Sorbonne Université
Master Ingénierie mathématique (ISDS)

Modèles à structure latente
Jean-Patrick Baudry – 2021-2022

TP 1 : Données, Kmeans, CHA

L'objectif de ce TP est de vous mettre en garde contre une application trop systématique ou aveugle de l'ACP dans une étude de clustering, et d'explorer et comparer le comportement des Kmeans et de la classification hiérarchique ascendante. Outre les questions posées ci-dessous, n'hésitez pas à laisser libre cours à votre curiosité.

Données, ACP

Travaillons sur les données *crabs*, qui me semblent bien illustrer le danger possible de l'utilisation de l'ACP.

1. Chargez les données (*data(crabs, package = "MASS")*), consultez l'aide associée. L'espèce et le sexe peuvent être considérés comme des classifications (possiblement à croiser pour obtenir quatre classes). Nous choisissons en tout cas de ne pas en tenir compte ici : supprimez les !
2. Étudiez successivement :
   (a) La proportion de variance de ces données expliquée respectivement par leur première et deuxième composante principale.
   (b) La projection de ces données sur leur première composante principale.
   (c) La projection de ces données sur leur deuxième composante principale.
   (d) La projection de ces données sur leurs deux premières composantes principales.
3. Concluez !

Kmeans et clustering hiérarchique ascendant

Les données *Seeds* Je vous propose de travailler sur les données *Seeds* que vous pouvez récupérer sur *UCI MLR Repository*. Vous pouvez choisir un autre jeu de données du même type (quelques variables continues, quelques centaines d'observations) si vous préférez !

Familiarisez-vous avec ces données, notamment en en considérant plusieurs représentations graphiques (histogrammes et nuages de points en dimensions 2 et 3 des données originales ou des données projetées sur les composantes principales...).

Fixez un objectif de clustering (c'est-à-dire : posez une question précise, à laquelle vous pensez essayer de répondre par l'étude de clustering qui vient. Votre analyse des résultats des différentes méthodes dépendra du choix de cet objectif).

Kmeans Utilisez la méthode des Kmeans pour obtenir une première classification. Quels sont les paramètres auxquels il faut faire attention lors de la mise en œuvre de cette méthode ?

Vous pourrez aussi représenter graphiquement l'évolution des affectations des classes et des centres de celles-ci au cours des itérations de l'algorithme, pour comprendre comment il fonctionne et visualiser sa convergence. Il n'est pas nécessaire de recoder l'aglorithme pour cela : il suffit de limiter le nombre d'itérations dans une fonction dédiée de R (et ainsi de représenter par exemple ce qu'il se passe après 1, 2, 3, ... itérations avec la initiatisation à chaque fois).

Clustering hiérarchique ascendant Mêmes objectifs, même question, pour le clustering hiérarchique ascendant. Vous pourrez considérer différentes fonctions de linkage, notamment celles vues en cours (Ward, single linkage, complete linkage, group average).

Comparer les résultats des Kmeans et du CHA En fonction de l'objectif de clustering que vous vous étiez fixé... Vous pourrez utiliser par exemple des représentations graphiques, des proportions d'accord entre différentes classifications (voir ci-dessous), l'indice de Rand (éventuellement dans sa version ajustée), etc.

Pour aller plus loin Vous pourrez essayer de comparer les résultats obtenus avec une procédure de miniminisation de la variance intra-classe exhaustive à ceux obtenus avec les Kmeans ou avec la CHA (avec un linkage bien choisi). Pour quelles tailles d'échantillon est-ce que cela reste possible ?