Université Grenoble Alpes

Master SSD, BDA et IC
Enseignant : Jonathan El Methni
Apprentissage Statistique

TP 3 : Examen mi-parcours noté

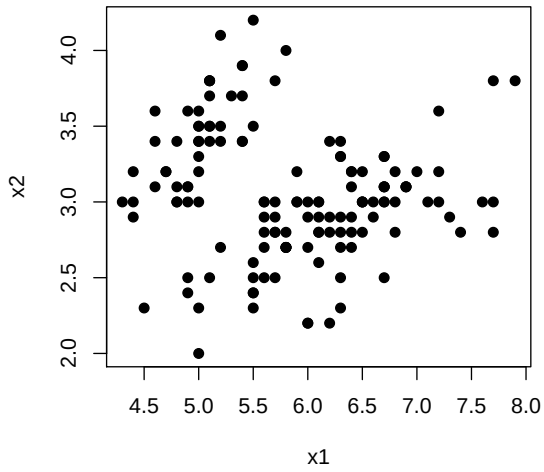
L'objectif de cet examen est de reproduire des analyses non-supervisées étudiées en cours et en TP. Vous serez évalués sur la qualité de vos interprétations et de vos sorties graphiques.

Vous devrez me rendre :

1. un compte-rendu au format pdf de 10 pages maximum (sans compter les éventuelles annexes),
2. le code mis en oeuvre, commenté un minimum, dans un fichier séparé (et pas dans le compte rendu).

Exercice 1 : k-means

L'objectif de cet exercice est d'apprendre à mettre en oeuvre l'algorithme k-means grâce à la fonction `kmeans` disponible dans `R`. Nous travaillerons pour cela à partir du jeu de données représenté ci-dessous.



1. Charger le fichier `exo-1.Rdata` et représenter le jeu de données, stocké dans la matrice `X` (une matrice à deux colonnes), tel que ci-dessus.
2. Consulter la documentation de la fonction `kmeans`. Quels sont les arguments obligatoires pour l'appel à cette fonction ? Quelle(s) donnée(s) clé renvoie t-elle ?
3. Réaliser un clustering du jeu de données en 3 clusters. Représenter le résultat obtenu en utilisant un code couleur pour refléter l'appartenance des points aux différents clusters. Faire apparaître les centroïdes en utilisant le même code couleur, mais un symbole différent.
4. A quoi sert l'argument `nstart` ? Quel est son intérêt ? Le démontrer empiriquement (et graphiquement) en considérant 5 clusters.

5. Quel critère permet de juger de la qualité du clustering ? Comment est-il défini ? Tracer son évolution quand on considère 1 à 10 clusters et commenter le résultat. Le choix de 3 clusters vous semble-t-il raisonnable ? Pourquoi ?
6. Le vecteur `y` contient les vraies catégories – entre 1 et 3 – des instances du jeu de données. Mesurer le taux d'agréement entre ces vraies catégories et le clustering proposé par k-means.
7. Enfin, la matrice `X.test` contient 12 nouvelles observations. Implémenter une procédure permettant de les affecter aux différents clusters obtenus précédemment.

Exercice 2 : clustering & formes

On considère les trois jeux de données représentés dans la figure ci-dessous. Chaque jeu de données est stocké dans un fichier texte à 3 colonnes : les deux premières colonnes contiennent les deux variables x_1 et x_2, et la troisième définit les catégories des points (correspondant aux couleurs dans la figure).

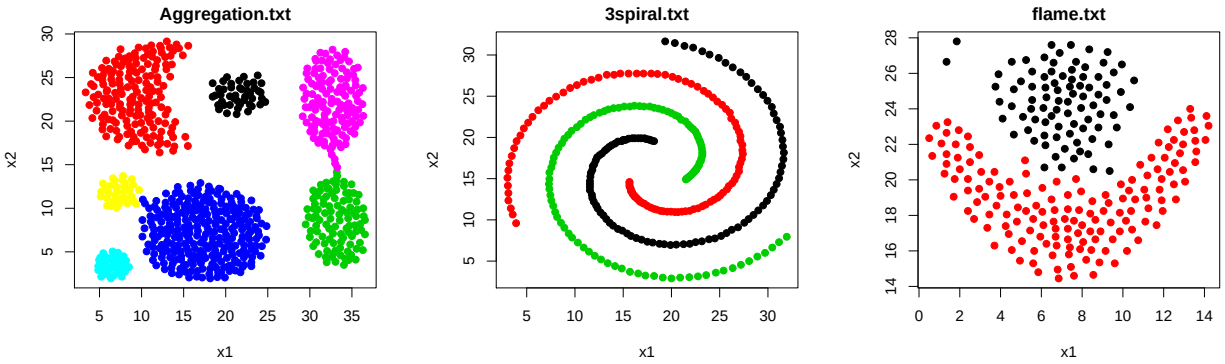


FIGURE 1 – Jeux de données considérés : `Aggregation.txt`, `3spiral.txt` et `flame.txt`.

Appliquer k-means et le clustering hiérarchique (en considérant différents critères de "linkage") sur les différents jeux de données. Identifier – si elle existe – une configuration permettant de retrouver la bonne catégorie des points, et expliquer pourquoi on n'y parvient pas dans certaines cas.

Exercice 3 : analyse exploratoire

On considère un jeu de données lié à la microbiologie. Chaque observation est une bactérie que l'on a caractérisé par spectrométrie de masse. Cette technologie vise à donner une empreinte protéomique de la bactérie, et conduit à 834 mesures.

Le jeu de données contient 429 bactéries dont on connaît l'*espèce* et le *genre*. Ces deux informations permettent de regrouper les bactéries en catégories, à différents niveaux de granularité : les 429 bactéries du jeu de données sont réparties en 8 genres bactériens, et 16 espèces bactériennes. Une espèce fait par définition partie d'un genre donné, et nous avons ici deux espèces par genre.

L'objectif de l'exercice est d'évaluer si on peut regrouper les bactéries en genres ou espèces par des approches (non-supervisées) de clustering.

Le jeu de données se constitue donc de 429 observations en 834 dimensions, réparties en 8 genres et 16 espèces bactériennes. Les données sont stockées dans deux fichiers texte :

— `spectra.txt` est un fichier tabulé contenant les spectres (429 lignes, 834 colonnes)
— `meta-data.txt` est un fichier tabulé contenant les méta-données (2 colonnes : genre et espèce bactérienne de chaque spectre - se référer au "header")

1. Charger le jeu de données, et représenter le nombre d'observations disponibles au sein de chaque genre.
2. Réaliser une analyse en composantes principales du jeu de données. Le représenter selon ses deux premières composantes, avec un code couleur indiquant le genre bactérien et commenter le résultat. Qu'observe t-on quand on considère les composantes d'ordre 3 et 4 ?
3. Réaliser un clustering hiérarchique du jeu de données. En extraire 8 clusters et analyser leur constitution selon les différents genres bactériens. Quelle stratégie d'agglomération vous semble la plus adaptée ? Ces résultats sont-ils cohérents avec l'ACP ?
4. Effectuer la même analyse avec l'algorithme k-means et comparer les résultats obtenus.
5. Enfin, identifier un genre bactérien pour lequel la séparation inter-espèces est claire et un genre bactérien pour lequel ce n'est pas le cas. Illustrer vos résultats par une (ou des) représentation(s) graphique(s).