A not so gentle introduction to machine learning in python using scikit-learn

Christos Aridas

https://www.linkedin.com/in/chris-aridas

Western Greece Software Development Group

Meetup #11

About me


ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

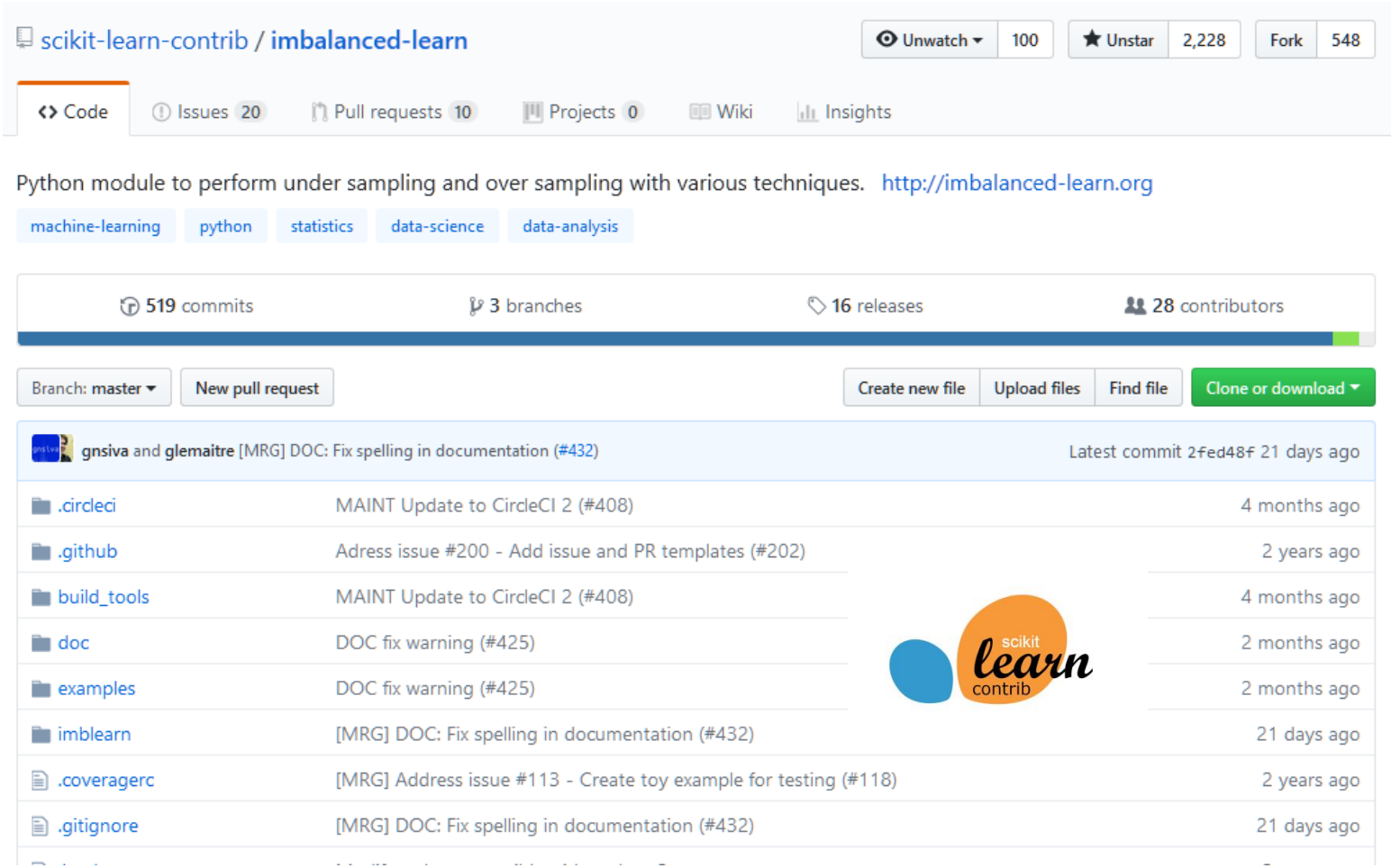

ΕΛΛΗΝΙΚΟ
ΑΝΟΙΚΤΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ


DYNACOMP
ΕΦΑΡΜΟΓΕΣ ΥΨΗΛΗΣ ΤΕΧΝΟΛΟΓΙΑΣ


ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
19 64
UNIVERSITY OF PATRAS


UNIVERSITY OF
PATRAS
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

About me


scikit-learn-contrib / imbalanced-learn
Unwatch 100
Unstar 2,228
Fork 548
Code
Issues 20
Pull requests 10
Projects 0
Wiki
Insights
Python module to perform under sampling and over sampling with various techniques. http://imbalanced-learn.org
machine-learning
python
statistics
data-science
data-analysis
519 commits
3 branches
16 releases
28 contributors
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
gnsiva and glemaitre [MRG] DOC: Fix spelling in documentation (#432)
Latest commit 2fed48f 21 days ago
.circleci MAINT Update to CircleCI 2 (#408) 4 months ago
.github Adress issue #200 - Add issue and PR templates (#202) 2 years ago
build_tools MAINT Update to CircleCI 2 (#408) 4 months ago
doc DOC fix warning (#425) 2 months ago
examples DOC fix warning (#425) 2 months ago
imblearn [MRG] DOC: Fix spelling in documentation (#432) 21 days ago
.coveragerc [MRG] Address issue #113 - Create toy example for testing (#118) 2 years ago
.gitignore [MRG] DOC: Fix spelling in documentation (#432) 21 days ago
scikit
learn
contrib

About you

- Machine Learning
- Python
- scikit-learn

What is Machine Learning?

- Machine learning is the process of extracting knowledge from data automatically
- Nowadays touch nearly every aspect of everyday life, from the face detection in our phones and the spam filtering to picking restaurants, partners, and movies.
- A classical example is a spam filter, for which the user keeps labeling incoming mails as either spam or not spam. A machine learning algorithm then "learns" a predictive model from data that distinguishes spam from normal emails, a model which can predict for new emails whether they are spam or not.

Machine Learning Concepts

- **Automating decision making** from data **without the user specifying explicit rules** how this decision should be made.
- **Generalization**

Machine Learning Concepts: Data

- Two-dimensional array (or matrix) of numbers.
- Each data point (aka *sample* or *training instance or object or example*) that we want to either learn from or make a decision on is represented as a list of numbers, a so-called feature vector, and its containing features (or attributes) represent the properties of this point.

Machine Learning Concepts: Data

Iris, a classic benchmark dataset in the field of machine learning, contains the measurements of 150 iris flowers from 3 different species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica.

Machine Learning Concepts: Data

Iris, a classic benchmark dataset in the field of machine learning, contains the measurements of 150 iris flowers from 3 different species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica.

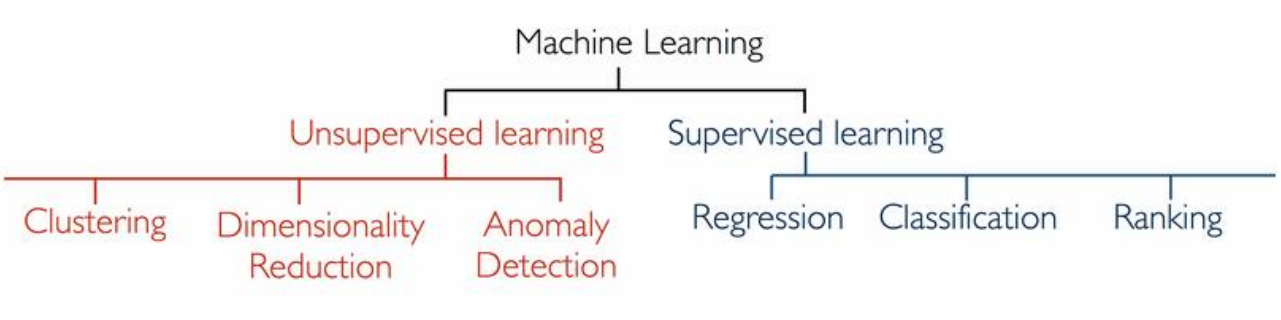
Machine Learning Concepts: Data

Iris, a classic benchmark dataset in the field of machine learning, contains the measurements of 150 iris flowers from 3 different species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica.

- Each flower sample represented as one row in our data array, and the columns (features or attributes) represent the flower measurements in centimeters.

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \dots & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \dots & x_4^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & \dots & x_4^{(150)} \end{bmatrix}.$$

Machine Learning Taxonomy



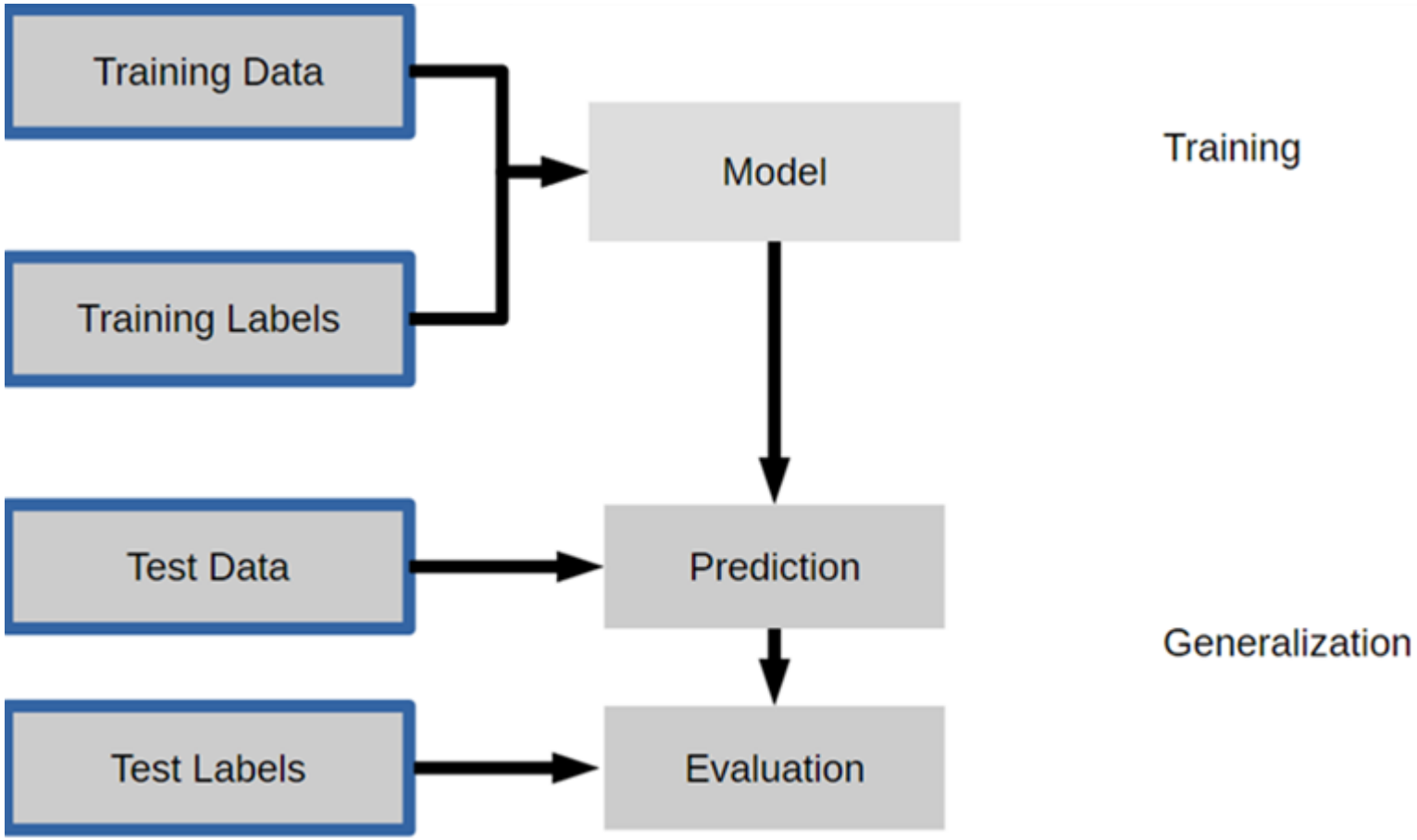
Supervised Learning: Classification and regression

- In Supervised Learning, we have a dataset consisting of both input features and a desired output.
- The task is to construct a model (or program) which is able to predict the desired output of an unseen object given the set of features.

Supervised Learning: Classification and regression

- **In classification, the label is discrete**
- **In regression, the label is continuous**

Supervised Learning: Classification and regression



Unsupervised Learning

- There is no desired output associated with the data
- We are interested in extracting some form of knowledge
- You can think of unsupervised learning as a means of discovering labels from the data itself.

Unsupervised Learning

- Is it more challenging than Supervised Learning?
  - Probably. But why?
  - Is often harder to understand and to evaluate

Why scikit-learn?

- **Python of course**
- **Commitment to documentation and usability**
- **Models are chosen and implemented by a dedicated team of experts**
- **Covers most machine-learning tasks**
- **Scales to most data problems**
- **Focus**

https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn

Why scikit-learn?

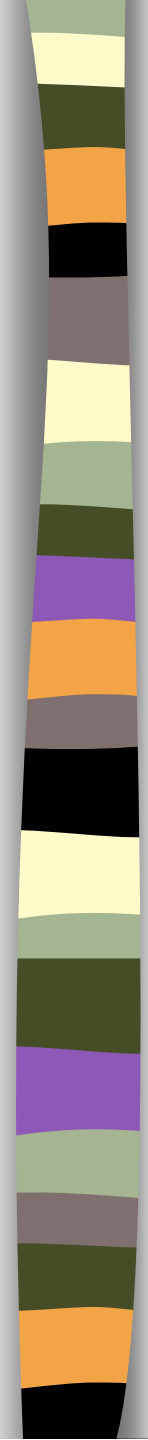
Classification
Regression
Clustering
Semi-Supervised Learning
Feature Selection
Feature Extraction
Manifold Learning
Dimensionality Reduction
Kernel Approximation
Hyperparameter Optimization
Evaluation Metrics
Out-of-core learning

https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn

Who is using scikit-learn?

The New York Times

Booking.com



Coding Time