Formulaire d'Analyse de Données

Pr. Carl FRÉLICOT – Dpt Info / Lab MIA

1. Description Unidimensionnelle de Données — tableau X $(n \times 1)$

1.1) Descripteurs de tendance centrale et localisation :

- la moyenne arithmétique $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ — *mean, average*
- le mode x_m est la valeur la plus fréquente — *modal*
- la médiane $Q_2 = x_{(\frac{n}{2}+1)}$ si n est impair ou $Q_2 = \frac{x_{(n/2)}+x_{(n/2+1)}}{2}$ si n pair — *median*
- les quartiles Q_1 et Q_3 sont les médianes des deux sous-ensembles que sépare la médiane — *quartiles*

1.2) Descripteurs de dispersion :

- la variance (empirique) $s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$ et l'écart-type $s = \sqrt{s^2}$ — *variance, standard deviation*
- le coefficient de variation $cv = \frac{s}{|\overline{x}|}$ — *variation coefficient*
- l'écart absolu moyen $\overline{e} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$ et médian $\widetilde{e} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Q_2|$ — *mean absolute deviation, median a.d.*
- l'étendue $E = |x_{(n)} - x_{(1)}|$, où $x_{(1)}$ et $x_{(n)}$ sont les plus petite et plus grande valeur — *range*
- l'étendue inter-quartile $EIQ = |Q_3 - Q_1|$ et l'épure $\left[Q_1 - \frac{3}{2}EIQ, Q_3 + \frac{3}{2}EIQ\right]$ — *inter quartile range, box*

1.3) Description graphique :

- diagramme en bâtons, en secteurs, histogramme — *bargraph, piechart, histogram*
- boîte à moustaches : la boîte porte les quartiles, et les moustaches vont jusqu'à la + petite et la + grande valeur observée situées dans l'épure ; toutes les valeurs en dehors de l'épure sont symbolisées (*) — *boxplot*

2. Analyse Bidimensionnelle de Données — tableaux X $(n \times 1)$ et Y $(n \times 1)$

2.1) Deux variables quantitatives

- la covariance (variance conjointe) $s_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i \times y_i - \overline{x} \times \overline{y}$ — *si $X = Y$, devient s_x^2*
- le coefficient de corrélation linéaire $r_{xy} = \frac{s_{xy}^2}{s_x \times s_y} \in [-1, 1]$ — *si $X = Y$, devient $r_{xx} = 1$*
- régression (simple) de Y par X : dans le cas d'un modèle *affine* ($Y = a \times X + b$), la solution des moindres carrés donne $\widehat{a} = \frac{s_{xy}^2}{s_x^2}$ et $\widehat{b} = \overline{y} - \widehat{a} \times \overline{x}$

2.2) Une variable qualitative Y et une variable quantitative X

- pour chaque modalité $j = 1, m$ de Y on peut calculer la moyenne $\overline{x}_j$ et la variance s_j^2 de X
- les statistiques globales sur X se décomposent comme suit :
  moyenne globale $\overline{x} = \frac{1}{n}\sum_{j=1}^{m} n_j\,\overline{x}_j$, et variance globale $s^2 = \frac{1}{n}\sum_{j=1}^{m} n_j\,(\overline{x}_j - \overline{x})^2 + \frac{1}{n}\sum_{j=1}^{m} n_j\,s_j^2 = s_B^2 + s_W^2$, où s_B^2 est la variance *inter-groupes*, c'est-à-dire la part de variance de X expliquée par Y, et s_W^2 est la variance *intra-groupes*, dite *résiduelle*, c'est-à-dire la part de variance de X non expliquée par Y
- $\frac{s_B^2}{s^2} \in [0, 1]$ mesure à quel point X dépend de Y

2.3) Deux variables qualitatives X (à l modalités) et Y (à c modalités)

- *table* de contingence $N = [n_{ij}]_{i=1,l;j=1,c}$, où n_{ij} est le nombre d'individus pour lesquels la modalité i de X et la modalité j de Y sont observées *conjointement* ;
  les *effectifs marginaux* $n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$ de X et $n_{\bullet j} = \sum_{i=1}^{l} n_{ij}$ de Y vérifient $\sum_{i=1}^{l} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j} = n$
- si X et Y sont indépendantes, alors la table de contingence (théorique) est $T = [t_{ij}]_{i=1,l;j=1,c}$, avec $t_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$;
  pour mesurer la dépendance, on somme les écarts entre N et T ainsi : $\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(n_{ij}-t_{ij})^2}{t_{ij}}$
- tableaux des l *profils-ligne* $L = [l_{ij}]_{i=1,l;j=1,c}$ où $l_{ij} = \frac{n_{ij}}{n_{i\bullet}}$ et des c profils-colonne $C = [c_{ji}]_{j=1,c;i=1,l}$ où $c_{ji} = \frac{n_{ij}}{n_{\bullet j}}$;
  L (resp. C) définit un nuage de l (resp. c) points en dimension $c - 1$ (resp. $l - 1$)

3. Analyse Multidimensionnelle de Données — tableau X $(n \times p)$

3.1) Produit scalaire $< x, y >= {}^t x\, y = \sum_{j=1}^{p} x_j \times y_j$, norme $||x|| = \sqrt{< x, x >}$ et $cos(x, y) = \frac{<x,y>}{||x||\,||y||}$;
si u est un vecteur directeur unitaire ($||u|| = 1$) définissant un axe, $< x, u >$ est la projection de x sur cet axe

3.2) Distances entre individus — (lignes)

- données quantitatives : Minkowski $d_q(x, y) = \left(\sum_{j=1}^{p}|x_j - y_j|^q\right)^{1/q}$
  - *Manhattan* si $q = 1$
  - *euclidenne* si $q = 2$; alors $d_2^2(x, y) =< x - y, x - y >= ||x - y||^2$
  - *Chebychev* si $q \to +\infty$; alors $d_\infty(x, y) = \max_{j=1,p}|x_j - y_j|$
- données binaires :
  - *Jaccard* $d_J(x, y) = 1 - \frac{b_{11}}{b - b_{00}}$ où b est le nombre de bits, b_{11} (b_{00}) est le nombre de 1 (0) en commun
  - *Hamming* $d_H(x, y) = \frac{b_{10}+b_{01}}{b}$, c'est à dire le % de 0 et de 1 qui différent
- Distances entre variables — (colonnes)
  - *cosinus* $d_{cos}(X, Y) = 1 - cos(X, Y)$
  - *correlation* $d_{correl}(X, Y) = d_{cos}(X'', Y'')$, où X'' et X'' sont centrées-réduites — $d_{correl}(X, Y) = 1 - r_{xy}$

4. Réduction de la Dimensionnalité (Facteurs) **tableau X $(n \times p)$ ou $X|Y$ $(n \times 1)$**

Passer de X à C' $(n \times q, q << p)$ décrivant mieux X (ou expliquant mieux Y) : $C' = X'U$ où U est une base $\perp$-normée.

4.1) Analyse en Composantes Principales (p variables quantitatives)

- U contient les vect. propres de la matrice de covariance $V = {}^tX'\,D\,X'$ (canonique, $D = \frac{1}{n}\,I_p$) ou de la matrice des corrélations $R = {}^tX''\,D\,X''$ (normée) ; les valeurs propres sont les variances des composantes (ordre décroissant)
- on retient q composantes principales par examen des % cumulés de variance expliquée
- on explique les composantes à l'aide des individus et des variables de X bien représentées (cos^2) uniquement

4.2) Analyse Discriminante Linéaire (p variables quantitatives, 1 variable qualitative à m modalités)

- U : vect. propres de $W^{-1}\,B$ où W et B sont les matrices de covariance intra- et inter-groupes de X ; $V = W + B$
- il n'y a que $q = min(m-1, p)$ comp. discriminantes ; la valeur propre ($\in [0,1]$) indique son pouvoir discriminant

4.3) Analyse des Correspondances (p variables qualitatives)

5. Classification Non Supervisée (*Clustering*) **tableau X $(n \times p)$**

Il s'agit de déterminer automatiquement, pour les observations x de X, une variable Y $(n \times 1)$ indicatrice d'appartenance à un groupe (*cluster*) : $Y_i = j$ $(j = 1, m)$.

5.1) Inerties de partition

- intra-groupes $I_W = \frac{1}{n}\sum_{j=1}^{m}\sum_{x_i:Y_i=j} d_2^2(x_i, \overline{x}_j) = tr(W)$
- inter-groupes $I_B = \frac{1}{n}\sum_{j=1}^{m} n_j d_2^2(\overline{x}, \overline{x}_j) = tr(B)$
- totale $I_T = I_W + I_B = \frac{1}{n}\sum_{i=1}^{n} d_2^2(x_i, \overline{x}) = tr(V)$ car $V = B + W$

5.2) Distances entre groupes

- $\mathcal{D}_{min}(C_i, C_j) = \underset{x \in C_i, y \in C_j}{min} d(x, y)$ *single*
- $\mathcal{D}_{max}(C_i, C_j) = \underset{x \in C_i, y \in C_j}{max} d(x, y)$ *complete*
- $\mathcal{D}_{moy}(C_i, C_j) = \frac{\sum_{x \in C_i}\sum_{y \in C_j} d(x,y)}{n_i \times n_j}$ *average*
- $\mathcal{D}_W(C_i, C_j) = \frac{n_i \times n_j}{n(n_i + n_j)} d^2(\overline{x}_i, \overline{x}_j)$ *Ward*

5.3) Adéquation d'une partition aux données, par ex.

- Indice de Dunn $DI(Y) \frac{min_{1 \leq i < i' \leq c} d(\overline{x}_i, \overline{x}_{i'})}{max_{j=1,c}\Delta_j}$ où $\Delta_j = max_{x_k, x_l \in C_j}\, d(x_k, x_l)$ est le *diamètre* du *cluster* C_j

5.4) Comparaison de deux matrices de partitions P $(n \times c)$ et Q $(n \times c')$ en c et c' *clusters*, à partir de la matrice d'accord $N(P,Q) = {}^tP\,Q = [n_{ij}]_{i=1,c;j=1,c'}$ de dimension $(c \times c')$

- Indice de Rand $RI(P,Q) = \frac{2\,t - (u+v)}{n\,(n-1)} + 1$
  $RI(P,Q) \in [0,1]$ et $RI(P,P) = 1$
- $t = \sum_{i=1}^{c}\sum_{j=1}^{c'} n_{ij}^2 - n$
- $u = \sum_{i=1}^{c} n_{i\bullet}^2 - n$, où $n_{i\bullet} = \sum_{i=1}^{c'} n_{ij}$
- $v = \sum_{j=1}^{c'} n_{\bullet j}^2 - n$, où $n_{\bullet j} = \sum_{j=1}^{c} n_{ij}$

5.5) Méthodes :

- algorithme des centres mobiles qui minimise $I_W = \frac{1}{n}\sum_{j=1}^{m}\sum_{x_i\,i=j} d_2^2(x, \overline{x}_j)$ *K-means*
- classification hiérarchique ascendante *Hierarchical (Agglom.) Clustering*

5.6) Remarques : on peut très bien classifier

- les individus dans l'espace défini par des facteurs, par ex. des composantes principales
- les variables à l'aide d'une distance adaptée, au lieu des individus

6. Classification Supervisée (Prédiction) **tableaux X $(n \times p)$ et Y $(n \times 1)$**

Il s'agit d'apprendre, à partir de X et Y (indicatrice), une règle pour classer une nouvelle observation x (prédiction).

6.1) Remarque : on peut très bien, au préalable

- projeter les données dans un espace de dim. réduite défini par des facteurs (voir 5.), par ex. des comp. principales

6.2) Méthodes

- Plus Proche Prototype *Nearest Prototype*
  (i) prédire pour x le groupe j $(j = 1, m)$ du centre le plus proche au sens d'une distance d
- K-Plus Proches Voisins *K-Nearest Neighbors*
  (i) trouver l'ensemble des K-PPV de x dans X au sens d'une distance d
  (ii) prédire pour x le groupe j $(j = 1, m)$ qui y est majoritairement représenté
- Analyse Discriminante Linéaire *Linear Discriminant Analysis*
  (i) trouver, à partir de X et Y, les facteurs discriminants qui définissent l'espace dans lequel projeter les données
  (ii) projeter, puis prédire pour x le groupe j $(j = 1, m)$ du centre le plus proche au sens de d_2 dans l'espace discriminant (Plus Proche Prototype)

6.3) Mesures de performance sur un jeu test (X_T, Y_T), une fois les prédictions Y_p réalisées sur X_T :

- taux d'accord $T = \frac{\sum_{i=1}^{m} c_{ii}}{n}$ et de désaccord $E = 1 - T$ où $C = [c_{ij}]_{i,j=1,m}$, appelée *matrice de confusion*, est la table de contingence croisant la prédiction Y_p et Y_T.
- binaires (groupe j) à partir des nombres de : *True Positives*, *True Negatives*, *False Positives* et *False Negatives*
  - *rappel* $R_j = \frac{TP}{TP+FN}$, *précision* $P_j = \frac{TP}{TP+FP}$ et *fausse alarme* $FA_j = \frac{FP}{TP+FP} = 1 - P_j$
  - *justesse* $A_j = \frac{TP+TN}{n}$ et *F1-score* $F1_j = 2\,\frac{R_j \times P_j}{R_j + P_j}$
- globales : $S = \sum_{j=1}^{m} \frac{n_j}{n} S_j$

6.4) Sélection d'attributs : mesures de pertinence pour chaque variable X_i $(i = 1, p)$, par ex.

- statistique de Fisher $F = \frac{s_B^2/(m-1)}{s_W^2/(n-m)}$ *AnOVa (Analysis Of Variance)*