CNAM - Paris

Master Science des Données

Synthèse du Cours STA211

Entreposage et Fouille de Données Massives

Cours : STA211 - Entreposage et Fouille de Données Massives

Enseignants : V. Audigier, N. Niang

Année : 2024-2025

Auteur de la synthèse : Maoulida Abdoullatuf

3 août 2025

Table des matières

1. Introduction au Data Mining et Fondements

1.1 Définitions Fondamentales

Le data mining (fouille de données) est défini dans le cours comme « l'application des techniques de statistique, d'analyse de données et d'intelligence artificielle à l'exploration et l'analyse sans a priori de grandes bases de données informatiques, en vue d'en extraire des informations nouvelles et utiles pour le détenteur de ces données » (Tufféry, 2007).

Une définition complémentaire selon Fayyad, Piatetsky-Shapiro, and Smyth (1996) : « Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data ».

Réflexion personnelle : Ces définitions soulignent l'aspect exploratoire du data mining, qui se distingue des approches statistiques classiques par son caractère non dirigé. Contrairement aux tests d'hypothèses traditionnels, le data mining part des données pour découvrir des structures, ce qui nécessite une vigilance particulière face au risque de sur-interprétation.

1.2 Le Processus KDD

Le data mining s'inscrit dans un processus plus large appelé KDD, qui se résume selon le cours à cinq étapes essentielles :

1. **Sélectionner** les données a priori pertinentes à l'objet de l'étude
2. **Pré-traiter** les données (correction d'erreurs, gestion des données manquantes)
3. **Transformer** les données (discrétisation, fusion de modalités)
4. **Extraire** l'information contenue au sein des données
5. **Interpréter et évaluer** les résultats des méthodes de fouille

Le cours précise que « le procédé de KDD n'est pas linéaire » et nécessite souvent de reprendre le processus en amont.

Application pratique : Dans un projet de détection de fraude bancaire, par exemple, la sélection porterait sur les transactions et profils clients, le pré-traitement gérerait les données manquantes sur les revenus, la transformation créerait des variables dérivées (ratios, agrégations temporelles), l'extraction utiliserait des algorithmes de détection d'anomalies, et l'interprétation validerait les règles découvertes avec les experts métier.

1.3 Évolution Historique et Enjeux Actuels

Le cours identifie quatre périodes clés :

— **Avant 1970 :** « la donnée était rare, on allait la récolter de façon à pouvoir répondre à des questions précises, définies en amont »

— **1970-1990 :** « l'ordinateur a commencé à envahir la société, avec lui on commence à accumuler davantage de données »

— **1990-2010 :** « on passe à une autre échelle avec la démocratisation du web », données de types nouveaux (images, sons, etc.)

— **Après 2010 :** « on passe à l'ère des données massives (le Big data) », jeux de données de l'ordre du Téraoctet

1.4 Secteurs d'Application et Cas d'Usage

Secteur Bancaire : « Un des secteurs où le data-mining joue un rôle des plus importants » avec le scoring pour déterminer la probabilité qu'un client soit un bon payeur. Les modèles intègrent des variables comportementales, démographiques et transactionnelles.

Grande Distribution : Utilisation des tickets de caisse et cartes de fidélité pour « identifier les règles d'association entre les achats de produits ». L'analyse du panier de la ménagère permet l'optimisation des promotions et de l'implantation magasin.

Téléphonie : Face à la saturation du marché, « évaluer la probabilité qu'un client parte chez un concurrent » (churn prediction). Les signaux précurseurs incluent la baisse d'utilisation, les appels au service client, et les comportements atypiques.

Secteur Médical : « Rechercher des associations entre gènes et pathologies, ou identifier des anomalies dans des résultats d'imagerie ». L'aide au diagnostic s'appuie sur l'apprentissage supervisé avec validation par des experts.

2. Gestion des Données Manquantes

2.1 Problématique Centrale et Enjeux Pratiques

Le cours souligne que « les données manquantes constituent un problème quasiment incontournable dans l'application des techniques de data-mining ». L'augmentation de la taille des données fait « mécaniquement augmenter la probabilité de rencontrer des données manquantes ».

Le cours donne un exemple concret : « supposons que les données manquantes soient disposées complètement au hasard et que 5% des valeurs soient manquantes pour chaque variable, alors un jeu composé de 50 variables contient en moyenne 8% d'individus complets ».

Analyse critique : Cette problématique illustre un paradoxe du Big Data : plus on collecte de variables, moins on a d'observations complètes. Cela nécessite un arbitrage constant entre richesse descriptive et complétude des données.

2.2 Taxonomie des Mécanismes

MCAR (Missing Completely At Random) : Données manquantes complètement aléatoires, indépendantes des valeurs observées et non observées. *Exemple :* Un dysfonctionnement technique aléatoire d'un capteur.

MAR (Missing At Random) : Données manquantes dépendant uniquement des variables observées, pas des valeurs manquantes elles-mêmes. *Exemple :* Les revenus manquants plus fréquents chez les jeunes (âge observé).

MNAR (Missing Not At Random) : Données manquantes dépendant des valeurs manquantes elles-mêmes ou de variables non observées. *Exemple :* Non-réponse volontaire aux questions sur les revenus élevés.

2.3 Méthodes de Traitement et Comparaison

Suppression par Liste : Le cours note que c'est « l'attitude généralement adoptée par les utilisateurs » et « utilisée par défaut dans la plupart des logiciels ». Cependant, « cette façon de procéder n'est pas satisfaisante » car les individus complets ne constituent pas nécessairement un échantillon représentatif.

Imputation Simple : Comprend l'imputation par la moyenne/mode, l'imputation par kNN utilisant la similarité entre individus, et l'imputation par régression.

Imputation Multiple : Technique avancée générant plusieurs jeux de données imputés pour tenir compte de l'incertitude liée à l'imputation.

Retour d'expérience : En pratique, le choix dépend du pourcentage de manquant, du mécanisme supposé, et de l'objectif final. Pour un modèle prédictif, l'imputation par kNN préserve souvent mieux les relations locales que l'imputation par la moyenne, trop simpliste.

3. Méthodes Non-Supervisées

Le cours précise que « parmi les méthodes non-supervisées, on retrouve essentiellement les méthodes de classification, les méthodes de recherche de règles d'association et les méthodes d'analyse factorielle ».

3.1 Classification Non-Supervisée (Clustering)

Objectif : « Identifier des groupes d'individus (ou de variables) homogènes » en se focalisant sur « les méthodes basées sur des distances qui sont essentiellement basées sur des approches géométriques ».

K-means : « définir une partition de l'ensemble des individus en K (défini à l'avance) groupes telle que la dispersion des individus au sein des groupes soit la plus faible possible »

Avantages : Simplicité, efficacité computationnelle, adapté aux clusters sphériques

Limites : Sensibilité à l'initialisation, nécessité de fixer K a priori, hypothèse de clusters convexes

Méthodes Hiérarchiques Ascendantes : « partent d'une partition très fine puis fusionnent les classes les plus voisines de proche en proche »

Avantages : Dendrogramme informatif, pas de K à fixer a priori

Limites : Complexité quadratique, sensibilité aux outliers

Méthodes Hiérarchiques Descendantes : « partent d'une classe regroupant l'ensemble des individus et la partitionnent de façon itérative »

Choix méthodologique : La CAH est préférable pour l'exploration initiale et la détermination du nombre de clusters, tandis que K-means est plus adapté aux gros volumes une fois K déterminé.

3.2 Cartes de Kohonen (SOM)

Le cours se focalise particulièrement sur cette méthode car « ces méthodes sont supposées connues du lecteur pour ce qui concerne la classification des individus ». Les cartes de Kohonen constituent un réseau neuronal de projection topologique préservant les relations de voisinage.

Spécificité : Contrairement aux méthodes précédentes, les SOM préservent la topologie des données en projetant sur une grille de faible dimension, facilitant la visualisation de structures complexes.

4. Arbres de Décision et Classification

Le cours définit les arbres comme faisant « partie des méthodes supervisées » et ayant « pour but d'expliquer et/ou prédire une variable réponse Y à partir d'un ensemble de variables explicatives $(X_1, \ldots, X_p)$ ».

4.1 Caractéristiques et Avantages

— « Méthodes non-paramétriques » où « la nature du lien entre X et Y n'est pas définie a priori »

— « Aucune hypothèse sur la distribution des données n'est nécessaire »

— Flexibilité : variables d'entrée et de sortie peuvent être quantitatives ou qualitatives

Atouts uniques : L'interprétabilité constitue l'avantage majeur des arbres. Contrairement aux méthodes « boîte noire », chaque prédiction est explicable par le chemin dans l'arbre, crucial pour les applications nécessitant de la transparence (médical, juridique, finance).

4.2 Principe de Segmentation et Algorithmes

Les arbres « font partie des méthodes dites de segmentation ». Le principe consiste à « partitionner l'espace des variables explicatives $X = (X_1, \ldots, X_p)$ en R régions et d'associer un modèle simple à chacune d'entre elles ».

Pour la régression : Minimisation de la variance intra-nœud

Pour la classification : Utilisation de l'indice de Gini ou de l'entropie

4.3 Construction, Élagage et Validation

Le cours souligne l'importance de l'élagage pour éviter le sur-ajustement, utilisant la validation croisée pour sélectionner le niveau optimal.

Dilemme biais-variance : Un arbre profond (faible biais, forte variance) surapprendra, tandis qu'un arbre peu profond (biais élevé, faible variance) sous-apprendra. L'élagage via validation croisée trouve le compromis optimal.

Cas d'usage recommandé : Les arbres simples conviennent pour l'exploration et l'interprétation, mais pour la prédiction pure, les méthodes d'ensemble (forêts, boosting) les surpassent généralement.

5. Méta-algorithmes : Bagging, Boosting et Forêts

Le cours introduit les méta-algorithmes en expliquant qu'« il est possible d'améliorer les performances [des méthodes supervisées] en utilisant des méthodes d'agrégation qui consistent à combiner plusieurs prédictions fournies par différents modèles ».

5.1 Bagging (Bootstrap Aggregating)

Le bagging « est aussi appelé agrégation bootstrap » et repose sur le résultat fondamental : « si on se donne un n-échantillon $(T_1, \ldots, T_n)$ tel que $\text{Var}[T_i] = \sigma^2$ pour tout $1 \leq i \leq n$, alors $\text{Var}[\frac{1}{n}\sum T_i] = \frac{\sigma^2}{n}$ ».

Objectif : « Procédure qui permet de réduire la variance d'une méthode supervisée ».

Principe pratique : Générer B échantillons bootstrap, entraîner B modèles, et moyenner les prédictions. Plus B est grand, plus la variance diminue, au prix d'un coût computationnel accru.

5.2 Boosting et Apprentissage Adaptatif

Le boosting constitue un apprentissage séquentiel où chaque modèle tente de corriger les erreurs des précédents :

— Pondération des observations mal classées

— Combinaison pondérée des modèles selon leur performance

— Réduction simultanée du biais et de la variance

Comparaison avec le bagging : Alors que le bagging réduit la variance en parallélisant l'apprentissage, le boosting réduit le biais en se concentrant séquentiellement sur les erreurs. Le boosting est plus sensible au bruit mais souvent plus performant.

5.3 Forêts Aléatoires : Synthèse Optimale

Les forêts aléatoires combinent le bagging avec la sélection aléatoire de variables à chaque nœud, réduisant ainsi la corrélation entre les arbres. Elles fournissent naturellement une mesure d'importance des variables basée sur la diminution de pureté ou la permutation OOB.

Avantages pratiques :

— Peu de paramètres à régler

— Robustesse aux outliers et surajustement

— Traitement naturel des variables mixtes

— Estimation d'erreur sans validation externe (OOB)

5.4 Extra-Trees et Extensions

Extension des forêts aléatoires avec seuils de coupure aléatoires, offrant plus grande randomisation, calcul plus rapide, et réduction supplémentaire du sur-ajustement.

Conseil pratique : Les forêts aléatoires constituent souvent un excellent choix par défaut, combinant performance, robustesse et facilité d'utilisation. Elles servent de référence pour évaluer d'autres méthodes plus complexes.

6. Éléments Fondamentaux des Méthodes Supervisées

Le cours précise que « le data-mining vise à identifier des structures au sein de données » en distinguant « deux types de structures : les modèles et les patterns ». Contrairement aux patterns, « les modèles visent à expliquer et/ou prédire une variable réponse, notée ici Y, à partir d'autres variables dites explicatives, notée X ».

6.1 Types de Problèmes et Choix Méthodologiques

Régression : « si Y est continue » - Objectif de minimiser l'erreur quadratique moyenne

Classification supervisée : « si Y est qualitative » - Objectif de maximiser le taux de bonne classification

Réflexion sur le choix : Le type de problème détermine les métriques d'évaluation et les algorithmes appropriés. Une variable ordinale peut être traitée en régression ou classification selon l'objectif métier.

6.2 Compromis Biais-Variance : Cœur de l'Apprentissage

Biais : Écart entre la vraie fonction et l'estimation moyenne

Variance : Variabilité des estimations autour de leur moyenne

Erreur irréductible : Bruit inhérent aux données

Enseignement fondamental : Ce compromis guide tous les choix méthodologiques. Les méthodes simples (régression linéaire) ont un biais potentiellement élevé mais une faible variance, tandis que les méthodes complexes (réseaux profonds) ont l'inverse. La régularisation permet de naviguer dans ce compromis.

6.3 Sélection de Modèles et Validation

Critères Pénalisés :

— AIC (Akaike Information Criterion) : $-2\log L + 2p$

— BIC (Bayesian Information Criterion) : $-2\log L + p\log(n)$

Approches Empiriques :

— Découpage test et apprentissage

— Validation croisée : « estimation plus robuste de la perte »

— Leave-one-out : Cas particulier de validation croisée

Bonnes pratiques : La validation croisée reste la référence pour des échantillons de taille modérée, tandis qu'un simple découpage suffit pour les gros volumes. L'important est de ne jamais évaluer sur les données d'entraînement.

7. Réseaux de Neurones

7.1 Perceptron Multi-Couches (MLP)

— **Couches d'entrée :** Réception des variables explicatives

— **Couches cachées :** Transformation non-linéaire via fonctions d'activation

— **Couche de sortie :** Prédiction finale

— **Connexions :** Entièrement connectées entre couches adjacentes

Architecture et capacité : Le théorème d'approximation universelle garantit qu'un MLP avec une couche cachée suffisamment large peut approximer toute fonction continue. En pratique, plusieurs couches plus petites sont préférables.

7.2 Apprentissage et Optimisation

— **Rétropropagation :** Calcul des gradients couche par couche via la règle de dérivation en chaîne

— **Descente de gradient :** Minimisation itérative de la fonction de coût

— **Défis :** Minima locaux, choix du taux d'apprentissage, initialisation des poids

7.3 Réseaux Convolutionnels (CNN)

— **Couches de convolution :** Application de filtres locaux préservant la structure spatiale

— **Couches de pooling :** Réduction de dimensionnalité et invariance partielle

— **Fonctions d'activation :** Non-linéarités (ReLU, sigmoid, tanh)

— **Applications :** Traitement d'images et signaux structurés

Avantage spécifique : Les CNN exploitent la structure locale des données (pixels voisins, séquences temporelles) contrairement aux MLP qui traitent toutes les entrées de manière équivalente.

8. Règles d'Association et Analyse Multi-blocs

8.1 Règles d'Association : Au-delà de la Corrélation

Le cours note que ces méthodes sont « globalement moins répandues que les méthodes précédentes, sont très utilisées dans le domaine du marketing ». Elles consistent à « comparer, pour certains produits, les probabilités d'achats simultanés aux probabilités obtenues dans le cas où il n'y aurait pas de lien entre les achats des produits ».

Mesures Fondamentales :

— **Support :** Fréquence conjointe $P(A \cap B)$ - mesure la popularité

— **Confiance :** Probabilité conditionnelle $P(B|A)$ - mesure la fiabilité

— **Lift :** Mesure d'indépendance, lift > 1 indique une dépendance positive

Algorithmes et Complexité :

— **Apriori :** Basé sur la propriété d'anti-monotonie (si un itemset est peu fréquent, tous ses sur-ensembles le sont aussi)

— **ECLAT :** Approche par intersection d'ensembles, plus efficace en mémoire

— **FP-Growth :** Utilisation d'arbres de fréquence, évite la génération explicite de candidats

Application pratique : Au-delà du marketing, les règles d'association s'appliquent à la détection de fraude (combinaisons de transactions suspectes), à la bioinformatique (co-occurrence de gènes), ou à l'analyse de logs (séquences d'événements).

8.2 Analyse Multi-blocs : Intégration de Sources Hétérogènes

« Les méthodes d'analyse factorielle sont complémentaires aux précédentes » et peuvent être « employées comme méthode de pré-traitement ».

— **Double ACP (DACP) :** Analyse de structures temporelles, adaptée aux données évolutives

— **STATIS :** Analyse par coefficient RV, compare les structures entre tableaux

— **Analyse Factorielle Multiple (AFM) :** ACP pondérée multi-groupes, équilibre les contributions

Enjeu moderne : Avec la multiplication des sources de données (capteurs, réseaux sociaux, transactions), l'intégration devient cruciale. Ces méthodes permettent de révéler des structures cachées dans des données hétérogènes.

9. Applications Pratiques et Outils

9.1 Environnements Logiciels et Écosystème

Python - Écosystème Scientifique :

— **Scikit-learn :** Bibliothèque généraliste, interface unifiée, excellente documentation

— **TensorFlow/Keras :** Réseaux de neurones, calcul distribué, déploiement production

— **XGBoost, CatBoost :** Boosting avancé, optimisé pour la performance

— **Pandas :** Manipulation de données, interface intuitive

R - Spécialisation Statistique :

— **randomForest, gbm :** Méta-algorithmes, implémentations de référence

— **rpart, tree :** Arbres de décision, visualisation claire

— **FactoMineR, ade4 :** Analyse factorielle, méthodes françaises

— **VIM, mice :** Données manquantes, diagnostic et imputation

Choix pratique : Python domine pour le déploiement et l'ingénierie, R excelle pour l'exploration et les méthodes statistiques avancées. Maîtriser les deux offre une flexibilité maximale.

9.2 Méthodologie de Projet : De la Théorie à la Pratique

1. **Compréhension du métier :** Définition des objectifs, contraintes, métriques de succès
2. **Compréhension des données :** Exploration, évaluation de la qualité, identification des biais
3. **Préparation des données :** Nettoyage, transformation, création de variables
4. **Modélisation :** Sélection d'algorithmes, optimisation des hyperparamètres
5. **Évaluation :** Validation des performances, test de robustesse
6. **Déploiement :** Mise en production, monitoring, maintenance

Retour d'expérience : En pratique, 80% du temps est consacré aux étapes 1-3, souvent sous-estimées. La qualité de ces phases détermine largement le succès du projet.

10. Démarche Méthodologique Intégrée

10.1 Workflow Type d'un Projet de Data Mining

Phase Exploratoire (Méthodes Non-Supervisées) :

1. Analyse factorielle (ACP/ACM) pour réduire la dimensionnalité et révéler les structures principales
2. Classification non-supervisée pour identifier des groupes naturels
3. Règles d'association pour découvrir des patterns comportementaux

Phase Prédictive (Méthodes Supervisées) :

1. Modèles simples (régression, arbres) pour établir une baseline interprétable
2. Méthodes d'ensemble (forêts, boosting) pour optimiser les performances
3. Méthodes avancées (réseaux de neurones) si la complexité est justifiée

10.2 Critères de Choix Méthodologique

Contraintes d'Interprétabilité :

— **Forte :** Régression linéaire, arbres simples, règles d'association

— **Modérée :** Forêts aléatoires (importance des variables), modèles linéaires généralisés

— **Faible :** Réseaux de neurones, SVM à noyau, ensembles complexes

Contraintes de Performance :

— **Données structurées classiques :** Forêts aléatoires, boosting souvent optimaux

— **Données séquentielles/images :** Réseaux convolutionnels incontournables

— **Petits échantillons :** Méthodes simples avec régularisation

Contraintes Computationnelles :

— **Temps réel :** Modèles linéaires, arbres peu profonds

— **Gros volumes :** Algorithmes parallélisables, approximations stochastiques

— **Mémoire limitée :** Méthodes incrémentales, échantillonnage

10.3 Validation et Mise en Production

Validation Statistique :

— Tests de significativité pour les modèles paramétriques

— Intervalles de confiance pour les métriques de performance

— Tests de stabilité temporelle (concept drift)

Validation Métier :

— Cohérence avec l'expertise domain

— Robustesse aux variations opérationnelles

— Acceptabilité éthique et réglementaire

11. Conclusion et Perspectives

11.1 Messages Clés du Cours

Le cours STA211 met en évidence plusieurs points fondamentaux :

Importance du Prétraitement : Le prétraitement constitue souvent la majorité du travail en data mining, avec une attention particulière à la qualité des données comme facteur déterminant de performance. Cette réalité contraste avec la perception populaire du data mining comme application directe d'algorithmes sophistiqués.

Complémentarité des Méthodes : Exploration puis prédiction via méthodes non-supervisées suivies de supervisées. Compromis interprétabilité-performance selon le contexte applicatif. Aucune méthode n'est universellement supérieure ; le choix dépend des contraintes spécifiques du problème.

Validation Rigoureuse : Validation croisée systématique pour éviter le sur-ajustement. Métriques adaptées au type de problème et aux contraintes métier. La validation constitue le garde-fou contre l'illusion de découverte dans des données complexes.

11.2 Vision Globale et Interdisciplinarité

Le cours présente le data mining comme une « démarche complète » située « à l'intersection de la statistique et des technologies de l'information ». Cette approche combine :

— **Rigueur scientifique :** Validation statistique des résultats, gestion de l'incertitude

— **Expertise métier :** Compréhension du domaine d'application, traduction des besoins

— **Savoir-faire technique :** Maîtrise des outils et méthodes, optimisation algorithmique

— **Sens critique :** Interprétation et remise en question des résultats, détection des biais

Réflexion personnelle : Cette vision holistique distingue le data mining de l'application mécanique d'algorithmes. Le praticien doit naviguer entre contraintes techniques, exigences métier, et rigueur scientifique.

11.3 Évolutions et Défis Futurs

Tendances Actuelles :

— **Deep Learning :** Réseaux de neurones profonds pour données complexes

— **AutoML :** Automatisation de la sélection de modèles et optimisation d'hyperparamètres

— **Explicabilité :** Techniques d'interprétation des modèles complexes (SHAP, LIME)

Défis Émergents :

— **Éthique et biais :** Équité des algorithmes, discrimination algorithmique

— **Vie privée :** Apprentissage préservant la confidentialité, anonymisation

— **Robustesse :** Résistance aux attaques adversariales, stabilité temporelle

— **Durabilité :** Efficacité énergétique des modèles, impact environnemental

11.4 Réflexion Finale

Comme le souligne le cours, le data mining constitue un champ d'analyse nouveau dont « l'évolution se poursuit encore aujourd'hui » avec l'émergence de nouvelles technologies et leur appropriation par la société.

Cette synthèse reflète la richesse du cours STA211, qui forme des praticiens capables de naviguer entre théorie et pratique, performance et interprétabilité, innovation et rigueur méthodologique.

Le data mining n'est pas seulement un ensemble de techniques, c'est une approche globale qui nécessite une compréhension profonde des données, des méthodes et du contexte applicatif pour extraire une connaissance utile et fiable.

L'évolution rapide du domaine exige une formation solide aux fondements plutôt qu'aux outils du moment. Les concepts de validation, compromis biais-variance, et démarche méthodologique restent invariants face aux innovations technologiques.

En définitive, le data mining réussi repose sur l'alliance de la rigueur scientifique, de la créativité technique, et de la compréhension métier – triptyque que ce cours STA211 s'efforce de transmettre.

Sources

— Audigier, V., Niang, N. (2024-2025). Documents de cours STA211

— Tufféry, S. (2007). Data Mining et statistique décisionnelle

— Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases