Université Rennes 2 - UFR Sciences sociales
Master 1 MAS

Année 2018-2019
Madison Giacofci

TP n° 1 - Classification
Méthodes hiérarchiques et centres mobiles

Les fichiers de données nécessaires au TP peuvent être téléchargés sur la plate-forme Cursus (espace cours "Classification", partie "TP").

1 Distances, similarités et dissimilarités

Dans les fonctions R de bases, la fonction adaptée au calcul de matrice de distances (ou dissimilarités) à partir d'un jeu de données est la fonction `dist()`.

1. En utilisant l'aide de R, identifier les différentes dissimilarités pouvant être calculées à l'aide de la fonction `dist`.

1.1 Données binaires

Le jeu de données `ZacharyBinaryData.txt` représente les relations sociales de 34 membres d'un club de karaté (interactions en dehors du club). Chaque ligne correspond à un membre de la communauté tandis que la $i$ème colonne indique si ce membre interagit ou non (valeur respective de 1 ou 0) avec le membre i.

2. Quel type de dissimilarité vous semblerait adaptée pour ce type de données ?
3. Créer une fonction permettant de calculer l'indice de Dice entre deux individus.
4. À l'aide des fonctions `outer` et `Vectorize`, calculer la matrice des dissimilarités de Dice associée au jeu de données `ZacharyBinaryData.txt`. On peut ensuite transformer la matrice de dissimilarités créée en un objet de type `dist` à l'aide de la fonction `as.dist`.

1.2 Données qualitatives

De la même manière, la distance du χ^2 entre profils-lignes n'est pas proposée dans la fonction `dist` et nous allons la programmer. Nous nous intéressons pour cela au jeu de données `breastData.txt` concernant des patientes atteintes de cancer du sein. Plusieurs variables génériques sur les patientes sont données (âge, statut ménopause) ainsi que des variables liées à la maladie (taille de la tumeur, nombres de noeuds lymphatiques metastasés, capsule lymphatique metastasée, degré de malinité, sein affecté, quart affecté, irradiation).

5. Pour l'ensemble du jeu de donnees, construire le tableau disjonctif complet associée à l'ensemble des variables (on pourra utiliser la fonction `acm.disjonctif` du package `ade4` ou la fonction `tab.disjonctif` du package `FactoMineR`).
6. Créer une fonction permettant de calculer la distance du χ^2 entre deux individus à partir du tableau disjonctif complet.
7. Quel est le lien avec la fonction `dist.dudi` du package `ade4` ?

1.3 Données quantitatives

Enfin, un dernier type de distance utile mais moins conventionnel : la distance issue de la corrélation.

8. Rappeler son intérêt en tant que mesure de distance.
9. On considère le jeu de données `loup_chien.dat` contenant des mesures pour 30 crânes de chiens domestiques de haute taille et 12 crânes de loups. Sur ces crânes, on a mesuré la longueur de la carnassière supérieure, la longueur et la largeur de la première molaire supérieure (respectivement notées LoC, LoM et LaM). Calculer la matrice de distance pour les données `loup_chien` pour la méthode des corrélations. On se servira des fonctions `cor` et `as.dist`.

2 Classification automatique

2.1 Données Iris

On s'intéresse dans un premier temps à un jeu de données test appelé `iris` (que l'on peut charger à l'aide de la commande `data(iris)`. Ce jeu de données concerne l'étude de fleurs pour lesquelles on dispose des variables `Sepal.Length`, `Sepal.Width`, `Petal.Length` et `Petal.Width`, mesurant respectivement les longueur et largeur des sépales et les longueur et largeur des pétales. La variable `Species` est la variable d'appartenance à une espèce. On souhaite savoir si la classification nous permet de retrouver la partition induite par la variable `Species`.

Classification ascendante hiérarchique

1. Réalisez une première analyse des données. Vous semble-t-il nécessaire de procéder à une standardisation des données ? Vous justifierez votre réponse.
2. La fonction `hclust` permet de réaliser une CAH à partir d'une matrice de distance et pour plusieurs critères d'agrégation. Effectuer la classification hiérarchique des données `iris` en utilisant la distance euclidienne et le critère de Ward. Analyser les sorties du logiciel et tracer le dendrogramme correspondant (fonction `plot` appliquée à l'objet issu de `hclust`).
3. Tracer la courbe de perte de l'inertie inter-classes. Combien de groupes choisiriez-vous ? Afficher les groupes avec la fonction `cutree`.
4. Faire apparaître les groupes sur le dendrogramme à l'aide de la fonction `cutree`.
5. Comparer la classification obtenue avec la variable `Species`. Commentez les résultats obtenus.
6. Représentez les résultats à l'aide de la fonction `clusplot` du package `cluster`. À quoi correspond cette représentation ?

Agrégation autour de centres mobiles

1. La fonction `kmeans` donne le résultat de l'algorithme d'agrégation autour des centres mobiles. En utilisant l'aide de la fonction, étudiez les principaux paramètres nécessaires à son utilisation.
2. Effectuer la partition des données `iris` en 3 groupes. Analyser les sorties de la fonction et représenter les résultats à l'aide de graphiques appropriés.
3. Relancer plusieurs fois la procédure. Que constatez-vous ?
4. Programmer une fonction permettant de faire tourner l'algorithme avec 50 différentes initialisations et de retenir celle conduisant à la meilleure inertie intra-classes.

2.2 Données loup/chien

1. Faire les représentations graphiques des variables deux à deux. Par la suite, pour les représentations graphiques, on représentera `LoC` en fonction de `LoM` et `LaM` en fonction de `LoM`.
2. Effectuer la CAH avec les méthodes du saut maximal et de Ward sur le jeu de données privé de l'individu pour lequel l'espèce est inconnue. Combien de classes choisirait-on ?
3. Représenter les résultats de la classification. Ajouter l'individu "inconnu". D'après vous, quelle pourrait être son espèce ?
4. Faire de même pour la méthode des k-means. Faire tourner l'algorithme plusieurs fois pour différents nombres de classes. Commentez vos résultats.

2.3 Données Citycrime

Le jeu de données `citycrime.dat` contient le nombre de crime pour 100000 habitants suivant leur type dans 16 villes américaines. Appliquer les différentes méthodes de classification comme pour les applications précédentes. On justifiera soigneusement les choix faits pour chacune d'entre elles. Interprétez les résultats obtenus.

2.4 Données de cancer du sein

On reprend à présent le jeu de données `breastData.txt` concernant des patientes atteintes de cancer du sein.

1. En reprenant la matrice de distance calculée dans la première partie du TP, effectuer le partitionnement des individus en classe.
2. Analyser les résultats obtenus en utilisant des projections factorielles. On rappelle que pour le cas de données qualitatives, la technique adaptée pour le calcul des projections factorielles est l'AFCM. On pourra utiliser les fonctions `dudi.coa` (`ade4`) ou `MCA` (`FactoMineR`).
3. Analyser les résultats obtenus variable par variable. On pourra utiliser des tables de fréquences, des tests du χ^2 ou les fonctions du package `FactoMineR`.