FL9BD01 **12 500 €** **26 jour(s)**

[Formation certifiante] 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 pour exercer le métier de data scientist
- 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

ATELIER

DATES ET LIEUX

Nous contacter pour les sessions à venir

PUBLIC / PREREQUIS

Ingénieur ou chef de projet souhaitant développer vos compétences dans le domaine de la gestion et l'analyse statistique des données massives pour 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 la formation MOOC Fondamentaux pour Big Data).

COORDINATEURS

Pietro GORI

Enseignant-chercheur au département Image, Données, Signal de Télécom Paris et au laboratoire LTCI. Ses recherches portent principalement sur l'anatomie computationnelle, l'analyse des formes, l'apprentissage statistique et l'imagerie médicale.

Fabian SUCHANEK

Enseignant-chercheur à Télécom Paris. Il a fait ses recherches à l'Institut Max Planck en Allemagne, chez Microsoft Research Cambridge/UK, chez Microsoft

- Désambiguation
- Fact extraction
- Web sémantique

Données multimédia

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

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
- Réseaux de neurones (ANN, CNN)
- Traitement d’images

Apprentissage non supervisé

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

Réseaux HMM / représentation vectorielles et modèles séquentiels

- Chaînes de Markov cachées
- Représentations vectorielles et modèles séquentiels pour le traitement du langage

Traitement du langage naturel

- Tokenisation
- Marquage de partie de discours
- Représentation de document
- Word Embeddings
- WordNet

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

Research Silicon Valley/USA, et à l’INRIA Saclay. Il est l’auteur principal de YAGO, une des plus grandes bases de connaissances publiques dans le monde.

- 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
- Index inversé
- Factorisation de matrice
- Échantillonnage
- PageRank

Retour sur la méthodologie du machine learning

Synthèse et conclusion