+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 29/11/2023

S'inscrire

Formation Analyse des Clusters (Partitionnement de données) avec Python

4 jours (28 heures)

Présentation

La clusterisation est l'une des méthodes incontournables de l'analyse de données. En effet, regrouper ces données en groupes homogènes et distincts peut bénéficier de nombreux domaines comme la santé, le marketing ou encore la finance.

En marketing, la création de clusters (aussi appelé segments) permet de catégoriser chaque client. Cette catégorisation a un effet positif sur les performances de vos campagnes, car vos messages sponsorisés seront personnalisés pour votre groupe cible.

La clusterisation peut également être utile pour la détection de fraudes grâce à la reconnaissance visuelle. Ce système est intéressant, notamment pour la reconnaissance de signature dans le domaine de la cybersécurité ou en finance.

Afin de réaliser nos partitions de données, nous utiliserons l'un des langages de programmation les plus utilisés dans le monde, Python. Grâce à sa librairie scikit-learn, Python possède toutes les fonctions pour la création efficace de clusters de données.

Notre formation analyse des clusters vous initiera à la programmation sur python pour l'analyse de données, vous connaîtrez l'intérêt et les cas d'usage des méthodes de clusterisation. À la fin de ce cours, vous saurez créer des clusters ainsi que les analyser avec Python.

Comme toujours, notre formation s'appuiera sur la dernière version en date du langage, Python 3.10.

Objectifs

- Utiliser Python pour l'analyse de données

- Comprendre l'utilité de la clusterisation
- Connaître les principaux types d'algorithmes de clusterisation
- Savoir préparer ses données avec Python
- Savoir représenter et analyser ses clusters

Public visé

- Data Analyst
- Data Scientist
- Data Engineer
- Machine learning engineer
- Chef d'entreprise
- Analyste
- Chargé de marketing

Pré-requis

Connaissances en mathématiques générales (probabilités, statistiques…).

Programme de notre formation Analyse des Clusters de Données

Introduction

- Qu'est-ce qu'un cluster ?
- La différence entre clusterisation et segmentation
- L'intérêt de l'analyse des clusters, les cas d'usage
- Les limites et les défis de la clusterisation

Les différentes méthodes de partitionnement de données

- K-means
- Mean-Shift
- Le DBSCAN (Regroupement spatial d'applications avec du bruit basé sur la densité)
- Algorithme espérance maximisation avec ou sans des modèles de mélange Gaussien (GMM)
- Regroupement hiérarchique

Présentation de Python

- Pourquoi utiliser Python ?
- Présentation de la librairie Scikit learn
- Utiliser des fonctions de librairies
- Gérer les modules et librairies

Commencer la programmation avec Python

- La syntaxe de Python
- Les variables
- Les différents types d'ensembles de données
  - Tuple
  - Liste
  - Set
  - Dictionnaire
- Les fonctions
- Écrire ses propres fonctions

Bien préparer ses données avec Python

- L’importance d’avoir des données intègres et préparées
- Lire et modifier des fichiers CSV
- Importer ses données
- Nettoyer et préparer ses données
- Formatage des données
- Construire des pipelines de données

K-Means Clustering

- Importer les modules sklearn
- Importer ses données
- Les paramètres
  - n_samples
  - centers
  - cluster_std
- La fonction make_blobs()
- Utiliser la standardisation
- Utiliser la fonction KMeans
- Les méthodes pour choisir le bon nombre de clusters
- Représenter les clusters graphiquement

Mean-Shift

- Importer MeanShift et make _blobs
- Déterminer les centres du cluster
- Représenter les données en 3D

DBSCAN

- Importer ses données
- Description des paramètres
- Clusteriser ses données
- Représenter ses regroupements de données graphiquement

Espérance-maximisation

- Concaténer des courbes gaussiennes
- Explication de l'algorithme d'espérance maximisation
- Représenter graphiquement ses partitions de données

Regroupement hiérarchique

- Préparer les données
- Calculer les informations de similarité entre chaque donnée
- Utiliser une fonction de liaison
- Déterminer la coupure de l'arbre hiérarchique

Analyser ses résultats

- Méthodes de validation des partitions
- Évaluation de la mise en grappes
- Améliorer ces clusters
- Mise en grappe basée sur les contraintes
  - Mesures établies sur l'appariement
  - Mesures basées sur l'entropie
  - Mesures par paires
- Mesures internes pour valider ses clusters
- La stabilité des grappes

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des

séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.