

Pensez comme un datascientist

Data science et Machine Learning

Durée : 4 jours

Réf : ML

Méthode pédagogique

Composée à 70% de pratique, cette formation utilise des exercices illustrés et didactiques.

Une évaluation quotidienne de l'acquisition des connaissances de la veille est effectuée.

Une synthèse est proposée en fin de formation.

Une évaluation à chaud sera proposée au stagiaire à la fin du cours. Un support de cours sera remis à chaque participant comprenant les slides sur la théorie, les exercices. Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

Un suivi et un échange avec les participants seront mis en place quelques jours après la formation.

Présentation

Plus d'un tiers des ventes du site Amazon.com est généré grâce à leur moteur de recommandation. Leurs algorithmes de Machine Learning leur a aussi révélé que les recommandations sont d'autant plus efficaces que lorsqu'elles sont faites lors d'une campagne mail plutôt qu'un affichage sur la page web.

Segmentation, prédiction, estimation, recommandation, ces méthodes d'analyse sont aujourd'hui utilisées dans l'objectif d'apporter de la valeur aux données détenues par l'entreprise.

Si l'implémentation de ces algorithmes reste de la responsabilité du mathématicien, leurs usages est maintenant rendu possible par l'apparition de bibliothèques masquant une partie de leur complexité. A l'issue de cette formation vous aurez appris les activités du datascientist : Choix et préparation des données, sélection des algorithmes, apprentissage, scoring, industrialisation.

Objectifs

- Comprendre le Machine Learning
- Catégoriser les différentes approches : clustering, classification, régression...
- Choisir, structurer et adapter les données pertinentes pour des résultats pertinents
- Apprendre le langage python et ses bibliothèques scientifiques (scikit-learn, pandas)
- Mettre en œuvre des cas concrets (prédiction, clustering)
- Mesurer la pertinence des modèles mis en œuvre
- Traiter des gros volumes de données en parallélisant les traitements : Spark et Dask
- Déployer un modèle en production
- Introduction au Deep Learning

Audience

Développeurs, Chefs de projet, Architectes

Prérequis

La connaissance d'un langage de programmation structuré est nécessaire

Le formateur

Le formateur est un expert du domaine qui intervient sur le sujet depuis plusieurs années en formation mais aussi en conseil. Doté d'une grande qualité d'écoute, sa pédagogie et sa compétence technique vous permettront d'acquérir les compétences sur le sujet. Il saura alterner entre théorie, pratique, et retours d'expérience.

Programme

Les bases du machine learning

- Quelques exemples en guise d'introduction
- Généraliser : un principe fondateur du Machine Learning pour permettre la prédiction et la segmentation
- Algorithmes supervisés et non supervisés
- Appropriation du vocabulaire du Data Scientist

Premiers pas

- Collecter et préparer les données
- Analyser, comprendre, nettoyer puis structurer les données : feature engineering
- Apprentissage et création d'un modèle

- Evaluation du modèle
- Amélioration du modèle

Le langage python

- Syntaxe
- Jupyter notebook
- Les librairies du Data Scientist : Pandas, scikit-learn
- Traitements statistiques pour analyser et comprendre les données
- Matplotlib et Seaborn : des librairies de datavisualisation pour Python

Choisir les algorithmes de machine learning

- Comprendre les enjeux
- Tour d'horizon des principaux algorithmes
- Classification : k-Nearest Neighbors (k-NN), Arbre de décision, Random Forest, XGBoost
- Régression : SVM, Ridge Regression
- Clustering : K-Means, DBScan
- Scoring

Concepts avancés

- Pipeline
- Validation croisée
- Ensemble Machine Learning : cumulez les algorithmes pour une meilleure précision

Traiter les données en parallèle

- Pourquoi paralléliser ?
- Adapter les algorithmes
- Une complexité complémentaire
- Les frameworks de distribution à disposition : Spark et Dask

Déployer en production

- Intégrer un pipeline à une chaîne de déploiement automatisée (continuous delivery)
- Packager un modèle : Predictive Model Markup Language
- Créer un endpoint REST avec python flask
- Déployer dans le cloud

Le Deep Learning

- Présentation générale : les réseaux de neurones
- Comprendre par l'exemple : Pytorch & Keras