Fiche de Revision - Data Mining

1. Definition du Data Mining

- Ensemble de techniques pour explorer/analyser de grandes bases de donnees.
- Types :
- - Descriptif : decouvrir motifs/associations
- - Predictif : predire des resultats (valeurs ou classes)

2. Etapes du processus KDD

- 1. Comprendre le domaine dapplication
- 2. Definir les objectifs et connaissances existantes
- 3. Selection des donnees
- 4. Nettoyage et pretraitement
- 5. Reduction/transformation des donnees
- 6. Application du data mining
- 7. Evaluation des resultats (patterns utiles, nouveaux, interpretables)

3. Domaines dapplication

- Marketing, assurance, medecine, finance, detection de fraudes, web, etc.

4. Taches principales du Data Mining

- - Descriptives : clustering, association (pas de variable cible)
- - Predictives : classification, regression (variable cible presente)

5. Techniques de classification

- KNN (k plus proches voisins) : base sur la distance, pas de modele appris
- Arbres de decision : ID3, C4.5, CART (utilisent le gain d'information, elagage necessaire)

6. Regression lineaire

- Formule : $Y = aX + b$
- Qualite du modele mesuree par R2 (plus R2 proche de 1, meilleur est lajustement)

7. Reseaux de neurones (ANNs)

- Inspires du cerveau humain
- Types : perceptron simple, multicouches (MLP + backpropagation)
- Limites : interpretabilite, risque doverfitting, temps dapprentissage eleve

8. Clustering (segmentation non supervisee)

- But : regrouper des elements similaires (K-means, hierarchique...)
- Qualite : forte similarite intra-cluster et faible inter-cluster

9. Logiciels de Data Mining

- Open source : R, Weka, Tanagra
- Professionnels : SAS, SPSS, KNIME, IBM Intelligent Miner

10. Challenges du Data Mining

- Qualite et volume des donnees (Big Data)
- Overfitting
- Interpretation des resultats