

EE-IRM83B2
Introduction au Data Mining

Système d'évaluation

CC+Examen

Crédits

3 ECTS

Liste des compétences visées : CG2, CG3, CG5, CS3, CS4, CS6.

Pré-requis: Algorithmique & structures de données (EE-IRM511), Probabilité & statistiques (EE-IRM621).

Mots clés: Data mining, Règles d'association, arbre de décision, ECD, Clustering.

Objectifs de l'enseignement : La fouille de données (data mining en Anglais) est l'aboutissement des avancées liées aux Bases de Données et à l'Intelligence Artificielle au sens de l'apprentissage. Sous la dénomination de Data Mining, on entend les procédés et les algorithmes utilisés pour découvrir les structures et les relations cachées dans de grandes bases de données. Ce procédé n'est pas limité aux données transactionnelles, mais il s'est étendu également aux données textuelles et aux données multimédias de type image et web. Le Data Mining est ainsi une discipline d'actualité, à fort potentiel pour les entreprises et est à cheval entre les mathématiques/statistiques et l'informatique. Le cours offre un large aperçu des principales tâches et techniques de Data Mining.

Contenu de l'enseignement :

1. Leçon 1 : Introduction au Data Mining
 - (a) Section 1 : Motivations : Contexte général et constat
 - (b) Section 2 : Intérêt du Data Mining
 - (c) Section 3 : Processus d'extraction des connaissances à partir des données (ECD)
 - (d) Section 4 : Etapes du processus d'ECD
 - (e) Section 5 : Caractéristique du processus d'ECD
 - (f) Section 6 : Processus de conduite de l'exploitation de données
 - (g) Section 7 : Evolution du cadre d'application du Data Mining
2. Leçon 2 : Tâches, techniques et logiciels de fouille de données
 - (a) Section 1 : Tâches de fouille de données
 - (b) Section 2 : Techniques de fouille de données
 - (c) Section 3 : Exemples réels d'applications
 - (d) Section 4 : Principaux logiciels de fouille de données
3. Leçon 3 : Fouille des itemsets intéressants et des règles d'association

- (a) Section 1 : Intérêt de la fouille des associations et cadres applicatifs
- (b) Section 2 : Contexte d'extraction
- (c) Section 3 : Propriétés et mesures de qualité d'un itemset
- (d) Section 4 : Notion de treillis et bordures
- (e) Section 5 : Fouille des règles d'association
- (f) Section 6 : Algorithmes d'extraction des itemsets intéressants et de dérivation des règles d'association
- (g) Section 7 : Problématique des cas réels
- (h) Section 8 : Caractérisation et fouille des représentations concises des motifs
- (i) Section 9 : Avantages et inconvénients de la technique

4. Leçon 4 : Fouille des arbres de décision

- (a) Section 1 : Données qualitatives versus Données quantitatives
- (b) Section 2 : Intérêt de la classification d'une population et cadres applicatifs
- (c) Section 3 : Généralités et notions de base associées aux arbres de décision (bases d'apprentissage et de test, taux d'erreur d'apprentissage et de classification, etc.)
- (d) Section 4 : Principe algorithmique des arbres de décision
- (e) Section 5 : Fonctions d'évaluation de la mélangeance d'une population
- (f) Section 6 : Algorithmes de constructions des arbres de décision
- (g) Section 7 : Validation croisée et taux d'erreur
- (h) Section 8 : Avantages et inconvénients de la technique

5. Leçon 5 : Fouille des clusters (regroupement)

- (a) Section 1 : Nécessité de segmenter une population en groupes et cadres applicatifs
- (b) Section 2 : Mesures de similarité et types de données
- (c) Section 3 : Mesure de distance entre objets
- (d) Section 4 : Méthodes de clustering
- (e) Section 5 : Avantages et inconvénients de la technique

Eventuels projets à réaliser: (à suivre ...)