Big Data Analytics avec Python

Durée

4 jours.

Participants

Ingénieurs, Data Analyst, Programmeurs désirant s'initier au machine learning.

Prérequis

Connaissances de base en Python. Connaissances de base en statistiques ou avoir suivi le stage "Statistiques, maîtriser les fondamentaux" (Réf. STA).

Objectifs pédagogiques

- Comprendre le principe de la modélisation statistique
- Choisir entre la régression et la classification en fonction du type de données
- Évaluer les performances prédictives d'un algorithme
- Créer des sélections et des classements dans de grands volumes de données pour dégager des tendances

Méthodes pédagogiques

Ce séminaire se base sur des présentations, des échanges et des études de cas. Des outils comme Lasagne ou Keras seront présentés.

Description

Data Analytics est un terme pour exprimer les démarches d'analyse de données, afin d'être en mesure de prendre des décisions. Le langage Python dispose d'un écosystème permettant les traitements statistiques : de la construction de modèles d'analyse, à leur évaluation jusqu'à leur représentation.

Programme

Introduction à la modélisation

- Introduction au langage Python.
- Introduction au logiciel Jupiter Notebook.
- Les étapes de construction d'un modèle.
- Les algorithmes supervisés et non supervisés.
- Le choix entre la régression et la classification.

Travaux pratiques Réflexion sur des problématiques des stagiaires. Quel type de données sont à disposition ? Quelle est la tâche à réaliser ? Quels modèles devront être utilisés ?

Gestion des données & Pré-traitement

- Utiliser PySpark pour la gestion des données.
- Utiliser des expressions régulières pour restructurer rapidement des données textuelles.

Travaux pratiques Récupérer un jeu de données et le modifier avec PySpark

Pré-traitementProcédures d'évaluation de modèles

- Les techniques de ré-échantillonnage en jeu d'apprentissage, de validation et de test.
- Test de représentativité des données d'apprentissage.
- Mesures de performance des modèles prédictifs.
- Matrice de confusion, de coût, la courbe ROC/AUC et courbe de lift.
- Utiliser PySpark pour le nettoyage et la préparation de données.
- Utiliser des expressions régulières pour restructurer rapidement des données textuelles.

Travaux pratiques Mise en place d'échantillonnage de jeux de donnes. Effectuer des tests d'évaluations sur plusieurs modèles fournis.

Les algorithmes supervisés : La régression

- Le principe de régression linéaire univariée.
- La régression multivariée.
- La régression polynomiale.
- La régression régularisée.

Travaux pratiques Mise en œuvre des régressions et des classifications sur plusieurs types de données en utilisant sklearn et XGBoost

Les algorithmes supervisés : La classification

- Régression logistique
- SVM
- Forêts aléatoires
- Techniques boostés

Travaux pratiques Mise en œuvre des régressions et des classifications sur plusieurs types de données en utilisant sklearn et XGBoost

Les algorithmes non supervisés : le clustering

- Cas d'utilisations du clustering.
- Le clustering K-means et dérivés.
- Le clustering hiérarchique.

- Les approches mixtes.

Travaux pratiques Traitements de clustering non supervisés sur plusieurs jeux de données.

Les algorithmes non supervisés : Réduction de dimensions

- Cas d'utilisations de la réduction de dimensions.
- Analyse en composantes principales.
- Classification hiérarchique sur composantes principales.
- Réductions non linéaires

Travaux pratiques Mise en œuvre de la diminution du nombre des variables et identification des facteurs sous-jacents des dimensions associées à une variabilité importante.

Les algorithmes non supervisés : Détection d'anomalies

- Cas d'utilisations de la détection d'anomalies.
- Détection d'anomalies en variance
- Technique basée sur les arbres : isolation forest
- Technique basée sur le voisinage : Local outlier factor

Travaux pratiques Mise en œuvre de la détection d'anomalies pour nettoyer un jeu de données.