AMBIENT-IT

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

Updated 07/27/2023

Sign up

R language training

5 days (35 hours)

Presentation

R is a programming language and software dedicated to statistics and data science. Created in 1993, it is used by statisticians, data miners and data scientists for statistical software development and data analysis. It compiles and runs on a wide variety of UNIX, Windows and MacOS platforms. In this Data Science course, we'll learn about the R language, the challenges and pitfalls of unsupervised learning and the rules of supervised learning. Then we'll analyze a model and discover how to process unstructured data. Finally, we'll finish with an introduction to Deep Learning. The course will use the latest stable version of the project (R version 4 to date).

Objectives

- Introduction to the R language
- Understanding unsupervised and supervised learning
- Anticipating Deep Learning

Target audience

Data scientists, Data handlers, Developers, Project managers, Architects

Prerequisites

Basic knowledge of statistics and a programming language.

R language training program: Data Science

Day 1 - Data science philosophy

- Quick history

- Formal foundations of machine learning.
- Distinction supervised, unsupervised, by reinforcement, trade off bias variance
- "Big Data: No ceiling, no floor
- Long tail theory applied to data
- 2 approaches: know the future or change it?
- A micro-decision strategy rather than a decision-making tool

Introduction to R

- Fundamentals
- Loading data with data.table
- Data exploration: synthesis, visualization. Selection/filtering exercises
- Categorical data processing, the notion of dummy variable
- Handling missing data
- Format management (including time and location)
- Generating new features: in-depth exploitation of datatable format

Day 2 - Unsupervised learning

- Synthesis approach
  - Summary by column: Dimension reduction: PCA / ICA
  - Line-by-line synthesis: clustering
  - Kmeans
  - Hierarchical (top down or bottom up)
  - Performance evaluation method: variance / silhouette indicator
- Missing values approach
  - SVD decomposition
  - SGD, ALS

Day 3 - Supervised learning

- Lin2ary regression
  - Formulation, conditions of use
- Performance analysis, pvalue, performance detection
  - Notion of overfitting
  - R2 and adjusted R2
- Feature selection: forward, stepwise approach
- Penalized approach
  - Ridge, Lasso, elastic net.
  - Geometric interpretation
- Decision trees
  - Construction principle
  - Prunning
  - Interpretation, operating context
- Random Forest
  - How to overcome the limits of the decision tree
  - Feature importance, local importance

- Gradient boosting
  - Principles
  - Settings
- XGBosst (extreme gradient boosting)
  - Principles, settings

Day 4 - Fine-tuning and model stripping

- Advanced model setting techniques
  - Cost functions, RMSE, roc curve and auc indicator
  - Adjustment precautions, pitfalls to avoid
    - Model skinning
- Where was the information?
  - Simplify model, advanced feature selection

Introduction to text mining and NLP

- Heaps' and Zipf's laws
- How to structure an unstructured source
  - Bag of words approach
  - Stop word and standardization TF IDF
- Towards NLP (natural language processing)
  - Semantic analysis
  - Deep learning approach

Day 5 - Deep Learning introduction

- Neural networks
- Network architecture
  - Convolution
  - LSTM
- Discover the Keras environment for deploying

Project management

- The different phases of a data project
- Adapting Agile project management to data projects
- Structuring the data science/business dialogue
- Managing the project
- How do you get projects off the ground? When to stop?

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.

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.