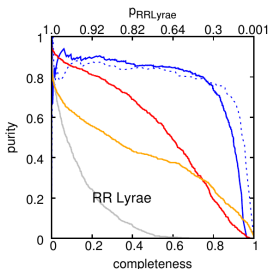
don't apply a classifier as a "black box"!

several concepts for **verification**

make usage of the training set \Rightarrow **10-fold cross-validation**

10 % held out \Rightarrow train on 90%, apply to 10% in turn

purity-completeness (or precision-recall) curves



completeness:

$\# \text{ selected true RR Lyrae} / \# \text{ true RR Lyrae}$

purity:

$\# \text{ selected true RR Lyrae} / \# \text{ all selected sources}$



scikit-learn is a popular Python package containing a collection of tools for **machine learning**

it includes algorithms used for classification, regression and clustering

it comes with an extensive **online documentation**:

<http://scikit-learn.org/stable/tutorial/basic/tutorial.html>

scikit-learn is built upon Python's NumPy (Numerical Python) and SciPy (Scientific Python) libraries, which enable efficient in-core numerical and scientific computation within Python.

scikit-learn uses 3 steps for **developing, applying and testing** machine learning algorithms:

- Train the model using an existing data set describing the phenomena you need the model to predict.
- Test the model on another existing data set to ensure it performs well.
- Use the model to predict phenomena as needed for your project.

Break & Questions

afterwards we continue with `notebook_5.ipynb` from the github repository

Motivation

Machine
Learning

scikit-learn