BE 3 –Analyses factorielles (AFC et ACM)

et classification non supervisée

Les objectifs de ce bureau d'étude sont de réaliser dans un premier temps une analyse factorielle des correspondances binaires (AFC), ainsi qu'une analyse factorielle des correspondances multiples (ACM). Dans un second temps, une classification non supervisée sera réalisée en utilisant deux méthodes : une classification ascendante hiérarchique, ainsi que l'algorithme des centres mobiles. Enfin, une combinaison des méthodes d'analyse factorielles et de classification non supervisée sera réalisée afin d'enrichir l'étude de tableaux de données.

1. AFC

La démarche à suivre pour réaliser une AFC est présentée en cours. Nous allons ici en reprendre les étapes pour l'analyse des tableaux de données fournis.

1.1) Lecture des données
Le fichier « csp.mat » contient les données à traiter (CSP des parents d'étudiants de diverses disciplines). La signification des codes utilisés est donnée en annexe.

1.2) Construction du tableau des fréquences relatives

1.3) Construction des tableaux des profils-lignes et des profils-colonnes

1.4) Calcul des matrices X, M et D
- X : matrice calculée à partir des fréquences relatives et des fréquences marginales
- D : matrice des poids des profils-lignes
- M : matrice des poids des profils-colonnes

1.5) Calcul des axes factoriels et des valeurs propres associées

1.6) Calcul des facteurs associés aux axes factoriels
- Pour le nuage des profils-lignes
- Pour le nuage des profils-colonnes

1.7) Calcul des indices d'aide à l'interprétation des résultats de l'analyse factorielle :
- Le pourcentage d'inertie associé à chaque axe (afficher le diagramme des valeurs propres)
- La contribution des profils-lignes et des profils-colonnes pour chaque axe
- La qualité de représentation des profils-lignes et des profils-colonnes pour chaque axe

1.8) Visualisation des résultats (par soucis de simplification, on ne s'intéressera qu'au premier plan factoriel)
- Projeter les profils-lignes sur le premier plan factoriel
- Projeter les profils-colonnes sur le premier plan factoriel

1.9) Interprétation des résultats
- Donner une interprétation aux deux premiers facteurs
- Réaliser une typologie des profils-lignes
- Réaliser une typologie des profils-colonnes
- Peut-on lier ces deux typologies ?

2. ACM

La démarche à suivre pour réaliser une ACM est présentée en cours. Nous allons ici en reprendre les étapes pour l'analyse des tableaux de données fournis.

2.1) Lecture des données

Le fichier « pommes.mat » contient les données à traiter (caractéristiques de variétés de pommes). La signification des codes utilisés est donnée en annexe.

2.2) Construction du tableau disjonctif complet (TDC)

2.3) Construction du tableau de Burt

2.4) Calcul des matrices X, M et D
- X : matrice calculée à partir du TDC
- D : matrice des poids des individus
- M : matrice des poids des modalités

2.5) Calcul des axes factoriels et des valeurs propres associées

2.6) Calcul des facteurs associés aux axes factoriels
- Pour le nuage des individus
- Pour le nuage des modalités

2.7) Calcul des indices d'aide à l'interprétation des résultats de l'analyse factorielle :
- Le pourcentage d'inertie associé à chaque axe (afficher le diagramme des valeurs propres)
- La contribution des individus, des modalités et des variables pour chaque axe
- La qualité de représentation des individus pour chaque axe

2.8) Visualisation des résultats (par soucis de simplification, on ne s'intéressera qu'au premier plan factoriel)
- Projeter les individus sur le premier plan factoriel
- Projeter les modalités sur le premier plan factoriel

2.9) Interprétation des résultats
- Donner une interprétation aux deux premiers facteurs
- Réaliser une typologie des individus
- Réaliser une typologie des modalités
- Peut-on lier ces deux typologies ?

3. Classification ascendante hiérarchique

La méthode de classification ascendante hiérarchique (CAH) est présentée en cours. Le programme que vous allez mettre en œuvre pour réaliser cette classification devra être constitué de 4 étapes :
1) Lecture des données
2) Calcul de la matrice des distances entre chaque paire d'éléments de cet ensemble de données (en utilisant la fonction Matlab « pdist »)
3) Construction de l'arbre hiérarchique par agrégations successives de deux éléments (un utilisant la fonction « linkage »)
4) Coupure de l'arbre hiérarchique pour obtenir une partition (en utilisant la fonction « cluster »)

Le résultat obtenu à l'issue de chacune de ces étapes devra être affiché.

4. Algorithme des centres mobiles

La méthode des centres mobiles a également été étudiée en cours. L'objectif ici est de la mettre en œuvre dans un programme. Pour cela, la fonction Matlab « kmeans » pourra être utilisée.

5. Travail à rendre

Comme cela a été vu dans le cours, il peut être très intéressant de combiner une analyse factorielle avec une méthode de classification. C'est l'objectif du travail que vous avez à rendre. Le principe est le suivant :
1) Réalisation d'une analyse factorielle
2) Classification à partir des facteurs

3) Positionnement des centres de classes dans le plan factoriel, et mise en évidence pour chaque point de leur classe d'appartenance (par une couleur différente par exemple)

Cette méthode mixte devra être réalisée pour quatre combinaisons :
- ACP + CAH
- ACP + centres mobiles
- ACM + CAH
- ACM + centres mobiles

Afin d'évaluer ces méthodes, vous pourrez analyser le jeu de données « population » utilisé lors du deuxième BE. L'ACM portant sur des variables qualitatives, les variables quantitatives de ces données devront donc être converties par une mise en classes, ce qui a l'avantage de permettre une analyse des données selon un point de vue différent.
Enfin, concernant la CAH, différentes stratégies peuvent être utilisées pour l'agrégation des éléments et la coupure de l'arbre. Les paramètres des fonctions « linkage » et « cluster » peuvent donc être modifiés pour mettre en œuvre la stratégie adoptée. Afin d'évaluer les influences de ces choix sur la classification, vous utiliserez plusieurs paramétrages (au moins deux).

Vous disposez de 15 jours pour terminer ce travail. Vous devrez alors rédiger un compte-rendu électronique au format pdf indiquant votre démarche, vos résultats (tableaux de données et graphes), vos commentaires et vos interprétations des données. Ce rapport ainsi que votre programme Matlab devront être regroupés dans un fichier archive (.zip) qui sera déposé sur le site pedagogie (rubrique « Travaux »➔ « Analyses factorielles et classifications non supervisées »).

6. Annexes

3.1) Jeu de données pour l'AFC

Analyse des types d'études en fonction des catégories socio-professionnelles (CSP) des parents.

Codes des CSP :
Age : Agriculteur exploitant
Oua : Ouvrier agricole
Ind: Industriel
Art: Artisan
Gco : Moyen et gros commerçant
Lib : Profession libérale
Ing : Ingénieur
Adm : Cadre de l'Administration
Mes : Profession médicale et salariée
Mso : Profession médicale et sociale
Tec : Technicien
Cam : Cadre administratif moyen
Ins : Instituteur
Emb : Employé de bureau
Con : Contremaître
Ous : Ouvrier spécialisé
Min : Mineur
Mar : Pêcheur
Man : Manoeuvre
Ser : Personnel de service
Arm : Armée, Police

Codes des disciplines :
dro : Droit
sci : Sciences
let : Lettres
med : medecine
pha : pharmacie
iut : IUT

3.2) Jeu de données pour l'ACM

Etude des propriétés de différentes variétés de pommes.

Codes des propriétés des pommes :
Arb : forme de l'arbre, caractérisée par 4 modalités (1=colonnaire, 2=spur, 3=étalé, 4=très étalé)
Rec : date de récolte, caractérisée par 3 modalités (1=précoce, 2=intermédiaire, 3=tardif)
Cal : calibre du fruit, caractérisé par 3 modalités (1=petit ou moyen, 2=gros, 3=très gros)
Coul : couleur dominante du fruit ou intensité de la couleur rouge additionnelle, caractérisée par 4 modalités (1=jaune ou vert, 2=rouge-orangé, 3=rouge, 4=rouge-violacé)
Pour : proportion de coloration rouge additionnelle, caractérisée par 4 modalités (1=absente, 2=faible, 3=moyenne, 4=forte)
Type : type de coloration du fruit, caractérisé par 3 modalités (1=lavé, 2=lavé-strié, 3=strié)
Form : forme du fruit, caractérisée par 3 modalités (1=allongé, 2=intermédiaire, 3=aplati)
Ferm : fermeté du fruit, caractérisée par 3 modalités (1=peu ferme, 2=moyen, 3=très ferme)
Gout : rapport sucre / acidité, caractérisé par 3 modalités (1=doux, 2=équilibré, 3=acidulé)