Best

Les bases de l'apprentissage Machine (Machine Learning)

Transformer des volumes massifs de données en informations utiles

Présentiel ou en classe à distance

Durée : 3 jours (21 h)

Réf. : BI105

Prix inter : 2.205,00 € HT

Forfait intra : 7.420,00 € HT

La maitrise du Data Mining et du Machine Learning est devenue une compétence nécessaire, voire même indispensable à toute personne souhaitant développer une expertise Big Data puisqu'elle permet d'explorer ou de fouiller de très importants volumes de données pour construire des modèles et répondre aux problèmes très variés des entreprises et organisations lorsque les méthodes statistiques traditionnelles deviennent inopérantes. Pour cela, les experts en Big Data doivent maitriser l'élaboration et l'étude des algorithmes permettant à des machines d'apprendre automatiquement à partir des données et à effectuer des tâches de façon autonome pour modéliser des tendances.

Les objectifs de la formation

- Comprendre les différences entre apprentissage automatique supervisé, non supervisé et méta-apprentissage
- Savoir transformer un gros volume de données à priori hétérogènes en informations utiles
- Maîtriser l'utilisation d'algorithmes d'auto-apprentissage adaptés à une solution d'analyse
- Comprendre comment exploiter de gros volumes de données textuelles
- Être capable d'appliquer ces différentes techniques aux projets Big Data

A qui s'adresse cette formation ?

Pour qui

- Ingénieurs, analystes, responsables marketing
- Data Analysts, Data Scientists, Data Steward
- Toute personne intéressée par les techniques de Data Mining et de Machine Learning

Prérequis

- Connaître l'utilité du Data Mining et les problématiques du Big Data dans le ciblage économique
- Disposez-vous des compétences nécessaires pour suivre cette formation ? Testez-vous !

Programme

1 - L'apprentissage machine (Introduction)

- Introduction
- Champs de compétences
- Focus Data Science (Data Mining)
- Focus Machine Learning

- Focus Big Data
- Focus Deep Learning
- Définition de l'apprentissage machine
- Exemples de tâches du machine Learning
- Que peuvent apprendre les machines
- Les différents modes d'entraînement

2 - Les fondamentaux de l'apprentissage machine

- Préambule : - Un problème d'optimisation - Quête de la capacité optimale du modèle - Relation capacité et erreurs - Un apport philosophique - Cadre statistique - Anatomie d'un modèle d'apprentissage machine
- Jeux de données d'entraînement : - Cadre statistique - Les variables prédictives - Chaîne de traitement des variables prédictives - Les variables à prédire
- Fonctions hypothèses : - Principe : jeux de fonctions hypothèses - Contexte de sélection des fonctions hypothèses - Caractéristiques des fonctions hypothèses - Modèles probabilistes Fréquentistes et Bayésiens
- Fonctions de coûts : - Les estimateurs - Principe du maximum de vraisemblance (MLE*) - MAP - Maximum A Posteriori - Le biais d'un estimateur - La variance d'un estimateur - Le compromis biais - variance - Les fonctions de coûts - La régularisation des paramètres
- Algorithmes d'optimisations : - Les grandes classes d'algorithmes d'optimisation - La descente de gradient (1er ordre) - Descente de gradient (détails) - Les approches de Newton (2nd ordre) - Optimisation batch et stochastique - Pour aller plus loin
- Lab : Mise en oeuvre de l'environnement de travail machine Learning

3 - La classification

- Introduction : - Choisir un algorithme de classification
- La régression logistique : - Du Perceptron à la régression logistique - Hypothèses du modèle - Apprentissage des poids du modèle - Exemple d'implémentation : scikit-learn - Régression logistique - Fiche Synthèse
- SVM : - Classification à marge maximum - La notion de marge souple (soft margin) - Les machines à noyau (kernel machines) - L'astuce du noyau (kernel trick) - Les fonctions noyaux - SVM - Maths - SVM - Fiche Synthèse
- Arbres de décision : - Principe de base - Fonctionnement - Maximisation du Gain Informationnel - Mesure d'impureté d'un noeud - Exemple d'implémentation : scikit-learn -Arbres de décision - Fiche Synthèse
- K plus proches voisins (kNN) : - L'apprentissage à base d'exemples - Principe de fonctionnement - Avantages et désavantages - kNN - Fiche synthèse
- Synthèse
- Lab : Expérimentation des algorithmes de classification sur cas concrets

4 - Les pratiques

- Prétraitement : - Gestion des données manquantes - Transformateurs et estimateurs - Le traitement des données catégorielles - Le partitionnement des jeux de données - Mise à l'échelle des données
- Ingénierie des variables prédictives (Feature Engineering) : - Sélection des variables prédictives - Sélection induite par régularisation L1 - Sélection séquentielle des variables - Déterminer l'importance des variables - Réduction dimensionnelle par Compression des données - L'extraction de variables prédictives - Analyse en composante principale (ACP) - Analyse linéaire discriminante (ADL) - l'ACP à noyau (KPCA)
- Réglages des hyper-paramètres et évaluation des modèles : - Bonnes pratiques - La notion de Pipeline - La validation croisée (cross validation) - Courbes d'apprentissage - Courbes de validation - La recherche par grille (grid search) - Validation croisée imbriquée (grid searchcv) - Métriques de performance
- Synthèse
- Lab : Expérimentation des pratiques du machine learning sur cas concrets

5 - L'apprentissage d'ensembles (ensemble learning)

- Introduction

- L'approche par vote
- Une variante : l'empilement (stacking)
- Le bagging
- Les forêts aléatoires
- Le boosting
- La variante Adaboost
- Gradient Boosting
- Fiches synthèses
- Lab : L'apprentissage d'ensemble sur un cas concret

6 - La régression

- Régression linéaire simple
- Régression linéaire multi-variée
- Relations entre les variables
- Valeurs aberrantes (RANSAC)
- Évaluation de la performance des modèles de régression
- La régularisation des modèles de régression linéaire
- Régression polynomiale
- La régression avec les forêts aléatoires
- Synthèse
- Lab : La régression sur un cas concret

7 - Le clustering

- Introduction
- Le regroupement d'objets par similarité avec les k-moyens (k-means)
- k-means : algorithme
- L'inertie d'un cluster
- Variante k-means ++
- Le clustering flou
- Trouver le nombre optimal de clusters avec la méthode Elbow
- Appréhender la qualité des clusters avec la méthode des silhouettes
- Le clustering hiérarchique
- Le clustering par mesure de densité DBSCAN
- Autres approches du Clustering
- Synthèse
- Lab : Le clustering sur un cas concret

Evaluation

- Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Les points forts de la formation

- Une formation très pratique : 70% du temps de la formation est dédié à la mise en pratique pour une meilleure assimilation de notions de base.
- Cette formation est basée sur des exercices principalement proposés par le formateur et tirés de l'ouvrage qui sert de support pour la formation.
- Les travaux pratiques sont principalement réalisés avec R et Python.
- Des consultants expérimentés partagent leur savoir-faire avec les participants.

- 100% des participants à cette formation se sont déclarés satisfaits ou très satisfaits au cours des 12 derniers mois.

Dates et villes 2024 - Référence BI105

A distance

du 27 mai au 29 mai

du 15 juil. au 17 juil. Session garantie

du 7 oct. au 9 oct. Session garantie

du 9 déc. au 11 déc.

Paris

du 27 mai au 29 mai

du 15 juil. au 17 juil. Session garantie

du 7 oct. au 9 oct. Session garantie

du 9 déc. au 11 déc.