TP Clustering
5e SDBD

MJ. HUGUET
homepages.laas.fr/huguet

Objectifs

Le but de ces TP est mettre en oeuvre et de comparer différents algorithmes de clustering tout d'abord à partir de quelques méthodes fournies par `scikit`***learn*** puis en utilisant une méthode externe :

- k-Means
- clustering hiérarchique (agglomératif)
- `DBSCAN`
- `HDBSCAN`

Nous utilisons des jeux de données "artificiels" en seulement 2 dimensions pour des raisons pédagogiques. En effet, en visualisant ces exemples, il est souvent assez évident de déterminer le bon nombre de clusters à obtenir.

Encadrants

Marie-José Huguet, Mohamed Siala, Julien Ferry, Hao Hu

1 Jeux de données

Les jeux de données sont disponibles sur le site : `https://github.com/deric/clustering-benchmark`. Seuls les jeux de données "artificiels" seront considérés dans ces TP.

Travail à réaliser

Le premier travail à réaliser est de lire ces jeux de données et de les visualiser sous forme d'une grille 2D avec les points. Vous pouvez "parser" les données en récupérant le package `arf` : `from scipy.io import arff`. La commande suivante permet de récupérer un tableau `numpy` :
`data = arff.loadarff(open('file.arff','r'))`

Note : dans les jeux de données, pour chaque exemple, la dernière colonne fournit le numéro de cluster (sans précision sur la méthode utilisée pour l'obtenir). Cela peut permettre d'afficher chaque cluster avec une couleur différente. En pratique, **vous ne devez pas utiliser cette colonne** car on suppose que les clusters ne sont pas connus.

2 Clustering k-Means

2.1 Intérêts de la méthode

Choisissez quelques (2 ou 3) jeux de données pour lesquels il vous semble que la méthode k-Means devrait identifier correctement les clusters.
Le travail à réaliser est le suivant :

- Appliquez la méthode k-Means en lui donnant directement le nombre de clusters attendus (utilisez l'initialisation k-means++)

On considère maintenant qu'il peut être possible de déterminer "automatiquement" le bon nombre de clusters. Identifiez pour cela dans la documentation de scikitlearn [1] les métriques qui vous semblent pertinentes.

- Appliquez itérativement la méthode précédente pour déterminer le bon nombre de clusters à l'aide de métriques d'évaluation sélectionnées
  - Mesurez le temps de calcul
  - Arrivez-vous à retrouver le résultat attendu à l'aide de ces métriques d'évaluation ?

2.2 Limites de la méthode

Choisissez quelques (2 ou 3) jeux de données pour lesquels il vous semble que la méthode k-Means aura des difficultés pour identifier correctement les clusters.

- Appliquez la méthode k-Means sur ces jeux de données pour confirmer vos choix.

Vous devez avoir identifié des limites de la méthode. Avez-vous besoin d'utiliser une autre métrique d'évaluation ?

1. `https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation`

3 Clustering agglomératif

3.1 Intérêts de la méthode

Choisissez quelques (2 ou 3) jeux de données pour lesquels il vous semble que la méthode de clustering aggomératif devrait identifier correctement les clusters.

- Appliquez une méthode de clustering agglomératif en lui donnant le nombre de clusters attendus
  - Considérez différentes manières de combiner des clusters (single, average, complete, ward linkage), uniquement pour la distance euclidienne. Par défaut l'option `connectivity` est laissée à `none`.
- Appliquez itérativement la méthode précédente pour déterminer le bon nombre de clusters à l'aide de métriques d'évaluation sélectionnées.
  - Mesurez le temps de calcul
  - Arrivez-vous à retrouver le résultat attendu à l'aide de ces critères d'évaluation ?

3.2 Limites de la méthode

Choisissez quelques (2 ou 3) jeux de données pour lesquels il vous semble que la méthode de clustering agglométarif aura des difficultés pour identifier correctement les clusters.

- Appliquez la méthode de clustering agglomératif sur ces jeux de données pour confirmer vos choix.

Vous devez avoir identifié des limites de la méthode. Avez-vous besoin d'utiliser une autre métrique d'évaluation ?

4 Clustering `DBSCAN`

4.1 Intérêts de la méthode

Choisissez quelques (2 ou 3) jeux de données pour lesquels il vous semble que la méthode `DBSCAN` devrait identifier correctement les clusters.

- Appliquez la méthode `DBSCAN` en lui donnant des valeurs "au hasard" pour les paramètres `min-sample` et `eps` et en laissant la métrique de distance à sa valeur par défaut
- Appliquez itérativement la méthode précédente pour déterminer des bonnes valeurs pour les paramètres `min-sample` et `eps`
  - Mesurez le temps de calcul

4.2 Limites de la méthode

Choisissez quelques (2 ou 3) jeux de données pour lesquels il vous semble que la méthode `DBSCAN` aura des difficultés pour identifier correctement les clusters.

- Appliquez la méthode de clustering agglomératif sur ces jeux de données pour confirmer vos choix.

Vous devez avoir identifié des limites de la méthode. Avez-vous besoin d'utiliser une autre métrique d'évaluation ?

5 Clustering `HDBSCAN`

Le code `Python` de cette méthode est accessible ici [2]. Elle est connue pour être insensible à la variabilité de densité dans les données.
Reprenez les expérimentations effectuées avec `DBSCAN`. Comparez les résultats de ces deux méthodes. Arrivez-vous à retrouver les qualités et les limites de ces deux méthodes sur les jeux de données sélectionnés ? Y-at-il des diférences de performances (en temps de calcul) ?

6 Synthèse

Dans cette partie, vous allez appliquer les différentes méthodes de clustering étudiées précédemment sur de nouveaux jeux de données

- un dataset de données générées aléatoirement à récupérer sur la page du cours. Ces données sont en dimension 2.
- le dataset `iris` ou `balance-scale` (dimension 4). Ces données sont accessibles sur le même site que celui utilisé pour le TP mais dans la catégorie `real world`. Choisissez un seul de ces jeux de données.

L'objectif est de réaliser une analyse expérimentale comparative de différentes méthodes de clustering (qualité des solutions obtenues, performances des méthodes, ...). Les différents algorithmes testés fournissent-ils des solutions de clustering similaires ?

Pour les jeux de données "réelles", vous disposez du résultat d'une autre méthode de clustering, vous pouvez alors comparer les résultats que vous obtenez aux résultats fournis.

En complément des méthodes étudiées, vous pouvez utiliser d'autres méthodes (par exemple disponibles dans scikitlearn).

Pour lancer votre analyse expérimentale vous pouvez utiliser différents serveurs de calcul (normalement accessible à distance via le vpn insa) : srv-ens-calcul ou srv-gei-gpu1 et srv-gei-gpu2.

7 Evaluation

L'évaluation est à déposer sur moodle (un rapport par binôme). La date limite est : **xxx novembre 2021**.

Consignes pour le rapport :

- fichier pdf (maximum 15 pages)
- fournir dans le rapport un lien vers votre code (un dépot git, un jupyter notebook).
- Les différentes visualisations des jeux de données ou des résultats peuvent être jointes en annexe.
- Plan :
  - Partie 1 : Points forts et points faibles identifiés pour les différentes méthodes de clustering étudiées (5 à 6 pages)
  - Partie 2 : Analyse comparative sur les nouvelles données fournies (6 à 8 pages)
  - Conclusion

2. `https://github.com/scikit-learn-contrib/hdbscan`