Pythagore F.D.
Apprendre à apprendre

CB050

Data Classification et Machine Learning

Durée: 2 jours

1320 €	12 au 13 mars	20 au 21 septembre
	31 mai au 1er juin	29 au 30 novembre

Public:
Chefs de projet, développeurs, data scientists, architectes souhaitant comprendre comment organiser le traitement des données et structurer les processus de Machine Learning.

Objectifs:
Savoir définir les étapes de préparation des données, comprendre et mettre en oeuvre l'apprentissage automatique, les techniques de classification de données, les apports des réseaux de neurones et du Deep Learning.

Connaissances préalables nécessaires:
Connaissances des principes du BigData, et des architectures techniques mises en oeuvre.

Programme:

Introduction : Zoom sur les données : format, volumes, structures, ...
et les requêtes, 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 oeuvre 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.

Pythagore F.D.

Apprendre à apprendre

CB050

Data Classification et Machine Learning

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, etc .. 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