ENSEIRB-MATMECA

Traitement De l'Information

semestre 5 — 2020-2021

Contents

Chapter 1

Introduction

1.1 Constitution du tableau de donnée

Classe	Survivant	Nom	Sexe	Age	Parent présents	Ref billet	Prix billet	Port	
1	1	Allen Elisabeth	F	29	0	null	211	B5	S
3	0	Dyker Adolf	N	23	1	null	7	null	C
3	1	Dyker Anna	F	22	1	null	null	null	C
3	0	Emir Farell	N	null	0	null	null	null	null

Toute les variables doivent être du même type (→ quels sont les types de var ?)

- Quid des données manquantes ? (→ les remplacer ? par la moyenne ?)
- Quid des donnée abérrantes ?
- Lignes redondantes ? (→ à supprimer ?)
- Colonnes redondantes ?

1.2 Les variables et les individus

1.2.1 les individus

Eléments de la population étudiée

1.2.2 les variables

Application definie par

$$\omega(\text{population}) \longrightarrow O(\text{espace d'observation}) \text{ struct de cet espace}$$

1^{ere} typologie

card O struct de O	O ensemble continu de $\mathbb{R}$	O fini ou denombrable	.
sans struct $=, \neq$	/	C8P lieu de residence	nominale
struct ordi. $\leq$	Age imperative	rang note	ordinal
corps ordonnée	salaire	/	mesurable
.	quantifiable	qualifiable	.

2^{eme}typologie : →variable d'incidence
→ variable relationelle

- variable d'incidence
  → attribut descriptif ,reponse oui = 1 OU non = 0
  → var. numérique espace d'observation = $\mathbb{R}$

- var. relationnelles
  → var rang → échelle de notes suffisamment fine pour que 2 individu n'aient pas (pas de suite c'est effacé)
  → var. présentant des modalités non ordonnées (bac, csp, ...)
  → var. mesures sur $\mathbb{R} \times \mathbb{R}$

1.3 Les types de tableaux

CN Tableaux doivent être homogènes 1 seul types de variable

1.3.1 Tableau de données quantitatives

(pourcentage de minerai par sondage / profondeur)

$$x_{i \leftarrow ligne}^{j \leftarrow colonne}$$

traitement

1.3.2 tableau de contingence

croisement de 2 var qualitatives
Traitement : Analyse fonctionelle des correspondances

1.3.3 tableaux binaires (logique d'incidence)

tableaux d'attributs descriptif de 1,0
Traitement : classifications hiérarchiques

1.3.4 Tableaux de préferences

Notes données à des marques de parfum
Traitement : Analyse fonctionelle des correspondances multiples

1.3.5 Tableaux de modalités

lisez vous tel journal ?

- 5 : tout le temps
- 4 : regulièrement
- 3 : parfois
- 2 : rarement
- 1 : jamais

1.3.6 tableaux de proximité

mesure sur $\mathbb{R} \times \mathbb{R}$

	r1	r2	r3
r1			
r2			
r3			

1.4 Changement de variables

→ pour rendre le tableau homogène

	Sexe	couleur yeux
pere	M	marron
mere	F	bleu
enfant	M	vert

codage disjonctif complet

	sexe M	sexe F	yeux marr	yeux bleu	yeux vert
pere	1	0	1	0	0
mere	0	1	0	1	0
enfant	1	0	0	0	1

⟶ Tableau de *Burt*

${}^tD \times D$

	sexe M	sexe F	Y m	Y b	Y v
sexe M	2	0	1	0	1
sexe F	0	1	0	1	0
Y m	1	0	1	0	0
Y b	0	1	0	1	0
Y v	1	0	0	0	1

Tableaux d'effectifs ⟶ Analyse fonctionnelle des correspondances

autre exemple : Tableau de Burt du Titanic

	classe 1	classe 2	classe 3	enf	adul	F	M	surv 0	surv 1
classe 1	325	0	0	6	319	147	180	122	203
classe 2	0	285	0	24	261	106	179	167	118
classe 3	0	0	706	79	627	null	null	null	null
etc	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

1.5 Elements descriptifs d'un tableau de données

n : individus/lignes
p : variable/colonnes

individu i $= x_i = \begin{pmatrix} x_i^1 & \dots & x_i^p \end{pmatrix} \in \mathbb{R}^p$ la variable j : $x_j = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \end{pmatrix} \in \mathbb{R}^n$

On va étudier le nuage $N_I = \{x_i, i = 1, \dots, n, \text{poids} p_i = \frac{1}{n}\}$

$N_J = \{x_j, j = 1, \dots, p, \text{poids} p_j = \frac{1}{p}\}$

Soit le nuage des variables
A chaque variable, on peut associer: sa myenne

$$\underline{x}^j = \sum_{i=1}^{n} p_i x_i^j$$

sa variance $Var\underline{x}^j = \sum_{i=1}^n p_i (x_i^j - \overline{x}^j)^2$

TROU

Pour le nuage $N_I$on peut calc le centre de garvité :

$$\overline{\underline{x}} = \sum_{i=1}^{n} p_i \underline{x_i}$$

$$\overline{\underline{x}} = \begin{pmatrix} \overline{x}^1 \\ \vdots \\ \overline{x}^p \end{pmatrix}$$

Sur le nuage N_j on peut calculer le centre de gravité

$$\overline{\underline{x}} = \begin{pmatrix} \overline{\underline{x}}_1 \\ \vdots \\ \overline{\underline{x}}_n \end{pmatrix}$$

L'inertie du nuage
des individus $I_{N_I} = \sum_{i=1}^{n} p_i d^2(x_i, \overline{\underline{x}}) \longrightarrow$ distance

1.6 Choix d'une mesure de ressemblance

définition 1 : indice de similarité

$$s\Omega \longrightarrow \mathbb{R}^+$$

s est un indice de similarité si :

1. s est symetrique, $s(x, y) = s(y, x)$
2. $s(x, x) = s(y, y) \geq s(x, y)$

définition 2 : indice de dissimilarité

$$s\Omega \longrightarrow \mathbb{R}^+$$

s est un indice de dissimilarité si :

1. symetrique
2. s(x, x)=0

definition 3 : une distance est un indice de dissimilarité
qui vérifie en plus

1. $s(x, y) = 0 \leftrightarrow x = y$
2. $s(x, z) \leq s(x, y) + s(y, z)$

- distance euclidienne (A en composantes principales)
- distance dite du χ^2 (Analyse factorielle des correspondances)

$$d^2(\underline{x}_i, \underline{x}_{i'}) = \sum_{j=1}^{p} \frac{1}{x_i^j} (\frac{x_i^j x_i}{-} \frac{x_{i'}^j}{x_{i'}})^2$$

- Tableau d'incidence
  TROU

Distance entre groupes
(indice d'aprépation)
distance du lien max
distance du lien min
distance des centres de gravité
trou mais fin de chap

distance fréquemment utilisé
$\delta(A, B) = I$ $\delta(A, B) = d(\overline{\underline{x}}, \overline{\underline{y}})$

Chapter 2

Analyse en composantes principales : *ACP*

Traite les tableaux de données quantitativement positives

2.1 Principe de la méthode

Le pb qui se pose dans la mesure où 1 partie des variables sont liées, c'est de passer d'un tableau X de dimension $n \times p$ à un tableau Y de dim $n \times q$ $q < p$ en réduisant le nbr de variable descriptive tout en perdant le moins possible d'info
la methode utilisé en ACP pour passer de p var à q var q¡p consiste à projeter le nuage des individus sur un sous espace W de dim q en deformant le moins possible le nuage lors de sa projection
Les composantes principales sont les nouvelles variables. elles vont s'interpreter comme des" 'synthèses' des variables initiales $\underline{x^j}$.

TROU flemme 3 lignes + fin du cours

Le principe de l'ACP, c'est de réduire le nbr de variables décrivant les individus en perdant le moins possible d'information

2.2 Formalisation du problème

Rechercher l'espace de projection W (a : origine, $u_1, \ldots, u_q$: système générateur de W) de telle façon que la perte d'info fut minimale

$$\forall i \underline{\widehat{x_i}} = \underline{a} + \sum_{k=1}^{a} y_i^k \underline{u_k}$$

$\underline{\widehat{x_i}}$ est la projection orthogonale de $\underline{x_i} sur\, W$
On chercher le referentiel W tq $\sum_W = \sum_{i=1}^n p_i d^2(\underline{x_i}, \underline{\widehat{x_i}})$

2.3 Resoplution du problème

2.3.1 Choix de a

2.3.2 choix de $[\underline{u_1}, \ldots, \underline{u_q}]$