FUN FRANCE UNIVERSITÉ NUMÉRIQUE
The freedom to study

You are here: Home › Courses
› Machine learning in python with scikit-learn

Machine learning in Python with scikit-learn

Ref. 41026

Computer science and programming Digital and technology

Build predictive models with scikit-learn and gain a practical understanding of the strengths and limitations of machine learning!

Effort: 36 hours Pace: Self paced

Languages: English

Enrollment

From Sep 18, 2023 to Oct 30, 2024

Course

From Nov 08, 2023 to Nov 07, 2024

Languages

English

Log in to enroll

What you will learn

At the end of this course, you will be able to:

- Grasp the fundamental concepts of machine learning
- Build a predictive modeling pipeline with scikit-learn
- Develop intuitions behind machine learning models from linear models to gradient-boosted decision trees
- Evaluate the statistical performance of your models

Description

Predictive modeling is a pillar of modern data science. In this field, scikit-learn is a central tool: it is easily accessible, yet powerful, and naturally dovetails in the wider ecosystem of data-science tools based on the Python programming language.

This course is an in-depth introduction to predictive modeling with scikit-learn. Step-by-step and didactic lessons introduce the fundamental methodological and software tools of machine learning, and is as such a stepping stone to more advanced challenges in artificial intelligence, text mining, or data science.

The course is more than a cookbook: it will teach you to be critical about each step of the design of a predictive modeling pipeline: from choices in data preprocessing, to choosing models, gaining insights on their failure modes and interpreting their predictions.

The training will be essentially practical, focusing on examples of applications with code executed by the participants.

The **MOOC is free of charge**, all the course materials are available at: https://inria.github.io/scikit-learn-mooc/.

The **authors of the course are scikit-learn core developers**, they will be your guides throughout the training!

Format

The course will cover practical aspects through the use of Jupyter notebooks and regular exercises. Throughout the course, we will highlight scikit-learn best practices and give you the intuition to use scikit-learn in a methodologically sound way.

Prerequisites

The course aims to be accessible without a strong technical background. The requirements for this course are:
- basic knowledge of Python programming : defining variables, writing functions, importing modules
- some prior experience with the NumPy, pandas and Matplotlib libraries is recommended but not required

For a quick introduction to these libraries, you can use the following resources : Introduction to NumPy and Matplotlib by Sebastian Raschka and 10 minutes to pandas.

Assessment and certification

Students' work in the course is assessed through quizzes after the lessons and programming exercises at the end of every modules.

An Open Badge for successful completion of the course will be issued on request to learners who obtain an overall score of 60% correct answers to all the quizzes and programming exercises.

Course plan

Introduction

- Machine Learning concepts

Module 1. The Predictive Modeling Pipeline

Tabular data exploration
Fitting a scikit-learn model on numerical data
Handling categorical data

Module 2. Selecting the best model

Overfitting and Underfitting
Validation and learning curves
Bias versus variance trade-off

Module 3. Hyperparameters tuning

Manual tuning
Automated tuning

Module 4. Linear Models

Intuitions on linear models
Linear regression
Modelling with a non-linear relationship data-target
Regularization in linear model
Linear model for classification

Module 5. Decision tree models

Intuitions on tree-based models
Decisison tree in classification
Decision tree in regression
Hyperparameters of decision tree

Module 6. Ensemble of models

Ensemble method using bootstrapping

Ensemble based on boosting

Hyperparameters tuning with ensemble methods

Module 7. Evaluating model performance

Comparing a model with simple baselines

Choice of cross-validation

Nested cross-validation

Classification metrics

Regression metrics

Other course runs

Archived

From May 18, 2021 to Jul 14, 2021

From Feb 15, 2022 to May 17, 2022

From Oct 18, 2022 to Jan 17, 2023

Course team

Arturo Amor

Arturo Amor is an engineer at Inria. He is in charge of broadening the scikit-learn documentation's accessibility to all kind of users.