Code

TD classification non supervisée (sujet)

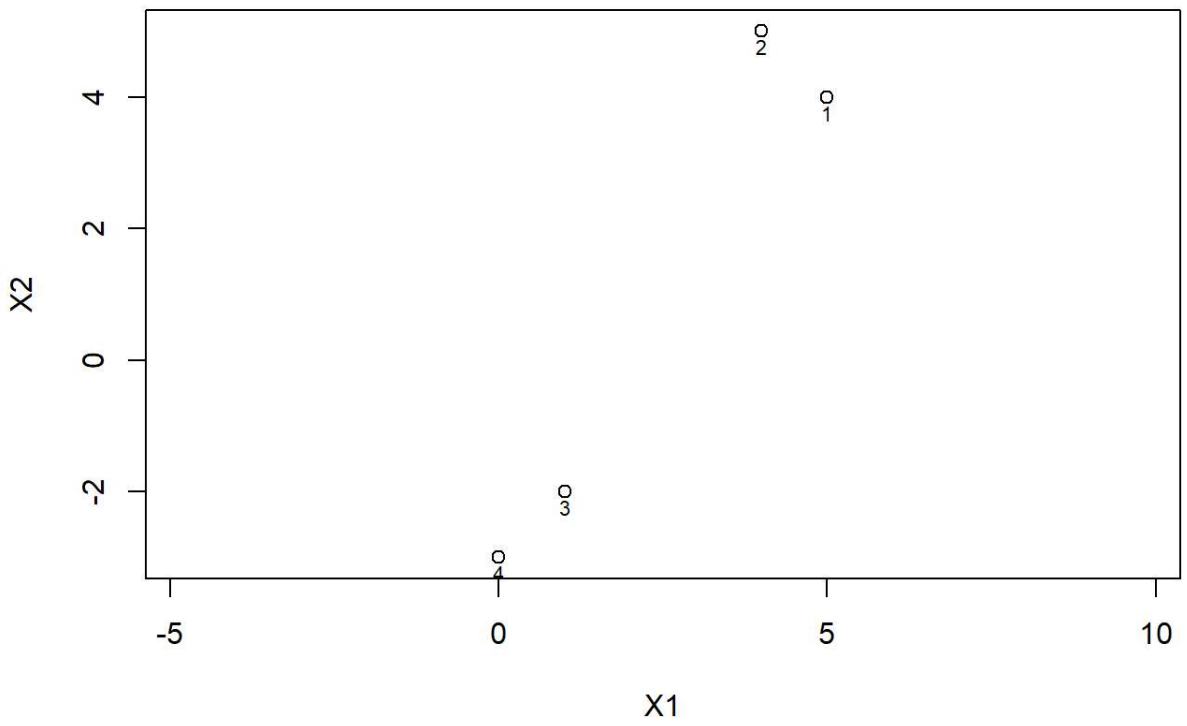
Marie Fourcot

01/02/2022

1 Clustering manuel

Soit X une matrice avec 4 observations pour 2 variables. Chaque observation est pondéré d'un poids $\omega_i = 1$.

	X1	X2
1	5	4
2	4	5
3	1	-2
4	0	-3



1.1 k-means

1 . Appliquez à la main l'algorithme des k-means avec $K = 2$ clusters et en prenant les individus 1 et 2 comme centres initiaux.

2 . Utilisez la fonction R `kmeans()` pour répéter la question précédente.

1.2 Classification hiérarchique ascendante

3 . Réalisez à la main une classification hiérarchique ascendante en utilisant la distance euclidienne et le lien maximal (complete linkage) comme distance d'agrégation.

4 . Utilisez les fonctions R `dist(), hclust() et plot()` pour répéter la question précédente.

5 . Construisez maintenant la classification hiérarchique ascendante en utilisant toujours le lien maximal comme distance d'agrégation, mais la distance de Manhattan.

2 Clustering avec R

2.1 Données

Les données proviennent d'une étude de nutrition chez la souris. Elles ont été publiées par Martin et al. en 2007 et sont disponibles dans le package `mixOmics` , je vous les ai mises sur Moodle vu que nous n'allons pas nous servir de ce package par ailleurs.
Pour 40 souris, nous disposons :

- des données d'expression de 120 gènes recueillies sur membrane nylon avec marquage radioactif,
- des mesures de 21 acides gras hépatiques.

Par ailleurs, les 40 souris sont réparties selon deux facteurs :

- Génotype (2 modalités) : les souris sont soit de type sauvage (wt) soit génétiquement modifiées (PPAR) ; 20 souris dans chaque cas.
- Régime (5 modalités) : les 5 régimes alimentaires sont notés `ref, efad, dha, lin, tournesol` ; 4 souris de chaque génotype sont soumises à chaque régime alimentaire.

L'objectif est de regrouper les gènes en classes homogènes afin d'aider le biologiste à les associer à des fonctions ou métabolismes de la lipogénèse.

2.2 Classification hiérarchique ascendante

2.2.1 Classification des gènes

Nous allons construire une classification hiérarchique ascendante des gènes en utilisant deux distances différentes :

- la distance euclidienne, classique
- la corrélation.

Nous allons donc obtenir deux arbres différents.
Dans les deux cas, nous utiliserons la méthode de Ward comme distance d'agrégation.

2.2.1.1 Distance euclidienne

6. Construisez l'arbre en utilisant la distance euclidienne et la distance de Ward comme distance d'agrégation.
   On utilise la transposée de notre tableau de données pour le calcul de la distance, en effet on souhaite classifier les gènes et non les souris.

7. Coupez l'arbre pour obtenir une classification.

Un choix important sur un dendogramme est la hauteur où on le coupe, qui va définir le nombre de classes obtenues.
On peut choisir à l'oeil, peu évident dans notre cas, ou s'aider d'un graphe de la hauteur des noeuds.

8. Triez les gènes selon leur classe et affichez les gènes de la deuxième classe.

2.2.1.2 Corrélations

Pour avoir une mesure qui ressemble plus à une distance, soit un éloignement, on va utiliser 1 - corrélation.

9. Construisez l'arbre en utilisant la distance basée sur la corrélation et la distance de Ward comme distance d'agrégation et le couper à un niveau pertinent.

10 (optionnel). On peut aussi utiliser une distance basée sur le carré de la corrélation, en effet, le signe de la corrélation ne nous intéresse pas forcément.

2.2.2 Classification des souris

Nous souhaitons maintenant obtenir une classification hiérarchique ascendante des souris.

11. Construisez l'arbre en utilisant la distance euclidienne et la distance de Ward comme distance d'agrégation.
    Le découper en un nombre de clusters qui vous semble approprié.

12. Utilisez la fonction `colored_bars` pour afficher sous le dendrogramme les variables diet et phenotype. Pensez-vous que ces variables puissent expliquer les clusters?

2.3 K-means

13. Utilisez les k-means (centres mobiles) avec 6 clusters et 5 itérations. Et interprétez la sortie fournie par un print() de la classification.

14. Comparez la classification obtenue avec la classification ascendante hiérarchique avec la distance euclidienne et les k-means.

2.3.1 Détermination du nombre de clusters

Les k-means, à la différence de la CAH, ne fournissent pas d'outil d'aide à la détection du nombre de classes. Nous devons les programmer sous R ou utiliser des procédures proposées par des packages dédiés.
Le schéma est souvent le même : on fait varier le nombre de groupes et on surveille l'évolution d'un indicateur de qualité de la solution c'est-à-dire l'aptitude des individus à être plus proches de ses congénères du même groupe que des individus des autres groupes. Pour cela, on étudie l'évolution de la proportion d'inertie expliquée par la partition, on cherche le «coude» dans le graphique.

2.4 Méthode mixte

On va réaliser un partitionnement avec le k-means puis une classification ascendante hiérarchique à partir des centres obtenus.