Ambient IT
128 Boulevard Macdonald
75019 Paris

Mail : formation@ambient-it.net
Tél : +33 (0)1 40 34 11 53

AMBIENT-IT

Formation Langage R

Durée 5 jours (35 heures)

Présentation

R est un langage de programmation et un logiciel dédié aux statistiques et à la science des données. Créé en 1993, il est utilisé par les statisticiens, les data miner, data scientist pour le développement de logiciels statistiques et l'analyse des données.

Il compile et fonctionne sur une grande variété de plates-formes UNIX, Windows et MacOS.

Dans cette formation Data Science, nous appréhendons le langage R, ensuite, nous apprendrons l'enjeu et les pièges de l'apprentissage non supervisé et les règles de l'apprentissage supervisé. Puis, nous analyserons un modèle et découvrirons le traitement de données non structurées. Enfin, nous finirons cette formation sur une initiation au Deep Learning.

La formation utilisera les dernière version stable en date du projet (R version 4 à ce jour).

Objectifs

- Initiation au langage R
- Comprendre l'apprentissage non supervisé et supervisé
- Anticiper le Deep Learning

Public visé

Data-scientists, Manipulateurs de la data, Développeurs, Chefs de Projets, Architecte

Pré-requis

Des connaissances de base en statistiques et à un langage de programmation

Programme de la formation langage R : Data Science

Jour 1 - Philosophie data science

- Historique rapide
- Fondements formel de l'apprentissage machine.
- Distinction supervisé, non supervisé, par renforcement, trade off biais variance
- « Big Data » : Ni plafond ni plancher
- Théorie de la longue traîne appliquée aux données
- 2 approches : connaître l'avenir ou le changer ?
- Une stratégie de microdécision plus qu'un outil de décision

Initiation à R

- Fondamentaux
- Chargement des données avec data.table
- Exploration des données : par synthèse, par visualisation. Exercices de sélection / filtrage
- Traitement des données catégorielles, notion de dummy variable
- Traitement des données manquantes
- Gestion des formats (dont temps et lieux)
- Génération de nouvelles features : exploitation approfondie du format datatable

Jour 2 - Apprentissage non supervisé

- Approche synthèse
  - Synthese par colonne : Réduction de dimension : PCA / ICA
  - Synthèse par ligne : clustering
  - Kmeans
  - Hiérarchique (top down ou bottom up
  - Méthode d'évaluation de performance : variance / indicateur de silhouette
- Approche valeurs manquantes
  - Décomposition SVD
  - SGD, ALS

Jour 3 - Apprentissage supervisé

- Régression lin2aire
  - Formulation, condition d'usage
- Analyse de performance, pvalue, détection de performance
  - Notion d'overfitting
  - R2 et R2 ajusté
- Sélection de feature : approche forward, stepwise
- Approche pénalisée
  - Ridge, Lasso, élastic net.
  - Interprétation géométrique
- Arbres de décision
  - Principe de construction
  - Prunning
  - Interprétation, contexte d'exploitation
- Random Forest
  - Comment dépasser les limites de l'arbre de décision
  - Feature importance, importance locale
- Gradient boosting
  - Principes
  - Réglages
- XGBosst (extreme gradient boosting)
  - Principes, réglages

Jour 4 - Réglage fin et dépouillement de modèle

- Approfondissement des techniques de réglage de modeles
  - Fonctions de coût, RMSE, courbe roc et indicateur auc
  - Précautions de réglage, pièges à éviter
- Dépouillement de modèle
  - Où était l'information ?
  - Simplifier le modèle, sélection de feature avancée

Initiation text mining et NLP

- Lois de Heaps et de Zipf
- Comment structurer une source non structurée
  - Approche bag of words
  - Stop word et normalisation TF IDF
- Vers le NLP (natural language processing)
  - Analyse sémantique
  - Approche deep learning

Jour 5 - Initiation Deep Learning

- Réseaux de neurone
- Architecture de réseaux
  - Convolution
  - LSTM
- Découverte de l'environnement Keras pour déployer

Gestion de projet

- Les différentes phases d'un projet data
- Adaptation de la gestion de projet Agile aux projets data
- Structurer le dialogue data science / métier
- Piloter le projet
- Comment faire émerger les projets ? Quand s'arrêter ?

Sociétés concernées

Cette formation s'adresse aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiple permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.