Mesures de qualité des partitions en classification automatique

Ricco RAKOTOMALALA
Université Lumière Lyon 2

PLAN

Evaluation d'une partition issue d'un processus de clustering

Comment mesurer la performance d'un algorithme de clustering

Par extension, comment comparer les résultats des algorithmes

Classification automatique

Typologie, apprentissage non-supervisé, clustering

Objectif : identifier des groupes d'observations ayant des caractéristiques similaires (ex. comportement d'achats de clients, propriétés de véhicules, etc.)

On veut que :

(1) Les individus dans un même groupe se ressemblent le plus possible

(2) Les individus dans des groupes différents se démarquent le plus possible

Pourquoi ?

→ Identifier des structures sous-jacentes dans les données

→ Résumer des comportements

→ Affecter de nouveaux individus à des catégories

→ Identifier les cas totalement atypiques

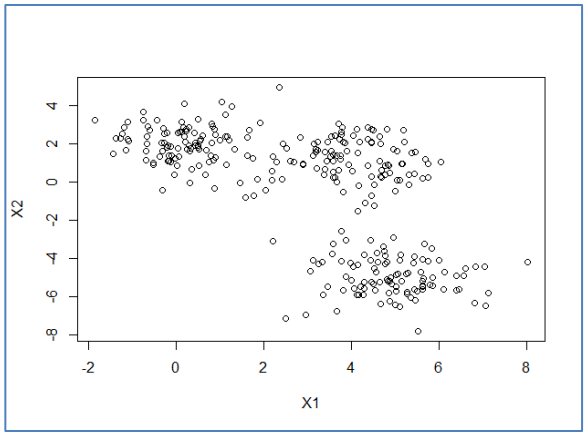
Modele	puissance	cylindree	vitesse	longueur	hauteur	poids	CO2
TWINGO	60	1149	151	344	143	840	143
PANDA	54	1108	150	354	154	860	135
YARIS	65	998	155	364	150	880	134
CITRONC2	61	1124	158	367	147	932	141
P1007	75	1360	165	374	161	1181	153
MODUS	113	1598	188	380	159	1170	163
CLIO	100	1461	185	382	142	980	113
CORSA	70	1248	165	384	144	1035	127
MERC_A	140	1991	201	384	160	1340	141
FIESTA	68	1399	164	392	144	1138	117
MUSA	100	1910	179	399	169	1275	146
AUDIA3	102	1595	185	421	143	1205	168
GOLF	75	1968	163	421	149	1217	143
CITRONC4	138	1997	207	426	146	1381	142
VECTRA	150	1910	217	460	146	1428	159
AVENSIS	115	1995	195	463	148	1400	155
P407	136	1997	212	468	145	1415	194
PASSAT	150	1781	221	471	147	1360	197
MONDEO	145	1999	215	474	143	1378	189
VELSATIS	150	2188	200	486	158	1735	188

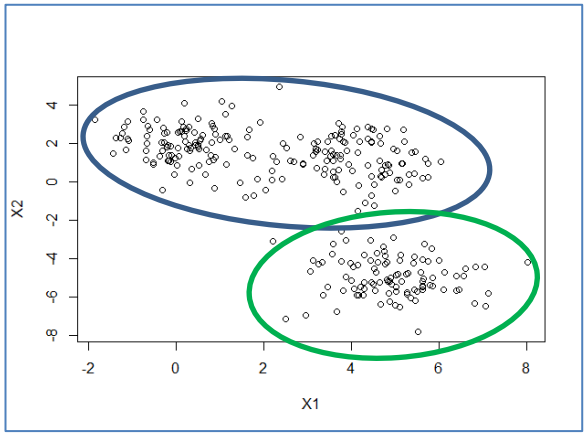
Exemple : Clustering en 3 groupes de véhicules à partir de leurs propriétés (motorisation, carrosserie, etc.)

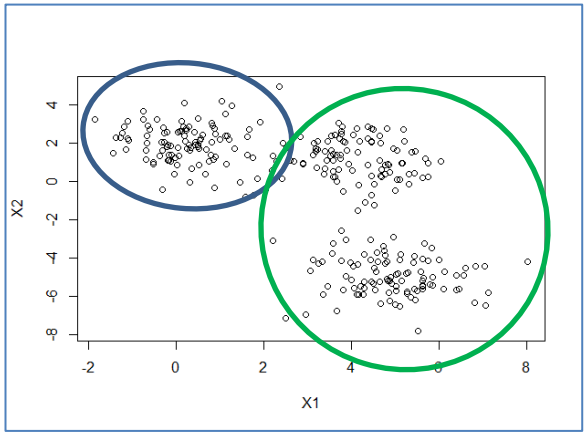
Evaluation des performances (1)

Quels critères numériques pour décider qu'une partition est « intéressante » ?

Exemple d'un nuage de points dans le plan (X1, X2)







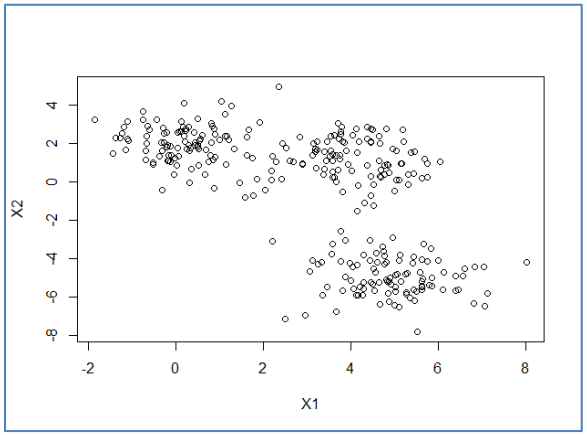
Laquelle des 2 partitions est meilleure ?

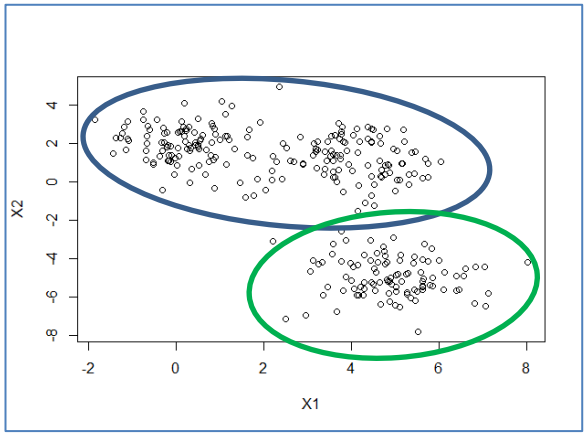
Sachant qu'à la différence de l'apprentissage supervisé, il n'y a pas de « vraies » classes d'appartenance qui feraient référence

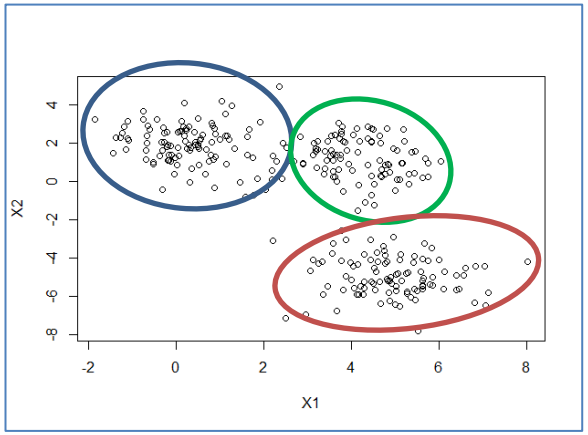
Evaluation des performances (2)

Comment comparer des solutions à nombre de classes différentes ?

Exemple d'un nuage de points dans le plan (X1, X2)







Parce qu'en plus de détourer les clusters, il faut aussi identifier leur nombre

Quelles mesures pour comparer des solutions avec un nombre de classes différentes ?

Mesures externes

Lorsque les « vraies » classes d'appartenance sont disponibles

Principe des mesures externes

Groupes issus du clustering vs. « vrais » groupes d'appartenance

On se positionne dans un cadre particulier où les véritables classes d'appartenance sont connues. Mesures = cohérence entre classes apprises et classes originelles.

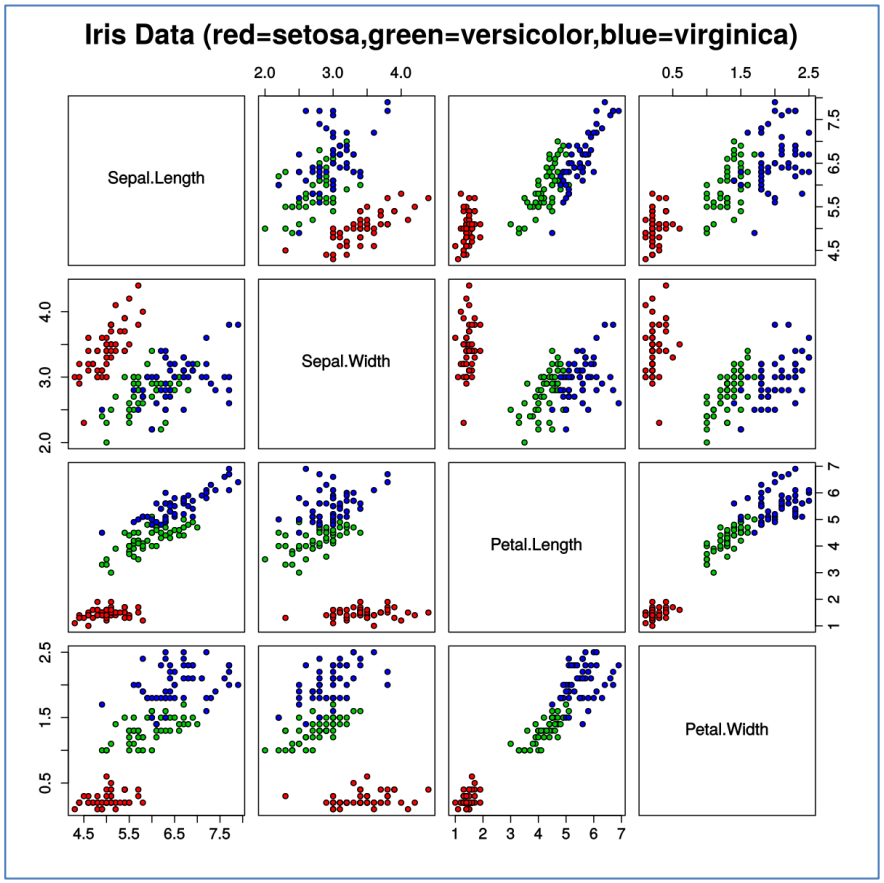
Cette situation n'arrive pas dans les études réelles, pourquoi s'embêter avec un algorithme de clustering sinon, on ferait mieux de passer à des techniques supervisées

Cette approche est mise en avant dans un cadre expérimental (en recherche souvent) où l'on souhaite comparer les mérites respectifs de différents algorithmes de clustering (ex. K-Means vs. CAH). Il faut bien un juge indiscutable.

On travaille à partir d'un tableau croisé, une « sorte de matrice de confusion », opposant les classes réelles et prédites. Sauf que les modalités sont dans le désordre (puisque les clusters sont numérotés arbitrairement)

Exemple du fichier IRIS

Dataset emblématique du machine learning



IRIS flower dataset : n = 150 fleurs, caractérisé par p = 4 variables, réparties en 3 espèces (Species = {setosa, versicolor, virginica}).

L'idée est de travailler « en aveugle » à partir des p = 4 descripteurs, identifier le nombre de classes, les circonscrire, puis de comparer les résultats avec les classes originelles définies par Species.

On remarque que « setosa » est facilement identifiable, les deux autres un peu moins.

Tableau croisé

Classes réelles (R) vs. classes attribuées (C) par le clustering

« K » n'est pas forcément égal à « L »

	c_1	…	c_k	…	c_K	Σ
r_1						
…						
r_l			n_{lk}			$n_{l.}$
…						
r_L						
Σ			$n_{.k}$			n

L'ordre des colonnes est aléatoire, pas forcément en connexion avec celui des lignes (contrairement à la matrice de confusion en supervisé)

(S1) K-Means en K = 2 classes

	c_1	c_2	Σ
setosa	50	0	50
versicolor	0	50	50
virginica	0	50	50
Σ	50	100	150

(S2) K-Means en K = 3 classes

	c_1	c_2	c_3	Σ
setosa	0	50	0	50
versicolor	39	0	11	50
virginica	14	0	36	50
Σ	53	50	47	150

(S3) Partition aléatoire en K = 3 classes

	c_1	c_2	c_3	Σ
setosa	20	17	13	50
versicolor	19	17	14	50
virginica	16	10	24	50
Σ	55	44	51	150

V de Cramer (1)

Mesure normalisée issue du KHI-2 d'indépendance

KHI-2 : Tester le lien statistique entre 2 variables qualitatives c.-à-d. la connaissance de l'une fournit des informations sur l'autre ? En pratique, mesurer l'écart à la situation d'indépendance. Permuter les lignes ou les colonnes du tableau ne change pas la mesure, transposer le tableau non-plus.

Effectifs théoriques sous indépendance

$$e_{lk} = \frac{n_{l.} \times n_{.k}}{n}$$

Disparités entre effectifs observés et théoriques. Statistique du KHI-2

$$\chi^2 = \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{(n_{lk} - e_{lk})^2}{e_{lk}}$$

V de Cramer, mesure normalisée du KHI-2, comprise entre 0 (indépendance) et 1 (déterminisme)

V de Cramer

$$v = \sqrt{\frac{\chi^2}{n \times \min(L-1, K-1)}}$$

V de Cramer (2)

Fichier IRIS

(S1) K-Means en K = 2 classes

	C_1	C_2	Σ
setosa	50	0	50
versicolor	0	50	50
virginica	0	50	50
Σ	50	100	150

$$v = 1.0$$

Pourtant : certes, connaître « R » permet de déterminer à coup sûr « C », mais l'inverse n'est pas vrai.

(S2) K-Means en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	0	50	0	50
versicolor	39	0	11	50
virginica	14	0	36	50
Σ	53	50	47	150

$$v = 0.79$$

Moins bien évaluée que la situation (S1), alors que le nombre de classes est « correct ».

(S3) Partition aléatoire en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	20	17	13	50
versicolor	19	17	14	50
virginica	16	10	24	50
Σ	55	44	51	150

$$v = 0.15$$

Partition aléatoire ne veut pas dire à coup sûr valeur minimale ($v = 0$) de la mesure

Information mutuelle

Mesure de liaison pour variables qualitatives

Toujours sur la même idée de déterminer le degré de liaison. On note la similitude avec le KHI-2 (écart à l'indépendance = produit des marges). On peut en dériver un test statistique également.

$$MI(R,C) = \sum_{l=1}^{L}\sum_{k=1}^{K} \frac{n_{lk}}{n} \log \frac{n \times n_{lk}}{n_{l.} \times n_{.k}}$$

$MI(R,C) \geq 0$; MI(R,C) = 0 seulement en cas d'indépendance.

Pour le fichier IRIS

$$MI(S_1) = 0.636$$

$$MI(S_2) = 0.724$$

$$MI(S_3) = 0.023$$

La partition aléatoire (S_3) est proche de 0. Les deux autres solutions se démarquent. S_2 semble être la meilleure, mais la mesure est connue pour favoriser les nombres de clusters plus élevés.

Information mutuelle normalisée

Mesure de liaison pour variables qualitatives

Avec la normalisation, la mesure présente l'avantage de varier entre 0 (indépendance) et 1 (déterminisme)

$$NMI(R,C) = \frac{MI(R,C)}{\frac{1}{2}[H(R) + H(C)]}$$

Où $$H(R) = -\sum_l \frac{n_{l.}}{n} \log \frac{n_{l.}}{n}$$

$$H(C) = -\sum_l \frac{n_{.k}}{n} \log \frac{n_{.k}}{n}$$

La structure de la formule n'est pas sans évoquer celle de la corrélation.

Pour le fichier IRIS

$$NMI(S_1) = 0.733$$

$$NMI(S_2) = 0.659$$

$$NMI(S_3) = 0.021$$

La partition aléatoire (S_3) toujours proche de 0. Les deux autres solutions se démarquent, la normalisation a modifié leur ordonnancement (mais les différences restent minimes).

Information mutuelle ajustée

Corrigée par la « chance » que deux individus atterrissent par hasard dans le même cluster

Même dans une partition réalisée aléatoirement, la probabilité que deux individus de la même classe appartiennent au même cluster n'est pas nulle. Il faut corriger la mesure avec la valeur de l'espérance de l'information mutuelle en cas de partition aléatoire E[MI(R,C)]. Cf. formule (réf. 1 et 2)

$$AMI(R,C) = \frac{MI(R,C) - E[MI(R,C)]}{\frac{1}{2}[H(R) + H(C)] - E[MI(R,C)]}$$

Attention : la mesure peut prendre des valeurs négatives dans certains cas extrêmes (proches de l'indépendance)

Pour le fichier IRIS

$$NMI(S_1) = 0.731$$

$$NMI(S_2) = 0.655$$

$$NMI(S_3) = 0.009$$

La partition aléatoire (S_3) toujours proche de 0. Les corrections sont minimes dans nos configurations.

Indice de Rand (1)

Cohérences et incohérences des classes d'appartenance réelles et attribuées

Construire un tableau intermédiaire binaire qui met en évidence les cohérences et incohérences des classes d'appartenance réelles (R) et affectées (C) par le clustering

	=	≠
=	a	b
≠	c	d

a : nombre de paires d'individus appartenant à la même classe dans R, et au même cluster dans C.

b : appartenant à la même classe dans R, mais dans des classes différentes dans C

c : classes différentes dans R, clusters identiques dans C

d : classes différentes dans R, différentes dans C

$$RI = \frac{a+d}{a+b+c+d}$$

Mais, astuce suprême, il n'est pas nécessaire de calculer (a, b, c, d) en considérant explicitement chaque paire d'individus (calculs prohibitifs !), on peut les obtenir à partir du premier tableau de contingence croisant R et C.

$$a = \frac{1}{2}\sum_l \sum_k n_{lk}(n_{lk} - 1)$$

$$b = \frac{1}{2}\sum_l \sum_k n_{lk}(n_{l.} - n_{kl})$$

$$c = \frac{1}{2}\sum_l \sum_k n_{lk}(n_{.k} - n_{kl})$$

$$d = \frac{1}{2}\sum_l \sum_k n_{lk}(n - n_{l.} - n_{.k} + n_{kl})$$

Indice de Rand (2)

Exemple : partition S2

(S2) K-Means en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	0	50	0	50
versicolor	39	0	11	50
virginica	14	0	36	50
Σ	53	50	47	150

	=	≠
=	2742	933
≠	942	6558

$$RI = \frac{2742 + 6558}{2742 + 933 + 942 + 6558} = 0.832$$

L'indice de Rand s'interprète comme la proportion des paires d'individus qui sont dans la même situation dans les 2 partitions (soit assignées dans la même classe dans les 2 cas, soit assignées dans des classes différentes). Il varie entre 0 et 1.

Et bien sûr, $(a + b + c + d) = C_n^2 = \binom{n}{2} = \frac{n!}{2!\,(n-2)!} = \frac{n(n-1)}{2} = 1175$

Correspond au nombre total de paires possibles…

Indice de Rand (3)

Comparaisons des partitions

(S1) K-Means en K = 2 classes

	C_1	C_2	Σ
setosa	50	0	50
versicolor	0	50	50
virginica	0	50	50
Σ	50	100	150

$$RI = 0.776$$

Setosa vs. C1 fait un sans faute, c'est plus mitigé pour les autres.

(S2) K-Means en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	0	50	0	50
versicolor	39	0	11	50
virginica	14	0	36	50
Σ	53	50	47	150

$$RI = 0.832$$

Mieux évaluée que la situation (S1), à raison ? À tort ?

(S3) Partition aléatoire en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	20	17	13	50
versicolor	19	17	14	50
virginica	16	10	24	50
Σ	55	44	51	150

$$RI = 0.569$$

Partition aléatoire, et pourtant l'indice est loin de la valeur 0

Indice de Rand ajusté

Ajusté toujours par la « chance » que les paires soient dans des groupes identiques

Même si l'affectation est aléatoire, la probabilité qu'une paire d'individus soit assigné à un même groupe dans les 2 partitions n'est pas nulle. Il faut corriger l'indice pour supprimer ce biais.

$$ARI = \frac{\sum_{lk}\binom{n_{lk}}{2} - \left[\sum_l \binom{n_{l.}}{2}\sum_k \binom{n_{.k}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_l \binom{n_{l.}}{2} + \sum_k \binom{n_{.k}}{2}\right] - \left[\sum_l \binom{n_{l.}}{2}\sum_k \binom{n_{.k}}{2}\right]/\binom{n}{2}}$$

Que l'on peut aussi lire : $ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$

E(RI) est l'espérance de l'indice en cas de partition au hasard.

⇨ L'indice ajusté peut prendre de valeurs négatives en cas de discordances fortes !!!

Pour le fichier IRIS

$ARI(S_1) = 0.568$

$NMI(S_2) = 0.620$

$NMI(S_3) = 0.0098$

La partition aléatoire (S_3) est maintenant proche de 0. Forte correction par rapport à l'indice non-ajusté.

Score d'homogénéité (1)

Basé sur les calculs d'entropie

Basé sur le gain d'information, la connaissance des clusters calculés (C) donne-t-elle de l'information sur l'appartenance originelle aux classes (R) ?

$$h = \frac{H(R) - H(R/C)}{H(R)} = 1 - \frac{H(R/C)}{H(R)}$$

Où

$$H(R) = -\sum_l \frac{n_{l.}}{n} \log \frac{n_{l.}}{n}$$

Entropie marginale

$$H(R/C) = -\sum_k \frac{n_{.k}}{n} \sum_l \frac{n_{lk}}{n_{.k}} \log \frac{n_{lk}}{n_{.k}}$$

Entropie conditionnelle

Interprétation : valeur de l'indice élevée si à chaque cluster calculé (c_k) correspond une et une seule classe d'appartenance initiale (r_l)

Indice varie entre 0 et 1.

Avec un peu d'imagination, on peut faire le parallèle avec variante inter = (variance totale – variance intra), normalisée par la variance totale.

Score d'homogénéité (2)

Fichier IRIS

	C_1	C_2	Σ
setosa	50	0	50
versicolor	0	50	50
virginica	0	50	50
Σ	50	100	150

$$h = 0.579$$

C1 est très bien, mais mesure plombée par C2. Pénalisé aussi ici parce que K < L, forcément colonnes non « pures ».

(S2) K-Means en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	0	50	0	50
versicolor	39	0	11	50
virginica	14	0	36	50
Σ	53	50	47	150

$$h = 0.659$$

C2 nickel (correspond à setosa), mitigé pour C1 et C3.

(S3) Partition aléatoire en K = 3 classes

	C_1	C_2	C_3	Σ
setosa	20	17	13	50
versicolor	19	17	14	50
virginica	16	10	24	50
Σ	55	44	51	150

$$h = 0.021$$

Ca va pas, et la mesure le dit bien.

Score de complétude

Basé sur les calculs d'entropie

Transposée de l'homogénéité c.-à-d. est-ce qu'à une classe réelle (r_l) correspond un et un seul cluster (c_k)

$$cs = \frac{H(C) - H(C/R)}{H(C)} = 1 - \frac{H(C/R)}{H(C)}$$

Où $$H(C) = -\sum_k \frac{n_{.k}}{n} \log \frac{n_{.k}}{n}$$

Dans quelle mesure on peut associer un cluster à chaque classe réellement observée

Pour le fichier IRIS

$$cs(S_1) = 1.0$$

$$cs(S_2) = 0.659$$

$$cs(S_3) = 0.0213$$

La partition (S_1) est parfaite dans le sens de la lecture en ligne du tableau de contingence : à « setosa », on associe C1 ; à « versicolor », sans aucune doute C2 ; à « virginica », sans aucun doute C2.

V-measure

Moyenne harmonique entre homogénéité et complétude

Faire un arbitrage entre « homogénéité » et « complétude » à travers une moyenne harmonique.

$$vm_\beta = \frac{(1+\beta) \times h \times cs}{\beta \times h + cs}$$

β = 1, même importance aux deux critères

β > 1, plus d'importance à « cs »

β < 1, plus d'importance à « h »

Remarque : ce critère n'est pas sans rappeler le F-Score qui permet d'arbitrer entre précision et rappel en apprentissage supervisé.

Pour le fichier IRIS, avec β =1

$$vm(S_1) = 0.733$$

$$vm(S_2) = 0.659$$

$$vm(S_3) = 0.0212$$

Pour la partition (S_1), complétude max., mais faible homogénéité (à cause de C2), la mesure combinée traduit cette situation.

Mesures internes

Mesurer la qualité intrinsèque de la partition avec 2 éléments clés :

(1) Compacité : Les individus dans un même groupe se ressemblent le plus possible

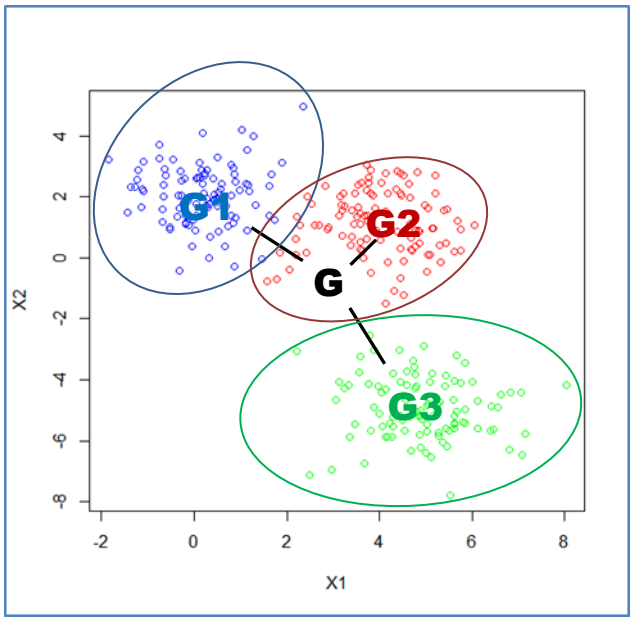
(2) Séparabilité : Les individus dans des groupes différents se démarquent le plus possible

⇨ Aux fins de comparaison de partitions bien sûr, mais aussi avec pour objectif sous-jacent de disposer d'outils pour identifier le nombre de classes ?

Inertie (1)

Mesure de dispersion multidimensionnelle

Rôle crucial des centres de classes



Relation fondamentale (Théorème d'Huygens)

$$\text{Inertie totale} = \text{Inertie inter - classes} + \text{Inertie intra - classe}$$

$$T = B + W$$

$$\sum_{i=1}^{n} d^2(i,G) = \sum_{k=1}^{K} n_k d^2(G_k,G) + \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i,G_k)$$

Dispersion des barycentres conditionnels autour du barycentre global. Indicateur de séparabilité des classes.

Dispersion à l'intérieur de chaque groupe. Indicateur de compacité des classes.

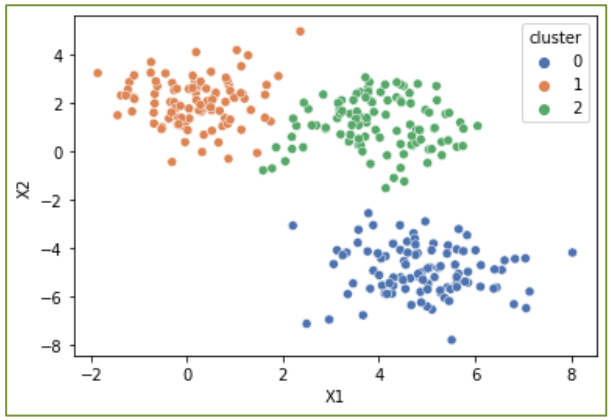
d() est une mesure de distance caractérisant les proximités entre les individus, distance euclidienne ou euclidienne pondérée par l'inverse de la variance

Certains algorithmes (K-Means, CAH critère de Ward) maximisent explicitement B ou, c'est équivalent, minimisent W

Inertie (2)

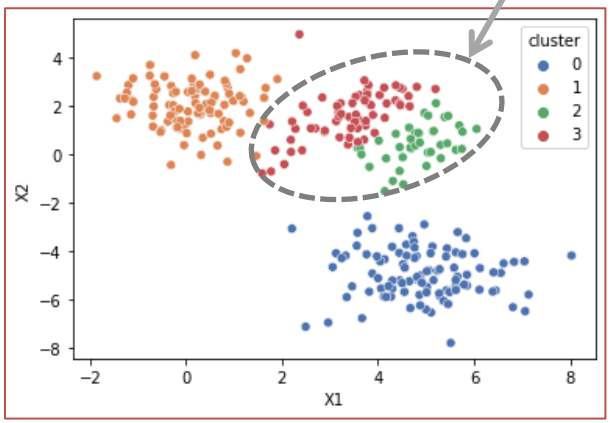
K (nombre de classes) augmente ➔ W diminue mécaniquement

W permet de comparer des partitions à nombre de classes égales, en revanche elle diminue mécaniquement quand K augmente (laissant à penser que la solution est meilleure)



W = 578.59

Justifié ?

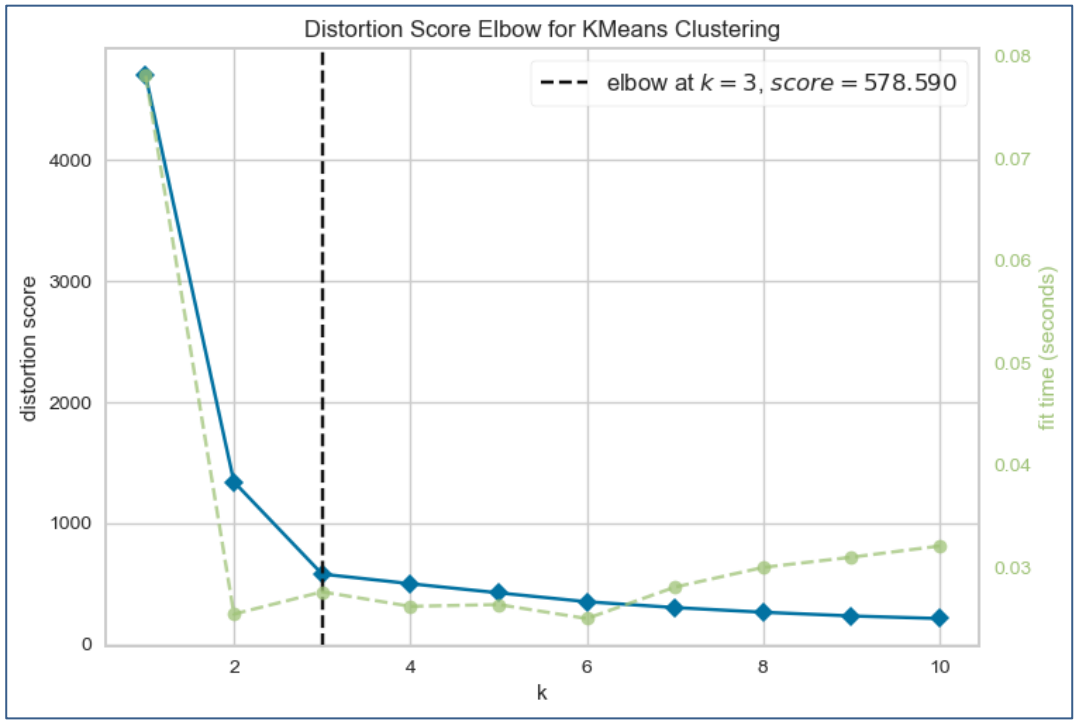


W = 499.23

La question que l'on peut se poser est : la décroissance est-elle « significative » lorsque l'on rajoute une classe supplémentaire ?

Inertie (3)

Stratégie pour l'identification du nombre de classes



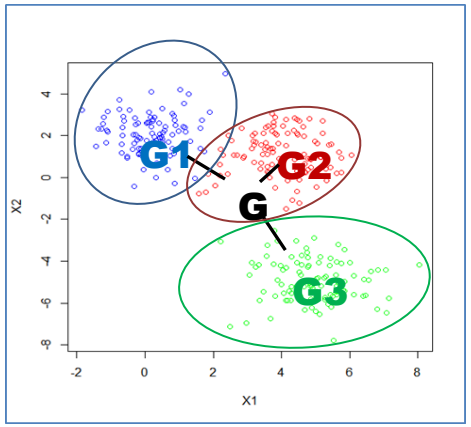
En bleu, courbe de décroissance de W en fonction de K

En vert le temps de calcul, ne nous intéresse pas ici.

Recherche du « coude ». Le passage de « K = 3 » à « K = 4 » n'amène pas une amélioration notable de la qualité de la partition (au sens de W). On décide « K = 3 ».

Indice de Calinski – Harabasz

Rapport entre inerties inter-groupes et intra-groupes

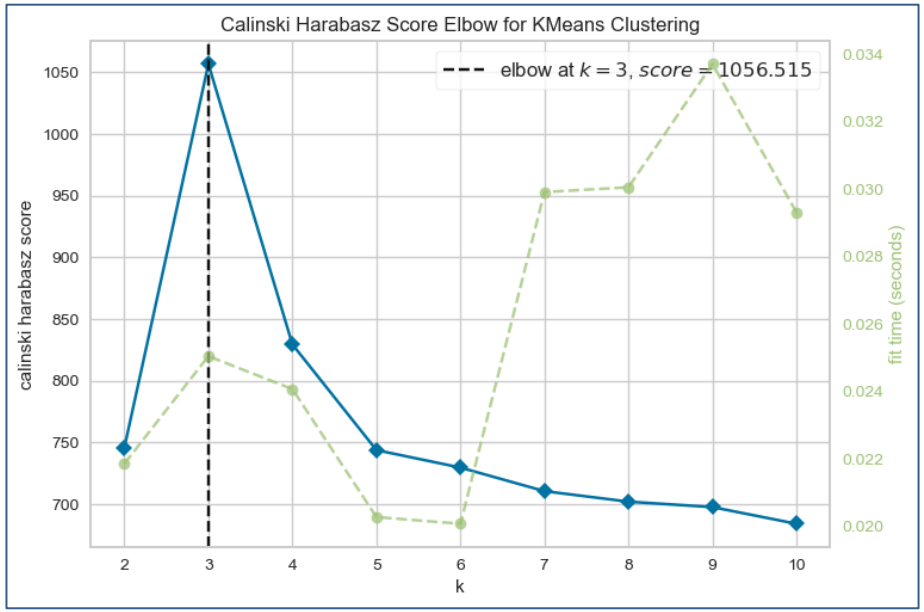


Toujours basé sur les inerties, mais la formule introduit explicitement le nombre de clusters (K). C'est une forme de correction vs. la propension à favoriser le nombre de classes élevé.

$$CH = \frac{(n-K) \times B}{(K-1) \times W} \qquad (CH \geq 0)$$

Plus CH est grand, meilleure est la partition

De fait, pour identifier le « bon » nombre de clusters (K), il suffit de choisir la configuration qui maximise le critère CH (ici K* = 3)



Critère silhouette (1)

Degré d'appartenance à sa classe d'un individu

Rousseeuw (1987) propose une mesure d'évaluation des partitions non-dépendante du nombre de classes : le score silhouette.

$$a(i) = \frac{1}{n_a - 1} \sum_{\substack{i'=1 \\ i' \neq i}}^{n_a} d(i,i')$$

Moyenne des distances du point i avec l'ensemble des points de sa classe d'appartenance C_a dont l'effectif est n_a.

$$d(i,C_k) = \frac{1}{n_k} \sum_{i'=1}^{n_k} d(i,i')$$

Moyenne des distances du point i avec l'ensemble des points d'une classe C_k – autre que C_a – dont l'effectif est n_k.

$$b(i) = \min_{k \neq a} d(i,C_k)$$

Distance à la classe la plus proche au sens de $d(i,C_k)$

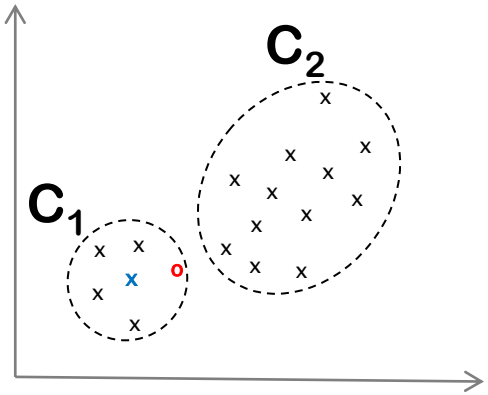
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Indique le degré d'appartenance à sa classe de l'individu en confrontant la distance moyenne à ses congénères avec la distance moyenne à la classe la plus voisine. s(i) est indépendant de K – nombre de classes – parce qu'on ne considère que la distance au voisin le plus proche !

s(i) → 1 : le point est bien positionné dans sa classe
s(i) ≈ 0 : le point est aussi proche des autres que de ses congénères
s(i) → -1 : le point est plus proche des autres que de ses congénères

Critère silhouette (2)

Indicateurs pour l'évaluation des classes et des partitions



s(x) > s(o): (1) parce que « x » est placé en position centrale au sein de C_1 ; (2) parce que « o » se rapproche de la classe C_2.

$$\bar{s}_k = \frac{1}{n_k} \sum_{i \in C_k} s(i)$$

Pour le cluster C_k : la silhouette moyenne caractérise à la fois la compacité du groupe et de son écartement (séparabilité) par rapport aux autres classes.

$$S_K = \frac{1}{n} \sum_{k=1}^{K} n_k \times \bar{s}_k$$

Pour caractériser la qualité globale de la partition en K classes. De manière empirique :

S ∈ [0,71 ; 1] : une structuration forte a été découverte

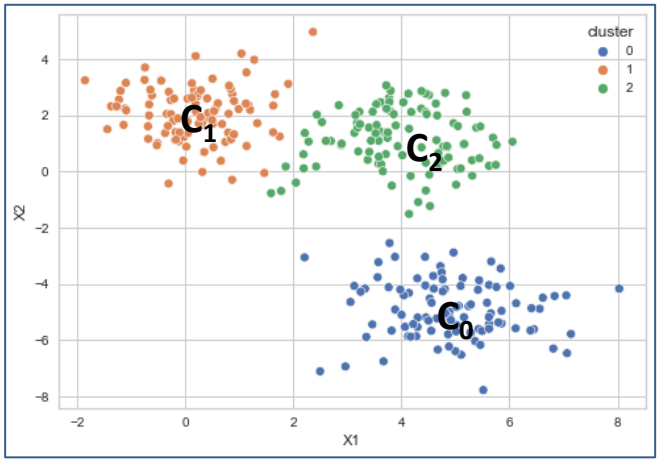
S ∈ [0,51 ; 0.70] : une structuration raisonnable existe

S ∈ [0,26 ; 0.50] : la structuration est faible, sujette à caution

S ∈ [0 ; 0.25] : pas de structuration des données

Critère silhouette (3)

Graphique silhouette



$\bar{s}_0 = 0.70$

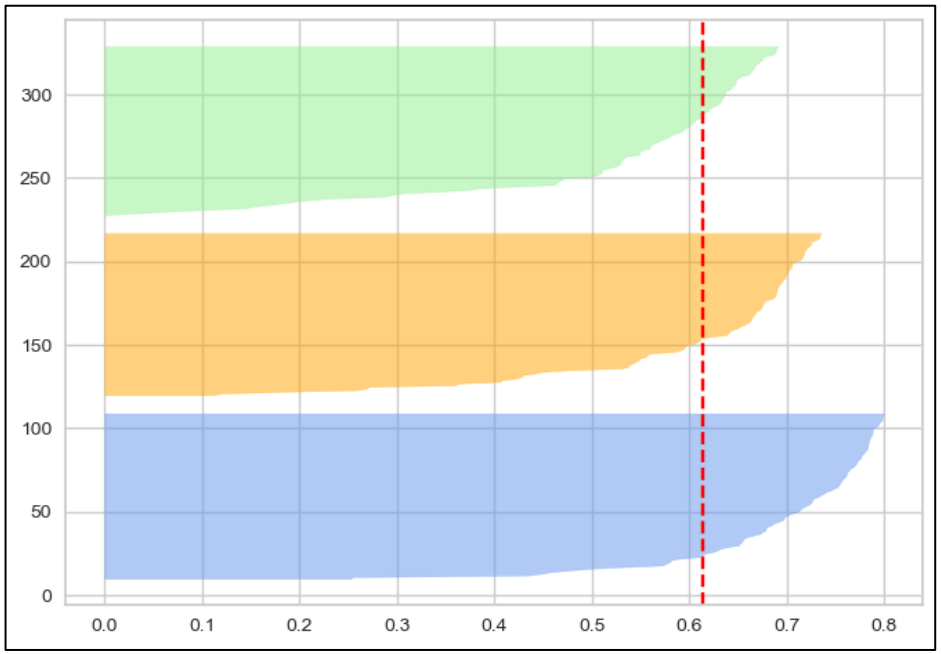
$\bar{s}_1 = 0.60$

$\bar{s}_2 = 0.53$

La classe C_0 (bleue) est celle qui se démarque le plus des autres, la valeur moyenne de l'indice silhouette est la plus élevée.

***Qualité globale** de la partition en K = 3 classes.*

$S_{K=3} = 0.61$



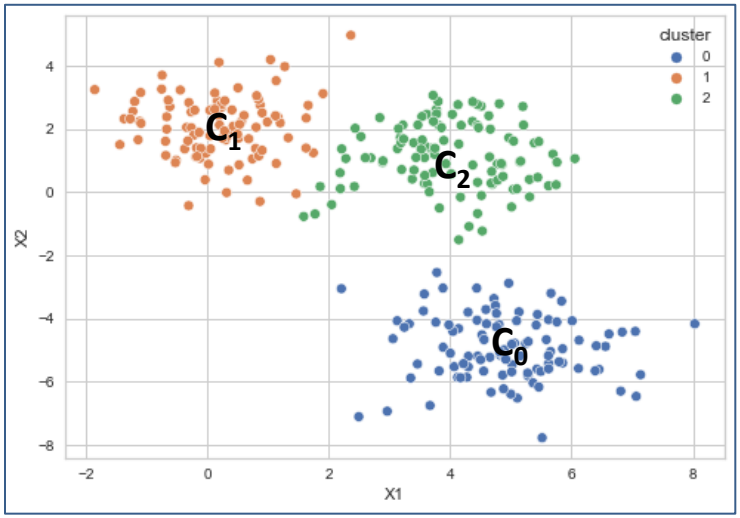
Les packages de calcul les plus connus proposent la représentation graphique connue sous l'appellation « Graphique silhouette » (silhouette plot).

Les individus de la classe C_0 (bleue) présentent une valeur de l'indice silhouette globalement plus élevé que l'indice moyen de l'ensemble de l'échantillon.

La perception visuelle est en accord avec les $\bar{s}_k$

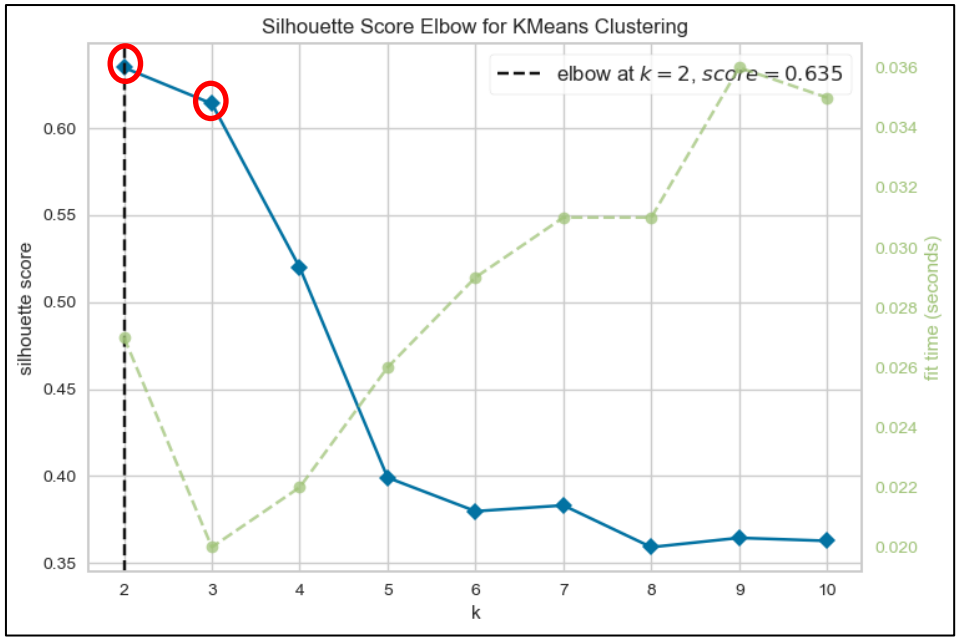
Critère silhouette (4)

Outil pour le choix du nombre de partitions



Le critère silhouette étant indépendant du nombre de classes, Il suffit de choisir la valeur K qui le maximise.

Tester les différentes valeurs de K et identifier les meilleurs configurations (partitions en K classes). Ici K= 2 ($S_{K=2}$ = 0.63) et ($S_{K=3}$ = 0.61) sont en concurrence. L'indicateur préfère la solution en K = 2 classes. Est-ce vraiment étonnant finalement (les deux solutions se tiennent…) ?



Indice de Davies-Bouldin (1)

Compacité vs. séparabilité

s_k Indicateur de compacité = moyenne des distances des points du cluster C_k avec leur barycentre G_k

d_{kj} Indicateur de séparabilité = distance entre les barycentres des clusters C_k et C_j

Opposition compacité vs. séparabilité entre paires de clusters (≥ 0 et symétrique). Plus faible est la valeur, plus les clusters sont discernables.

$$R_{kj} = \frac{s_k + s_j}{d_{kj}}$$

Indice de Davies-Bouldin

$$DB = \frac{1}{K}\sum_{k=1}^{K} \max_{k \neq j} R_{kj}$$

Concrètement, R est une matrice symétrique de taille (K, K), on cherche le maximum par ligne (en ignorant la diagonale), et on calcule la moyenne de ces valeurs.

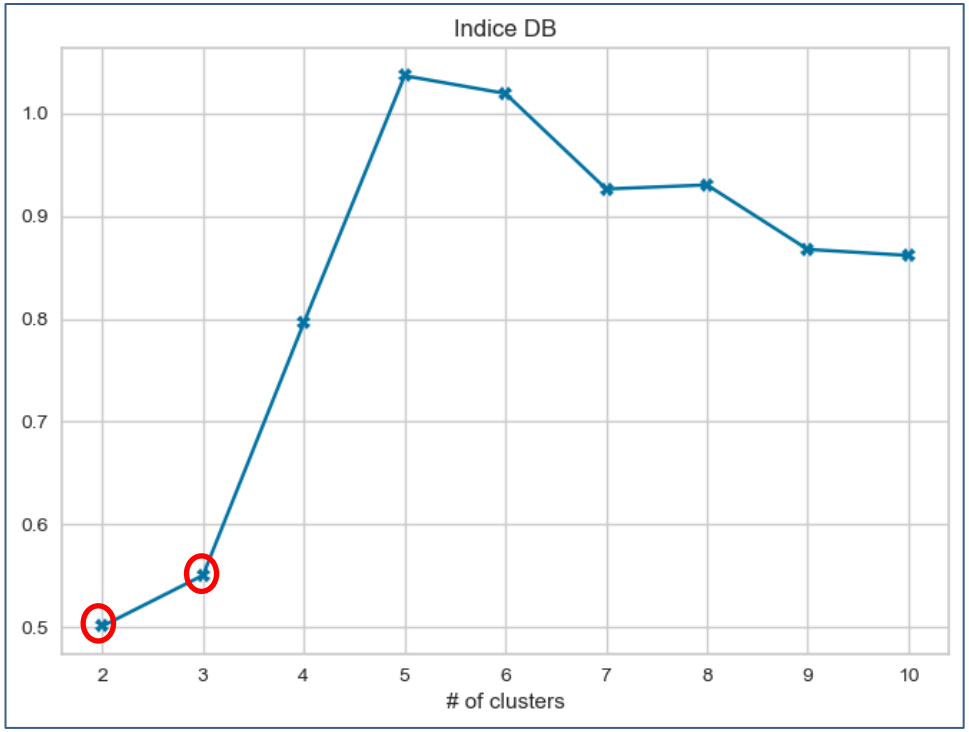
⇨ (DB ≥ 0), plus faible est sa valeur, plus les clusters sont compacts et discernables. DB n'est pas influencé par le nombre de classes (pourvu que K ≥ 2) puisqu'il prend le max par ligne dans la matrice R

Indice de Davies-Bouldin (2)

Outil pour le choix du nombre de classes

Puisque l'indice de Davies-Bouldin n'a pas de relation mécanique avec K (le nombre de clusters), on peut l'exploiter pour choisir la partition qui minimise la mesure.

Encore une fois, (K =2) et (K = 3) se tiennent pour nos données.



Conclusion

Conclusion

- S'appuyer sur des critères numériques pour évaluer les performances des algorithmes est inhérent au machine learning.

- Les mesures externes confrontent les classes calculées avec les véritables classes d'appartenance connues à l'avance. Cela n'arrive jamais dans une utilisation pratique des méthodes de « clustering ». Elles sont surtout utilisées en recherche pour confronter différentes approches.
- Et les mesures étudiées peuvent servir à comparer les partitions induites par différents algorithmes.

- Les mesures internes sont fondées à des degrés divers sur les notions de « compacité » et de « séparabilité ». Certaines sont insensibles au nombre de classes, nous pouvons les exploiter directement pour identifier la solution « optimale ».

Bibliographie

Ouvrages

Chandon J.L., Pinson S., « Analyse typologique – Théorie et applications », Masson, 1981.

Gan G., Ma C., Wu J., « Data Clustering – Theory, Algorithms and Applications », SIAM, 2007.

L. Lebart, A. Morineau, M. Piron, « Statistique exploratoire multidimensionnelle », Dunod, 2000.

Webographie

« Clustering performance evaluation », documentation Scikit-Learn / Python, chapitre « Clustering » (version 1.2.0, janvier 2023).

« Cluster analysis – Evaluation and assessment », page « Cluster analysis », Wikipédia (janvier 2023).