Cours d'Analyse des Données Multidimensionnelles

Chapitre 0 : Introduction

Zaineb Smida

5ème année GEA
INSA Lyon

14 novembre 2024

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Rappels sur le vocabulaire de base de la statistique

- *individu* (ou *observation*) : entité de base en statistique (ex : étudiants, entreprises, régions, années...)
- *population* : ensemble des individus
- *effectif total* : nombre d'individus de la population
- *échantillon* : un sous-ensemble (de taille n) de la population
- *variable* : caractéristique mesurée sur des individus (ex : âge, nombre de salariés, sexe, secteur d'activité...)
- *modalités* : valeurs observées de la variable

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Définitions

- La statistique descriptive ou exploratoire consiste à traiter et à interpréter les informations recueillies par le biais de données.
- Analyse des données : ensemble de méthodes statistiques exploratoire multidimensionnelle (dans le sens où plusieurs variables sont mesurées simultanément sur les mêmes individus).
- Le but de la statistique descriptive (uni-ou bi-dimensionnelle) ou des méthodes d'analyse de données est de synthétiser et de résumer l'information contenue dans les données.
- Chaque méthode d'analyse de données correspond à **un type de données** et à **un objectif précis**.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Les principales méthodes d'analyse des données se séparent en deux groupes :

- Les méthodes de classification visant à réduire la taille de l'ensemble des individus en formant des groupes homogènes ;

- les méthodes factorielles qui cherchent à réduire le nombre de variables en les résumant en un petit nombre de composantes synthétiques. Selon que l'on travaille avec un tableau de variables quantitatives ou des variables qualitatives, on utilisera des méthodes différentes (resp. ici l'analyse en composantes principales (ACP) ou bien l'analyse factorielle des correspondances (AFC)).

- Objectif du cours : présenter quatre de ces méthodes
  1. Analyse en composantes principales (ACP),
  2. Analyse factorielle des correspondances (AFC),
  3. Méthodes de classification non supervisée : agrégation autour des moyennes mobiles (AMM) et classification ascendante hiérarchique (CAH) .

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Aujourd'hui, on parle plutôt de *data mining* (prospection des données afin de les transformer en connaissance). Le *data mining* regroupe un certain nombre de techniques :

- Informatique : gestion des bases de données,
- Statistique : analyse de données,
- Intelligence Artificielle : réseaux de neurones.

Ces techniques existent depuis plusieurs années mais deviennent aujourd'hui **une réalité industrielle** car

- les données sont produites (codes barre, cartes de crédit, achats par internet,...),
- les données sont archivées (*data warehouse*),
- la puissance de calcul nécessaire est abordable,
- le contexte est ultra-concurrentiel (télécommunications, assurances, services financiers,...),
- les produits commerciaux pour le *data mining* sont disponibles.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

- *Big data* : ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de bases de données ou de gestion de l'information.

  Les 5 V du *Big data* :
  - Volume : données massives (= mégadonnées),
  - Vélocité : fréquence élevée avec laquelle les données sont générées, capturées, partagées et mises à jour,
  - Variété : données hétérogènes = données complexes provenant du web (*web mining*), au format texte (*text mining*) et images (*image mining*),
  - Véracité (ou fiabilité des données) : est notamment menacée par les comportements déclaratifs (sur formulaires), par les diversités des points de collecte, par la multiplication des formats de données et par l'activité des robots et faux profils innombrables sévissant sur Internet,
  - Valeur : dans un contexte d'infobésité, il s'agit d'être capable de se concentrer sur les données ayant une réelle valeur et étant actionnables.

Applications de l'analyse des données

- Marketing quantitatif : études de marchés, enquêtes de satisfaction, études de typologies.
- Economie, sciences sociales, médecine, météorologie...

Exemples :

- dans les banques :
  - score d'appétence : mesure quantitative de la propension d'un client à être intéressé par un produit.
  - crédit scoring : la banque attribue des notes aux clients pour savoir s'il est raisonnable de leur accorder un prêt.
- en téléphonie mobile : score d'attrition (mesure le risque de voir le client partir à la concurrence)

Caractéristiques des méthodes d'analyse des données

- Méthodes statistiques ***descriptives*** par opposition aux méthodes ***inférentielles*** (basées sur des modèles probabilistes). Dans ce cours, présentation descriptive (pas de tests)
- En statistique multivariée, ***descriptif*** s'oppose aussi à ***explicatif***.
  - dans les méthodes descriptives, toutes les variables sont considérées de la même façon.
  - dans les méthodes explicatives, une variable au moins est *à expliquer* et les autres sont explicatives.
- Méthodes descriptives d'analyse des données : ACP, AFC, Classification non supervisée (*clustering*),
- Méthodes explicatives d'analyse des données : Analyse Factorielle Discriminante, Analyse Canonique.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

Caractéristiques des méthodes d'analyse des données

Remarque :

Autres méthodes explicatives : régression linéaire multiple, régression logistique, arbres de classification et de régression, forêts aléatoires (les deux dernières méthodes sont des algorithmes de machine learning ou apprentissage automatique en français).

Présentation des données : tableau de données brutes

- Se présente sous la forme *individus* / *variables*.
- une ligne = un individu, une colonne = une variable
- On considère n individus et p variables.

Var. / Ind.	Sexe	Âge	...	X^j	...	X^p
1						
2	H	25				
⋮						
X_i				X_i^j		
⋮						
X_n						