TP5 : Clustering

Objectif

L'objectif de ce TP est de mettre en pratique les algorithmes de classification non supervisée:

- K-means
- PAM
- Classification Hiérarchique
- DBSCAN

---

1. Exploration de la base de données IRIS

La base de données `IRIS` contient cinq variables :

- `Sepal.Length`
- `Sepal.Width`
- `Petal.Length`
- `Petal.Width`
- `Species` (3 classes : `setosa`, `versicolor`, `virginica`)

Taper la commande suivante pour plus d'informations:

```
str(iris)
```

Exercice 1

1. Appliquer l'analyse en composantes principales (commande `prcomp`) sur iris (stocker le résultat dans un objet nommé `iris_pca`).
2. Visualiser les données en utilisant les deux premières composantes principales et en les colorant selon `Species`.
3. Comment interpréter la séparation entre les classes?

---

2. Clustering par K-means

Exercice 2

1. Appliquer la standardization sur la base des données `iris` (stocker le résultat dans un objet nommé `iris_std`).
2. Appliquer l'algorithme des K-Moyennes (K-Means) en utilisant la commande `kmeans` avec k = 3 sur les données normalisées.
3. Visualiser le résultat (nuage de points coloré par cluster).
4. Le clustering correspond-il aux vraies classes ? Pourquoi ?

---

3. Clustering par PAM (Partitioning Around Medoids)

Exercice 3

Faire la même chose avec l'algorithme PAM (commande `pam` de la bibliothèque `cluster`).

---

4. Classification Hiérachique

Exercice 4

Faire la même chose avec l'algorithme de la classification hiérarchique (commande `hclust`):

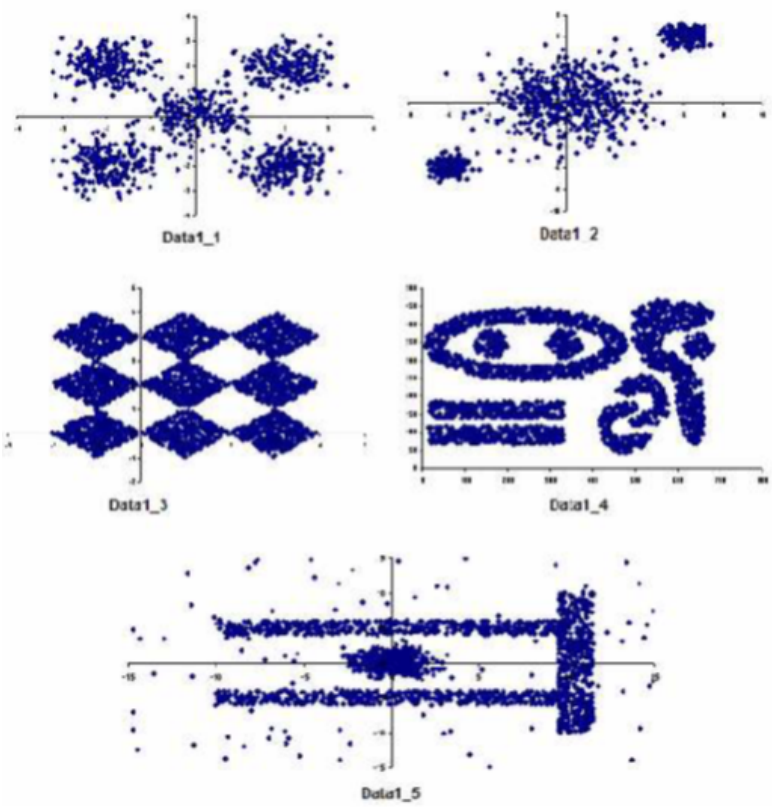
- Utiliser la commande `dist` pour créer la matrice des distances.
- Utiliser la commande `cutree` pour couper l'arbre et obtenir 3 clusters.

---

5. Bases de données synthétiques

Description :

- **Data1_1** : 5 clusters sphériques, 1000 points
- **Data1_2** : 3 clusters sphériques densités différentes, 900 points
- **Data1_3** : 9 clusters, 3000 points
- **Data1_4** : 9 clusters formes variées, 3030 points
- **Data1_5** : 2 clusters + 5% de bruit, 2100 points



Utiliser la commande suivante pour charger les fichiers csv dans R:

```
library(readr)
library(dplyr)
data1_1 <- read_csv("path/to/Data1_1.csv") %>%
    mutate(class = as.factor(class))
summary(data1_1)
```

Exercice 5

1. Appliquer `kmeans`, `pam` et `hclust` (en choisissant le niveau de coupure adéquat) et tracer les quatres graphiques (avec la classification originale) pour chaque dataset.
2. Le clustering correspond-il aux vraies classes ? Pourquoi ?
3. Appliquer DBSCAN (commande `dbscan` de la bibliothèque `dbscan`) sur chaque dataset.
