?

FL9BD01 **11 500 €** **26 jour(s)**

CES Data Scientist / Data Science : Analyse et gestion de grandes masses de données

OBJECTIFS

- Mettre en œuvre les techniques récentes de gestion et d'analyse de grandes masses de données.
- Identifier et prendre en compte les différents formats des données, modèles, méthodes d'extraction de descripteurs (features) structurels et sémantiques.
- Utiliser et adapter les algorithmes et les techniques d'analyse des données et d'apprentissage statistique.
- Prendre en compte les problématiques de volumétrie et mettre en œuvre les techniques de passage à l'échelle.

PROGRAMME

Introduction à l'apprentissage statistique

- Objectifs et enjeux de l'apprentissage statistique
- Nomenclature des problèmes
- Formalisme probabiliste
- Régression logistique - loi/vraisemblance conditionnelle - Newton Raphson
- Analyse discriminante linéaire/quadratique
- Le perceptron de F. Rosenblatt
- Méthode des *k*-plus proches voisins

Bases de données NoSQL

- Concepts de base autour des bases de données distribuées
- MapReduce
- Bases de données clés-valeurs
- Bases de données orientées colonne
- Bases de données orientées document
- Bases de données orientées graphe
- Flux de données

Extraction d'informations du web

- Reconnaissance d'entités nommées
- Désambiguation
- Fact extraction
- Web sémantique

Données multimédia

- Initiation à l'indexation des images
- Initiation à l'indexation des sons
- Étude de cas

DATES ET LIEUX

Nous contacter pour les sessions à venir

PUBLIC / PREREQUIS

Cette formation s'adresse à des ingénieurs, chefs de projet avec des bonnes connaissances en mathématiques (probabilités, optimisation, algèbre linéaire) et une bonne expérience de la programmation, souhaitant évoluer vers un poste de Data Scientist, Data Analyst ou ingénieur Big Data.

De bonnes connaissances en mathématiques (optimisation, probabilités/statistique, algèbre linéaire) et une bonne expérience de la programmation sont indispensables pour suivre avec profit cette formation (voir MOOC Fondamentaux pour Big Data).

COORDINATEURS

Anne SABOURIN

Maître de conférence au sein du Département "Image, Données et Signal" de Télécom Paris, elle consacre ses recherches à l'apprentissage statistique sur des événements rares. Les applications directes de ses recherches concernent la détection d'anomalie et la gestion des risques liés aux valeurs extrêmes.

Fabian SUCHANEK

Professeur à Télécom Paris. Il a fait ses recherches à l'Institut Max Planck en Allemagne, chez Microsoft Research Cambridge/UK, chez Microsoft Research Silicon Valley/USA, et à l'INRIA Saclay. Il est l'auteur principal de YAGO, une des

Apprentissage supervisé : de la théorie aux algorithmes

- Éléments de la théorie de Vapnik-Chervonenkis
- Arbres de décision
- Réseaux de neurones
- Support Vector Machines
- Boosting
- Lasso
- Apprentissage par renforcement

Techniques avancées pour l'apprentissage : Noyaux et Deep Learning

- Apprentissage en ligne
- Apprentissage statistique distribué
- Techniques d'échantillonnage

Apprentissage non supervisé

- Variables latentes
- Clustering
- Analyse des affinités
- Détection d'anomalies

Réseaux bayésiens/HMM

- Chaînes de Markov cachées
- Réseaux bayésiens

Visualisation de données

- Principes de base de la visualisation d'information
- Critique des techniques de visualisation appliquées à une donnée particulière pour une tâche donnée
- Évaluation des systèmes de visualisation
- Conception de nouveaux outils de visualisation

Stockage à l'échelle du Web

- SGBD relationnels distribués classiques
- Systèmes de fichiers distribués HDFS/GFS
- Stockage à grande échelle
- Stockage clés-valeurs par table de hachage distribuée (Dynamo)
- Stockage par arbre distribué (BigTable, HBase)
- Systèmes NewSQL (Google Spanner, SGBD en mémoire, MySQL Cluster)

Calcul distribué

- MapReduce avancé
- Au-delà de MapReduce : Spark, Stratosphere
- Message Passing Interface
- Calculs distribués sur des graphes : GraphLab, Pregel, Giraph

Apprentissage distribué - Fouille de graphes

- Distribution d'algorithmes d'indexation, d'apprentissage et de fouille

plus grandes bases de connaissances publiques dans le monde.

MODALITES PEDAGOGIQUES

Suivre le MOOC gratuit « Fondamentaux pour le Big Data » en prérequis de la formation
Cours et travaux pratiques
Mini-projets inter-sessions
Retours d'expérience de professionnels

- Index inversé
- Factorisation de matrice
- Échantillonnage
- PageRank

Retour sur la méthodologie du Machine Learning