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

Updated 09/01/2025

Sign up

BentoML training

2 days (14 hours)

Presentation

Our BentoML training course will teach you to master and understand the creation, management and deployment of Machine Learning models. Deepen your knowledge and skills in production, while simplifying processes.

With our training, enter the era of "intelligent deployment" BentoML is the ideal solution for simplifying the deployment of your machine learning models. This open tool lets you control your models via APIs, while guaranteeing secure, flexible project management.

During this course, you'll discover how to modernize ML model deployments by transforming your application systems to enable efficient horizontal scaling and seamless integration into your existing infrastructures.

As with all our training courses, this one will introduce you to the latest version of BentoML V1.3.19.

Objectives

- Understand the benefits and features of BentoML for ML model deployment
- Transform ML models into production-ready application services
- Master model packaging and management tools with BentoML
- Enable automatic scaling of prediction services
- Optimize dependency and resource management for heavy models
- Develop advanced performance monitoring for models in production
- Automate model deployment and version updates

Target audience

- DevOps
- Developers
- Data Scientists
- Machine Learning Engineers

Prerequisites

- Good Python skills
- Knowledge of machine learning

BentoML training program

Introduction to BentoML

- What is BentoML?
- BentoML architecture analysis
- Technical requirements
- BentoML installation and environment configuration
- CLI interface and associated libraries

Model Packaging

- Packaging scikit-learn, TensorFlow, and PyTorch models with BentoML
- Dependency management in a Python environment with specific libraries
- Introduction to BentoBundle

Creating and deploying a simple service

- Creating a REST/GRPC service around an ML model
- Addition of pre-treatment and post-treatment pipelines
- Local service testing
- Error debugging
- Local deployment via Docker
- Kubernetes integration overview

Cloud deployment

- Deployment on AWS, GCP, and Azure
- BentoML with Kubernetes for scaling
- Configuring Docker for multi-environment deployment

- Setting up metrics with Prometheus and Grafana
- Performance analysis
  - latency
  - response time
- Managing model versions in production

Performance optimization and case studies

- Horizontal and vertical scaling
- API optimization for heavy loads
  - Batching
  - Multi-threading
- Using GPUs for heavy models
- Integrating a recommendation model into a web application
- Production of a classification model with a complete pipeline
- Handling common errors
  - Dependencies
  - Incompatibilities

Companies concerned

This 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 enabling 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 with regard to 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, 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.