Intitulé du Master : Science de Données et Intelligence Artificielle

Semestre : S1 - Unité d'Enseignement : UEF1

Intitulé de la matière : Data Mining 1 Code : DAMI1

Nombre d'heures d'enseignement : 63 H au total sur 14 semaines

Nombre de crédits : 6 - Coefficient de la Matière : 3

Cours : 1H30/semaine - TD : 1H30/semaine - TP : 1H30/semaine

Mode d'évaluation : Examen (60%), contrôle continu (40%)

Objectifs de l'enseignement

- Comprendre les principales tâches de data et comment ils peuvent être abordés par les différents algorithmes de data (clustering, classification, extraction de règles d'association)

- Comprendre le fonctionnement interne des algorithmes de data et sous quelles conditions ils devraient être utilisés.

- Maîtriser le langage python

- Être en mesure d'aborder des problèmes réels et complexes du data

Connaissances préalables recommandées :

Des notions de probabilités et statistiques.

Contenu de la matière :

<table>
<tr><td colspan="4">Introduction générale (Chapitre 1)</td></tr>
<tr><td colspan="4">Méthodes de prétraitement et préparation de données (Chapitre 2)</td></tr>
<tr><td rowspan="4">Techniques et méthodes du Data Mining</td><td rowspan="2">Supervisées (Prédictives)</td><td>Régression</td><td>• Régression linéaire (Simple et Multiple)
• Régression non linéaire
• Régression logistique</td></tr>
<tr><td>Classification (Chapitre 3)</td><td>• KNN
• Arbre de décision (ID3, CART …)
• Classification de bayes
• SVM</td></tr>
<tr><td rowspan="2">Non Supervisées (Descriptives)</td><td>Clustering (Chapitre 4)</td><td>• K-means
• Classification hiérarchique (Agglomerative, Divisive)
• Clustering par voisinage dense (DBSCAN)</td></tr>
<tr><td>Règles d’association (Chapitre 5)</td><td>• Apriori
• Fptree</td></tr>
</table>

Chapitre 1 : Introduction générale

Bienvenue dans le cours de Data Mining, une discipline fondamentale au cœur de l'intelligence artificielle et de l'analyse de données. Dans ce chapitre introductif, nous explorons les concepts clés du Data Mining, son importance croissante dans le paysage technologique actuel et son rôle essentiel dans l'ère de l'Intelligence Artificielle.

L'abondance de données numériques créées chaque jour dans le monde moderne a donné naissance à une nouvelle ère où les données deviennent un actif inestimable. Cependant, ces données ne sont que des chiffres et des caractères au moins qu'elles ne soient transformées en connaissances exploitables. C'est là qu'intervient le Data Mining.

1. Qu'est-ce que le Data Mining ?

Le Data Mining, également appelé fouille de données, est l'art et la science d'extraire des connaissances significatives et utiles à partir de grandes quantités de données. Il s'appuie sur des techniques provenant de divers domaines tels que les statistiques, l'apprentissage automatique, la visualisation et la base de données. L'objectif ultime du Data Mining est de découvrir des modèles cachés, des tendances, des relations et des informations précieuses à partir des données brutes.

Le Data Mining est un processus inductif, itératif et interactif, et il est accessible par sa capacité à découvrir des modèles de données valides, innovants, utiles et compréhensibles au sein de vastes bases de données. Ses principes fondamentaux comprennent :

- Inductif : En partant des données brutes, il déduit des modèles et des tendances, plutôt que de valider des hypothèses préexistantes.

Exemple : Imaginons que vous ayez une grande base de données contenant des enregistrements de ventes pour différentes régions et produits. En utilisant le Data Mining de manière inductive, vous pourriez découvrir que les ventes de produits sont liées aux saisons de l'année, même si vous n'aviez pas supposé cette relation au départ.

- Itératif : Le processus se répète, développe progressivement les modèles par des cycles d'exploration, de modélisation et d'évaluation.

Exemple : Supposons que vous travailliez pour une entreprise de marketing qui analyse les habitudes d'achat des clients. Au fil du temps, en appliquant itérativement des techniques de Data Mining, vous affinez vos modèles prédictifs pour mieux anticiper les comportements d'achat futurs, en ajustant régulièrement vos stratégies marketing en conséquence.

- Interactif : l'interaction entre expertise humaine et algorithmes guide le processus, avec des experts formulant des questions, interprétant des résultats et ajustant les paramètres.

Exemple : Dans le domaine médical, des experts ont travaillé avec des algorithmes de Data Mining pour identifier les facteurs qui contribuent aux maladies. Les experts traduisent des questions ciblées, interprètent les résultats des analyses et ajustent les paramètres des algorithmes en fonction de leurs connaissances spécifiques, aboutissant à des modèles plus précis et pertinents.

- Valides : Les modèles sont applicables à de nouvelles données avec une certaine fiabilité.

Exemple : Dans le secteur financier, les modèles de détection de fraude intégrés par Data Mining sont valides lorsqu'ils identifient de manière fiable des transactions suspectes dans de nouvelles données, permettant ainsi aux institutions de sécurité de réagir rapidement aux activités frauduleuses.

- Nouveaux : Ils présentent des informations non évidentes et inattendues.

Exemple : Prenons l'exemple du marketing en ligne, où l'application du Data Mining peut révéler que les utilisateurs qui achètent des produits de jardinage en ligne sont également enclins à acheter des produits de cuisine, une corrélation non évidente au départ.

- Utiles : Les modèles fournissent des connaissances concrètes pour la prise de décisions et la résolution de problèmes.

Exemple : Dans le domaine de la logistique, en utilisant le Data Mining, une entreprise de livraison peut optimiser ses itinéraires de livraison en fonction des modèles de trafic et des préférences des clients, ce qui permet d'économiser du temps et des ressources.

- Compréhensibles : Ils sont facilement interprétables par les êtres humains, favorisant leur utilisation pratique.

Exemple : Dans le secteur de la santé, les médecins peuvent utiliser des modèles de Data Mining pour prédire les risques de maladies chez les patients. Les modèles simples et compréhensibles aident les médecins à prendre des décisions éclairées et à fournir des soins personnalisés aux patients.

2. Rôle du Data Mining dans l'Intelligence Artificielle

L'Intelligence Artificielle vise à doter les systèmes informatiques de la capacité de penser, d'apprendre et de résoudre des problèmes de manière similaire à celle des êtres humains. Dans cette quête, le Data Mining joue un rôle crucial en fournissant les insights nécessaires pour alimenter les algorithmes d'apprentissage automatique, de

traitement du langage naturel, de vision par ordinateur et d'autres domaines de l'IA. Les modèles découverts par le Data Mining servent de fondement à la prise de décision automatisée et à la génération de prédictions précises.

3. Le Processus de Découverte de Connaissances (KDD) dans le Contexte du Data Mining

Le processus de découverte de connaissances (KDD), également appelé "knowledge data discovery", est une démarche globale qui enveloppe le Data Mining au sein de son champ d'action. Il représente une méthodologie systématique pour extraire des connaissances significatives à partir de vastes bases de données.

D'autre part, le Data Mining occupe une position cruciale dans le processus de découverte de connaissances. Ce processus, se déroule sous la forme d'une séquence itérative comprenant plusieurs étapes, allant de la préparation des données à la présentation des connaissances extraites :

- Nettoyage des données (pour supprimer le bruit et les incohérences)
- Intégration des données (fusion de sources de données multiples)
- Transformation des données (mise en forme adéquate pour l'analyse)
- Sélection des données (récupération des données pertinentes)
- Data Mining (application de méthodes intelligentes pour extraire des modèles)
- Évaluation des motifs (identification des motifs intéressants)
- Présentation des connaissances (recours à la visualisation et à la représentation pour communiquer les connaissances extraites)

Ces étapes mettent en lumière la relation du Data Mining avec le processus de découverte de connaissances, où il occupe une place essentielle en exposant des modèles dissimulés pour l'évaluation. En pratique industrielle, dans les médias et dans la recherche, le terme "Data Mining" est souvent employé pour désigner l'ensemble du processus de découverte de connaissances, peut-être en raison de sa brièveté. Par conséquent, nous adoptons une définition élargie du Data Mining : c'est le procédé qui consiste à découvrir des motifs et des connaissances intéressantes à partir de volumes conséquents de données, issues de diverses sources telles que les bases de données, les entrepôts de données, le Web, et bien d'autres.

4. Contexte et Importance :

Le data joue un rôle central dans le contexte actuel en raison de la profusion de données dans tous les secteurs. Les organisations collectent une quantité considérable d'informations provenant de diverses sources, comme les transactions commerciales, les interactions en ligne et les réseaux sociaux. Cependant, ces données brutes ne sont utiles que lorsqu'elles sont transformées en connaissances exploitables, et c'est là que le data entre en jeu. Les avancées technologiques ont facilité la collecte massive de données, allant des historiques d'achats aux données de production. Les entreprises qui maîtrisent le data gagnent un avantage concurrentiel en découvrant des modèles

cachés, ce qui les aide à prendre des décisions éclairées et à personnaliser l'expérience client. De plus, le data permet de créer des modèles prédictifs en analysant les données historiques pour anticiper les comportements futurs, comme évaluer le risque de crédit ou prédire les pannes d'équipement. En identifiant les tendances émergentes et en comprenant les fluctuations du marché, le data aide les organisations à s'adapter rapidement aux changements et à prendre des décisions plus éclairées.

5. Applications :

Le Data Mining est largement appliqué dans divers domaines pour extraire des connaissances significatives à partir de grandes quantités de données. Voici des exemples concrets d'application dans différents secteurs :

- Marketing : Il joue un rôle crucial en aidant les entreprises à comprendre les préférences des clients et à cibler leurs efforts. Par exemple, dans le secteur de la vente au détail, il analyse les historiques d'achats pour identifier les modèles de comportement, conduire à des recommandations de produits personnalisés, à l'optimisation des campagnes et à des ajustements de tarification basés sur les tendances.

- Finance : Dans ce domaine, le Data Mining est utilisé pour évaluer le risque de crédit, détecter la fraude et optimiser les investissements. Les institutions financières peuvent prédire la probabilité de remboursement en analysant les données historiques des emprunteurs. Il aide également à repérer des schémas insuffisants dans les transactions, contribuant à la détection de fraudes.

- Santé : Il contribue à l'analyse des dossiers médicaux et à la recherche de modèles de maladies. L'analyse des données médicales des patients permet d'identifier des facteurs de risque, facilitant la prévention et le traitement précoce. Il peut également prédire les épidémies et évaluer l'efficacité des traitements.

- Sciences sociales et comportementales : Le Data Mining analyse les interactions en ligne, les médias sociaux et les enquêtes pour comprendre les tendances et les comportements. Par exemple, il surveille les opinions des clients, détecte les tendances émergentes et adapte les stratégies de marketing en conséquence.

- Sciences et recherche : Il est utilisé pour analyseur de grandes quantités de données expérimentales ou de simulations. En astronomie, il peut identifier de nouveaux types de galaxies ou détecter des événements astronomiques rares à partir de données de télescopes.

6. Processus de Data Mining :

Le processus de data comporte plusieurs étapes essentielles qui permettent de transformer des données brutes en informations exploitables. Voici les étapes typiques du processus de datamining :

- Collecte des données : Rassembler des données provenant de diverses sources, qu'elles soient structurées ou non, et les organiser en vue d'une analyse ultérieure.

- Préparation des données : Nettoyer, filtrer et transformer les données pour éliminer les valeurs aberrantes, gérer les données manquantes, normaliser les échelles et préparer les données pour les analyses.

- Exploration des données : Appliquer des techniques d'exploration pour rechercher des tendances, des motifs et des relations initiales dans les données, ce qui peut orienter le choix des méthodes de Data Mining.

- Modélisation des données : Utiliser des techniques et des algorithmes de Data Mining tels que la régression, la classification, le clustering, etc., pour construire des modèles prédictifs ou des regroupements basés sur les données.

- Évaluation des modèles : Tester la qualité et la validité des modèles créés en utilisant des données non utilisées précédemment, afin de s'assurer qu'ils généralisent bien au-delà des données d'entraînement.

- Déploiement et interprétation : Intégrer les modèles dans des applications ou des décisions concrètes, et interpréter les résultats pour en tirer des connaissances exploitables pour l'entreprise ou le domaine d'application.

7. Difficultés et Défis :

L'application du data comporte divers défis nécessitant une approche méthodique pour des résultats précis :

- Qualité des données : Les données brutes peuvent être erronées, contenir des doublons ou des valeurs aberrantes. La fiabilité des résultats dépend des données de qualité, incluant un nettoyage minutieux.

- Surapprentissage : Des modèles trop complexes créés lors de l'entraînement peuvent mener à une mauvaise généralisation, nécessitant des méthodes de régularisation et de validation croisée.

- Complexe d'interprétation : Modèles de complexes d'apprentissage automatique générés par des résultats difficiles à comprendre. L'interprétation est cruciale pour des décisions éclairées, nécessitant des méthodes explicatives et de visualisation.

- Sélection de variables : Le choix entre variables peut être difficile, car trop ou trop peu peut nuire aux modèles. Équilibrer l'inclusion et l'exclusion est un défi.

- Biais des données : Biais inhérents dans les données pouvant entraîner des modèles injustes. Identifier et corriger les biais est vital pour des résultats équitables.

- Gestion des données volumineuses : l'ampleur des données peut poser des problèmes de stockage et de traitement. Des technologies distribuées et parallèles sont utilisées pour gérer cette masse.

- Choix des méthodes : Sélectionner des méthodes d'analyse adaptées à chaque problème est complexe. Comprendre avantages et limites est essentiel pour l'adaptation aux données et aux objectifs.

8. Techniques et Méthodes :

8.1. Supervisées (Prédictives) :

Apprentissage supervisé est une approche d'apprentissage automatique où l'algorithme est entraîné sur un ensemble de données étiquetées, ce qui signifie que les données d'entraînement comprennent à la fois les caractéristiques d'entrée et les étiquettes de sortie. L'objectif est de créer un modèle capable de prédire les étiquettes des nouvelles données en se basant sur les informations apprises lors de l'entraînement. Par conséquent, on parle de tâches de prédiction.

Méthodes de l'apprentissage supervisé :

- Classification : La classification consiste à attribuer des étiquettes prédéfinies à des objets ou à des exemples en fonction de leurs caractéristiques. C'est utile pour la catégorisation, comme la prédiction de la classe d'appartenance d'une observation inconnue. Par exemple, la classification peut être utilisée pour prédire si un e-mail est un spam ou non.

- Régression : La régression vise à établir une relation entre une variable cible continue et des variables explicatives. Elle permet de prédire une valeur numérique basée sur les valeurs des autres variables. Par exemple, la régression peut être utilisée pour prédire le prix d'une maison en fonction de ses caractéristiques.

8.2. Non Supervisées (Descriptives) :

L'apprentissage non supervisé est une approche où l'algorithme est utilisé pour explorer la structure intrinsèque des données sans avoir d'étiquettes ou de réponses prédéfinies. Il est souvent utilisé dans l'analyse de données pour découvrir des schémas, des groupes ou des structures cachées dans les données. On parle de tâches descriptives car l'objectif principal est de décrire et d'analyser les données sans nécessairement prédire une variable cible.

Méthodes de l'apprentissage non supervisé :

- Regroupement (Clustering) : Le regroupement implique la création de groupes d'objets similaires en fonction de leurs caractéristiques. Cela aide à identifier les structures intégrées dans les données sans avoir de classes préétablies. Par exemple, le

regroupement peut être utilisé pour segmenter les clients en groupes similaires en fonction de leurs habitudes d'achat.

- Association : l'association cherche à découvrir des relations révélées entre les éléments d'un ensemble de données. Cela peut être utilisé pour révéler des associations intéressantes dans les données transactionnelles. Par exemple, l'association pourrait révéler quels produits sont souvent achetés ensemble dans un panier d'achat.

- Détection d'anomalies : La détection d'anomalies vise à identifier des observations inhabituelles ou des points de données aberrantes qui diffèrent du reste de l'ensemble de données. Cela peut être utilisé pour identifier des fraudes, des erreurs ou des incidents inhabituels dans les données.

- Prévision et séries chronologiques : Ces techniques ciblées à prédire les valeurs futures en fonction des tendances passées. Elles sont souvent utilisées pour des données séquentielles, telles que les données financières, météorologiques ou de ventes.

Chaque technique de data a ses propres forces et limites, et leur choix dépend du type de données, des objectifs de l'analyse et des questions spécifiques que vous essayez de résoudre. Une compréhension approfondie de ces techniques et de leurs applications est essentielle pour choisir la méthode la mieux adaptée à chaque situation.

9. Data mining et la machine learning :

L'apprentissage automatique (machine learning) et l'exploration de données (data mining) sont deux domaines interconnectés de l'informatique qui se concentrent sur l'extraction de connaissances utiles à partir de données. L'apprentissage automatique se focalise sur le développement de modèles et d'algorithmes qui permettent aux ordinateurs d'apprendre à partir de données et de prendre des décisions basées sur ces connaissances. En revanche, l'exploration de données se concentre sur la découverte de modèles, de tendances et de relations dans les données, souvent en utilisant des techniques statistiques et de traitement des données. Ces deux domaines se chevauchent, car l'apprentissage automatique peut être considéré comme une sous-discipline de l'exploration de données, où l'accent est mis sur la construction de modèles prédictifs. En fin de compte, ils travaillent ensemble pour aider à extraire des informations précieuses à partir de grandes quantités de données, ce qui est essentiel dans le monde actuel axé sur les données.

10. Outils et Technologies :

Il existe plusieurs outils, langages de programmation et bibliothèques qui sont largement utilisés dans le domaine du data pour faciliter l'analyse et l'extraction d'informations à partir des données. Voici quelques-uns des plus populaires :

10.1. Outils de Data Mining :

A. Rapidminer : Une plateforme open-source pour l'analyse prédictive et le data, offrant des fonctionnalités de traitement des données, de modélisation et de visualisation.

B. Weka : Un logiciel open-source avec une large gamme d'algorithmes pour le data et l'apprentissage automatique, ainsi que des outils de prétraitement des données.

C. Knime : Une plateforme open-source pour l'analyse de données et le data, permettant de créer des flux de travail personnalisés.

D. Orange : Un environnement de data visuel open-source qui permet aux utilisateurs de créer des analyses interactives à l'aide d'une interface conviviale.

10.2. Langages de Programmation :

A. Python : Un langage de programmation polyvalent avec de nombreuses bibliothèques d'apprentissage automatique, telles que Scikit-learn, tensorflow et Keras.

B. R : Un langage statistique largement utilisé dans l'analyse de données, avec de nombreuses bibliothèques pour le data et la visualisation, comme Caret et ggplot2.

10.3. Bibliothèques d'Apprentissage Automatique :

A. Scikit-learn : Une bibliothèque Python populaire pour l'apprentissage automatique, offrant une variété d'algorithmes pour la classification, la régression, le clustering, etc.

B. Tensorflow : Une bibliothèque d'apprentissage automatique open-source développée par Google, principalement utilisée pour les tâches liées au deep learning.

C. Pytorch : Une autre bibliothèque d'apprentissage automatique open-source axée sur le deep learning, largement utilisée pour la recherche et le développement de modèles avancés.

10.4. Logiciels de Base de Données et de Manipulation de Données :

A. SQL (Structured Query Language) : Utilisé pour la manipulation de données dans les bases de données relationnelles.

B. Pandas : Une bibliothèque Python pour la manipulation et l'analyse de données, souvent utilisée pour préparer les données avant l'analyse.

C. Les fichiers Excel : ils peuvent jouer un rôle dans la manipulation, le stockage, la structuration et la préparation des données, mais ils présentent des limites, ce qui peut rendre difficile la manipulation et la préparation de données complexes.

Ces outils et technologies offrent des solutions puissantes pour effectuer des analyses de données et des opérations de data, qu'il s'agisse de nettoyer les données, de créer des modèles prédictifs ou d'extraire des informations significatives à partir de données volumineuses et complexes.

Conclusion :

Pour conclure, le chapitre d'introduction au Data Mining explore l'essence de cette discipline cruciale dans l'ère de l'Intelligence Artificielle et de l'analyse de données. Le Data Mining consiste à extraire des connaissances utiles à partir de grandes quantités de données, en identifiant des modèles cachés et des tendances. Il repose sur diverses techniques provenant de domaines comme la statistique et l'apprentissage automatique. Le processus de découverte de connaissances (KDD) guide le Data Mining, en le situant dans un processus itératif allant de la collecte des données à la présentation des connaissances. Les applications du Data Mining sont vastes, allant du marketing à la santé, et ses techniques incluent la classification, le clustering, et les règles d’associations.