

Data Classification et Machine Learning

Détails

- Code : DB-MCH
- Durée : 2 jours (14 heures)

Public

- Chefs de projets
- Architectes
- Développeurs
- Data Scientists

Pré-requis

- Connaissances des principes du Big Data et des architectures techniques mises en œuvre

Objectifs

- Savoir définir les étapes de préparation des données
- Comprendre et mettre en œuvre l'apprentissage automatique, les techniques de classification de données, les apports des réseaux de neurones et du Deep Learning

Programme

Introduction

- Zoom sur les données : format, volumes, structures, ...
- Requêtes et attentes des utilisateurs
- Etapes de la préparation des données
- Définitions, présentation du data munging
- Le rôle du data scientist

Gouvernance des données

- Qualité des données
- Transformation de l'information en donnée
- Qualification et enrichissement
- Sécurisation et étanchéité des lacs de données
- Flux de données et organisation dans l'entreprise
- De la donnée maître à la donnée de travail
- MDM
- Mise en œuvre pratique des différentes phases : nettoyage, enrichissement, organisation des données

Traitements statistiques de base

- Introduction aux calculs statistiques
- Paramétrisation des fonctions
- Applications aux fermes de calculs distribués
- Problématiques induites
- Approximations
- Précision des estimations

Data Mining

- Besoin, apports et enjeux
- Extraction et organisation des classes de données
- Analyse factorielle

Machine Learning

- Apprentissage automatique

- Définition, les attentes par rapport au Machine Learning
- Les valeurs d'observation et les variables cibles
- Ingénierie des variables
- Les méthodes : apprentissage supervisé et non supervisé
- Classification des données
- Algorithmes : régression linéaire, k-moyennes, k-voisins, classification naïve bayésienne, arbres de décision, forêts aléatoires, ...
- Création de jeux d'essai, entraînement et construction de modèles
- Prévisions à partir de données réelles
- Mesure de l'efficacité des algorithmes
- Courbes ROC
- Parallélisation des algorithmes
- Choix automatique

IA

- Introduction aux réseaux de neurones
- Réseaux de neurones à convolution
- Modèles de CNN
- Les types de couches : convolution, pooling et pertes
- L'approche du Deep Learning
- DeepLearning4j sur Spark

Les risques et écueils

- Importance de la préparation des données
- L'écueil du "surapprentissage"

Visualisation des données

- L'intérêt de la visualisation
- Outils disponibles
- Exemples de visualisation avec R et Python

Modalité

- Stage pratique en présentiel
- Stage pratique en distanciel

- Nombre de stagiaires minimum : 4
- Nombre de stagiaires maximum : 10

Méthodes pédagogiques

- Exposés
- Cas pratiques

Profils des intervenants

- Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

Modalités d'évaluation

- Evaluation des acquis de la formation par le biais de cas pratiques et/ou mises en situation.
- Attestation de formation remise à chaque participant.

Démarche qualité

- Questionnaire d'évaluation de satisfaction à chaud complété par chaque participant à l'issue de la formation.

Moyens pédagogiques

- Salle équipée de PC (1 poste par stagiaire), vidéo-projecteur.
- Espace de pause.

Dernière mise à jour le 08/12/2020