Updated on 10/04/2024

Sign up

Python training for scientists with Scikit-Learn
2 days (14 hours)

Presentation

Scikit-learn (or Sklearn) is a free Python library dedicated to machine learning. It has been developed by numerous contributors, including French higher education and research institutes such as Inria and Télécom ParisTech.

It includes functions for estimating random forests, logistic regressions, classification algorithms and support vector machines. It is designed to harmonize with other free Python libraries, notably NumPy and SciPy.

This Python for Scientists course will give you the knowledge you need to interact with Sklearn. It will cover its syntax, tools and best development practices, so that you can take full advantage of this library. We'll present the algorithms, supervised and unsupervised methods, classification, regression and clustering functionalities.

Like all our training courses, this one will introduce you to the latest stable version and its new features (Scikit-learn 1.4 released in June 2023 at the time of writing, and Python 3.12).

Objectives

- Implementing machine learning algorithms in Python with Scikit-Learn
- Choose a scalability strategy for Scikit-Learn
- Use Scikit-Learn in conjunction with other toolkits in the Python universe (e.g. Skimage or OpenCV)
- Be able to name the different components of Scikit-Learn, their use, find quickly their documentation

Target audience

- Data analysts
- Data Miners
- Statisticians
- Scientists

Prerequisites

- Programming in Python 2 or 3
- Good knowledge of data processing and manipulation
- Good knowledge of linear algebra

Scientific Python with Scikit.Learn training program

Introduction

- What is Machine Learning?
- Machine Learning ?
- Scikit-learn installation
- Clean Data
- Over-adjustment and generalization
- Cross-validation
- Bias-Variance Trade-Off
- Checking and updating scikit-learn installation
- Solutions and benefits: the state of the art
  - Tensorflow and Keras
  - Scikit?learn
  - Deeplearning for Java
  - pytorch and fastai

SCIKIT basic concepts

- Algorithms, vocabulary
- Scikit-learn overview
- Learning process
- Toolkit scope of application
- Structure
- Data representation and API principles
- Documentation

MACHINE LEARNING METHODS

- Supervised methods
  - KNN
  - SVM
  - Neural networks
  - Naive Bayesian network
  - Regression penalized
  - Boosting
  - Random forest
- Unsupervised methods
  - Clustering
  - Association and sequence rules
  - Text mining

Problem categories and implementation

- Classification
  - Definition of decision trees
  - Decision trees
  - Overall models
  - Random Forest Classifier
- Regression
  - Decision trees
  - Regression models
  - Definition of a hyperplane
  - Anatomy of an estimator
  - Simple linear regression
  - The COST function
  - What is R-Squared?
  - Model evaluation
  - Regularization
  - Multivariate linear regression
  - Applied linear regression
  - Visualizing our DataSet
- Clustering
  - Leading the K-Means Cluster
  - The K-Means algorithm
  - Applying K-Means clustering
- Other algorithms available

The models

- Model validation
- Model preparation
  - Images
  - Texts
  - Size reduction
  - Other preparations

Optimization

- How to set your ML algo?
- HPO = Hyper Parameter Optimization
- AutoML
- Introducing a kaggle master
  - Using Bayesian optimization

Using SCIKIT management parameters

- Choosing the right estimator and parameters
- Interoperability with the scientific python ecosystem
- Performance and scalability
  - Strategy
  - Tools

Companies concerned

This training course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced computer technology, or to acquire specific business knowledge or modern methods.

Positioning on entry to training

Positioning at the start of training complies with Qualiopi quality criteria. As soon as registration is finalized, the learner receives a self-assessment questionnaire which enables us to assess his or her estimated level of proficiency in different types of technology, as well as his or her expectations and personal objectives for the training to come, within the limits imposed by the selected format. This questionnaire also enables us to anticipate any connection or security difficulties within the company (intra-company or virtual classroom) which could be problematic for the follow-up and smooth running of the training session.

Teaching methods

Practical course: 60% Practical, 40% Theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical input from the trainer, supported by examples, with brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Sanction

A certificate will be issued to each trainee who completes the course.