M1 Informatique AMU – TP4 ISD — 2023/24 — Rapport. Date : 14 novembre 2023.
ZEMMOURI Yasmine, Groupe 3 TP# 5.

TP5 - Pratique du Clustering 2

Partie 1 : Classification non-supervisé.

Production d'un dataframe issu de l'initial :

La dataframe produit ne conserve que les attributs jugés importants dans TP précédant, entre autre, les attributs 'danceability', 'loudness', 'duration_ms', 'instrumentalness','valence' et 'acousticness'. Une classification sur ces colonnes avait montré que les résultats étaient similaires par rapport au DataSet initial.

➔ Comme nous connaissons le nombre de valeurs uniques de l'attribut 'popularity' (10 classes, de 0 à 9), il serait plus judicieux de d'opter pour le meme nombre de clusters., ce qui devrait conduire à des mesures de qualité des clusters plus précises.

Mais, si nous ne connaissions pas au préalable la classe, les clusters peuvent etre au nombre de 3 par exemple : pour exprimer le degré de non populairté.

✓ Néanmoins, au cours de ce TP, la décision a été prise d'utiliser 10 clusters.

Mesure d'évaluation d'un clustering :

➢ Si nous avions accès à l'impopularité de chaque chanson :

Mesure de similarité interne : l'indice de silhouette est une mesure populaire qui évalue la cohésion des points à l'intérieur des clusters par rapport à la séparation entre les clusters. Un score de silhouette élevé indique une bonne qualité de clustering.

Mesure de similarité externe : Nous pouvons utiliser des mesures telles que l'indice de Rand ou l'information mutuelle pour comparer les clusters obtenus avec les vraies classes.

➢ Si nous n'y avions pas accès :

Mesure de séparation entre les clusters : La distance entre les centroids des clusters peut également être utilisée pour évaluer la séparation entre les clusters.

✓ Les distances intra-clusters et extra-clusters peuvent également etre utilisées pour indiquer la qualité d'un clustering :

a- *Distance intra-cluster :* mesure la cohésion d'un cluster en calculant la somme des carrés des distances entre chaque point du cluster et le centroïde du cluster. Une faible inertie intra-cluster indique que les points à l'intérieur d'un cluster sont proches les uns des autres.

b- *Distance extra-cluster :* mesure la séparation entre les clusters en calculant la somme des carrés des distances entre les centroïdes des clusters. Une grande inertie extra-cluster suggère une bonne séparation entre les clusters.

Ainsi, si la distance intra-cluster est faible et la distance extra-cluster est élevée, cela indique une bonne cohésion interne des clusters et une bonne séparation entre eux, ce qui est souhaitable. Si la distance

intra-cluster est élevée, cela peut indiquer que les points à l'intérieur des clusters sont dispersés, ce qui peut signaler une mauvaise qualité du clustering.

Partie 2 : Mesures de qualité d'un regroupement.

Le code se trouve dans le fichier part1.py joint avec ce rapport.

Qualité d'un clustering k-moyennes :

Afin d'assurer des divisions (splits) différentes, il est possible d'atteindre cet objectif en définissant le paramètre random_state avec n'importe quelle valeur, à l'exception de 42.

En répétant 10fois l'apprentissage et la prédiction d'un regroupement en k clusters en utilisant les k-moyennes, on obtient les résultats suivants :

```
Moyenne des scores sur le jeu d'apprentissage (DB, Silhouette, Rand Index): 1.2461520040540999 0.22170080634387204 -0.001393024
7712197026
Moyenne des scores sur le jeu de test (DB, Silhouette, Rand Index) 1.2703510597708707 0.21391546746997533 -0.000580389265424514
2
```

i. Le score moyen sur le jeu d'apprentissage (1.246) et sur le jeu de test (1.270) indique la compacité et la séparation des clusters. Un score plus bas est préférable, ce qui suggère une meilleure qualité du clustering. Dans ce cas, les scores sont relativement proches, indiquant une cohérence du modèle entre les ensembles d'apprentissage et de test.

ii. La silhouette mesure à quel point les objets d'un cluster sont similaires entre eux par rapport aux clusters voisins. Un score élevé (plus proche de 1) indique une bonne séparation entre les clusters. Les scores moyens sur le jeu d'apprentissage (0.2217) et le jeu de test (0.2139) indiquent une séparation relativement correcte des clusters, bien que les valeurs soient assez modestes.

iii. Le Rand Index mesure la similarité entre les paires d'échantillons dans les vrais clusters et les clusters prédits. Un score de 0 indique une correspondance aléatoire, tandis qu'un score de 1 indique une correspondance parfaite. Les scores moyens sur le jeu d'apprentissage (-0.0014) et le jeu de test (-0.0006) suggèrent une correspondance faible, voire aléatoire, entre les vrais clusters et les clusters prédits.

En conclusion, les résultats indiquent une certaine cohérence entre les performances sur le jeu d'apprentissage et le jeu de test, mais les scores généraux semblent relativement modestes. De plus, le Rand Index proche de zéro indique des difficultés dans la correspondance entre les clusters réels et prédits.

Qualité d'un clustering mélange de gaussiennes :

En appliquant les memes étapes que la section précédante avec l'algorithme des mélanges gaussiens, on obtient les résultats suivants :

```
Moyenne des scores sur le jeu d'apprentissage (DB, Silhouette, Rand Index): 3.1830461894546 0.00876658542350062 -0.006502103071
889982
Moyenne des scores sur le jeu de test (DB, Silhouette, Rand Index) 3.649678038402873 -0.007479239670740602 -0.00901930065656881
6
```

i. Les scores moyens sur le jeu d'apprentissage (3.183) et le jeu de test (3.649) indiquent des valeurs plus élevées par rapport au clustering KMeans. Un score plus élevé dans le cas de DB suggère une dispersion et une séparation moins claires entre les clusters, montrant une qualité de clustering inférieure.

ii. Les scores moyens sur le jeu d'apprentissage (0.0088) et le jeu de test (-0.0075) sont relativement bas. Un score négatif pour la silhouette suggère une mauvaise séparation entre les clusters, qui montre une faible qualité de clustering et une certaine superposition entre les clusters.

iii. Les scores moyens sur le jeu d'apprentissage (-0.0065) et le jeu de test (-0.0090) sont également négatifs, suggérant une faible concordance entre les clusters réels et les clusters prédits. Ce qui indique des difficultés dans la capacité du modèle à reproduire la structure des données.

En comparaison avec le clustering KMeans, les mélanges gaussiens semblent montrer des performances relativement inférieures, avec des scores indiquant une séparation moins claire entre les clusters et une correspondance moindre avec les vrais clusters.

Comparaison de deux méthodes :

Le code se trouve dans le fichier part2.py joint avec ce rapport.

En entrainant les deux méthodes sur toutes nos données, on obtient un résultat de 0.213847 pour la mesure Rand Index.

➔ Une valeur positive du Rand Index indique que les clusters générés par K-Means et GMM partagent certaines similarités. Cependant, le fait que le Rand Index ne soit pas très proche de 1 suggère qu'il existe des différences importantes entre les clusters produits par les deux algorithmes.

On peut dire qu'il existe une concordance modérée entre les deux méthodes de regroupement. Cependant, cette valeur suggère également des différences substantielles entre les clusters générés par K-Means et ceux produits par GMM. Ces différences peuvent être attribuées aux hypothèses distinctes sur la structure des données entre les deux algorithmes, notamment la flexibilité de GMM pour modéliser des clusters de formes variées.

✓ On en conclue, d'après les résultats obtenus dans les sections précédentes, que la méthode des k-moyennes est plus performante pour ce jeu de données.

Pour savoir dans combien de clusters les exemples de popularité se trouvent, on peut construite la matrice suivante :

popularity	0	1	2	3	4	5	6	7	8	9
KMeans_Cluster										
0	289	91	29	15	4	4	2	14	15	1
1	557	161	44	29	14	4	6	29	31	1
2	232	29	5	0	2	0	0	4	3	0
3	4	2	0	0	0	0	0	0	0	0
4	389	158	51	18	13	5	2	21	25	1
5	89	35	21	2	3	4	1	1	6	0
6	485	130	40	16	9	5	2	24	37	1
7	238	66	13	6	5	3	1	4	2	0
8	163	111	24	5	5	4	1	2	4	0
9	130	53	13	2	1	1	1	0	0	0

Par exemple, les instances de popularité 3 se trouvent dans les clusters 0 (15), 1 (29), ‘4 (18), 5 (2), 6 (16), 7 (6), 8 (5), et 9 (2).

On remarque que aucune instance de popularité 3 n’a été classé dans un cluster d’étiquette 3.

➔ Rand index de ce regroupement par rapport à la vérité terrain est de -0.0018.

Ce résultat est signe d’une correspondance très faible entre les clusters prédits et les véritables classes. Un Rand Index négatif suggère que les clusters générés ne concordent pas du tout avec les classes réelles. En conséquence, il semble y avoir une absence de relation significative entre les attributs décrivant les chansons (utilisés pour le regroupement) et leur véritable popularité.

Ces résultats pourraient indiquer que les attributs choisis ne capturent pas de manière significative les nuances de popularité, ou que d'autres facteurs non pris en compte par les attributs utilisés influent davantage sur la popularité des chansons.