+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on 29/11/2023

Sign up

Training Cluster Analysis (Data Partitioning) with Python

4 days (28 hours)

Presentation

Clustering is one of the essential methods of data analysis. Indeed, grouping data into distinct, homogeneous groups can benefit a wide range of fields, including health, marketing and finance.

In marketing, the creation of clusters (also known as segments) makes it possible to categorize each customer. This categorization has a positive effect on the performance of your campaigns, because your sponsored messages will be personalized for your target group.

Clustering can also be useful for fraud detection through visual recognition. This system is particularly interesting for signature recognition in cybersecurity or finance.

To create our data partitions, we'll be using one of the world's most widely used programming languages, Python. Thanks to its scikit-learn library, Python has all the functions needed to create data clusters efficiently.

Our cluster analysis training course will introduce you to Python programming for data analysis, and show you the benefits and use cases of clustering methods. By the end of the course, you'll know how to create clusters and analyze them with Python.

As always, our training will be based on the latest version of the language, Python 3.10.

Objectives

- Using Python for data analysis

- Understanding the benefits of clustering
- Understand the main types of clustering algorithms
- Preparing data with Python
- Represent and analyze clusters

Target audience

- Data Analyst
- Data Scientist
- Data Engineer
- Machine learning engineer
- Company manager
- Analyst
- Marketing manager

Prerequisites

Knowledge of general mathematics (probability, statistics, etc.).

Program of our Data Cluster Analysis training course

Introduction

- What is a cluster?
- The difference between clustering and segmentation
- The benefits of cluster analysis, use cases
- The limits and challenges of clustering

Data partitioning methods

- K-means
- Mean-Shift
- DBSCAN (spatial clustering of applications with density-based noise)
- Hope maximization algorithm with or without Gaussian Mixture Models (GMM)
- Hierarchical grouping

Introducing Python

- Why use Python?
- Introducing the Scikit learn library
- Using library functions
- Managing modules and libraries

Getting started with Python

- Python syntax
- Variables
- The different types of data sets
  - Tuple
  - List
  - Set
  - Dictionary
- The functions
- Write your own functions

Preparing your data with Python

- The importance of data integrity and preparation
- Reading and editing CSV files
- Import data
- Clean and prepare your data
- Data formatting
- Building data pipelines

K-Means Clustering

- Import sklearn modules
- Import data
- Parameters
  - n_samples
  - centers
  - cluster_std
- The make_blobs() function
- Using standardization
- Using the KMeans function
- How to choose the right number of clusters
- Graphical representation of clusters

Mean-Shift

- Import MeanShift and make _blobs
- Determining cluster centers
- 3D data representation

DBSCAN

- Import data
- Parameter description
- Cluster your data
- Represent data groupings graphically

Expectation-maximization

- Concatenate Gaussian curves
- Explanation of the expectation-maximization algorithm
- Graphical representation of data partitions

Hierarchical grouping

- Preparing data
- Calculate similarity information between each data item
- Using a link function
- Determine the cut-off point of the hierarchical tree

Analyze your results

- Score validation methods
- Evaluation of clustering
- Improving these clusters
- Constraint-based clustering
  - Measurements based on matching
  - Entropy-based measurements
  - Measurements in pairs
- Internal measurements to validate clusters
- Cluster stability

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 inputs from the trainer supported by examples and

brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire is used to check that skills have been correctly acquired.

Sanction

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