Digital Bridge
RDC

K-Means Clustering

Digital Bridge RDC

Contenu

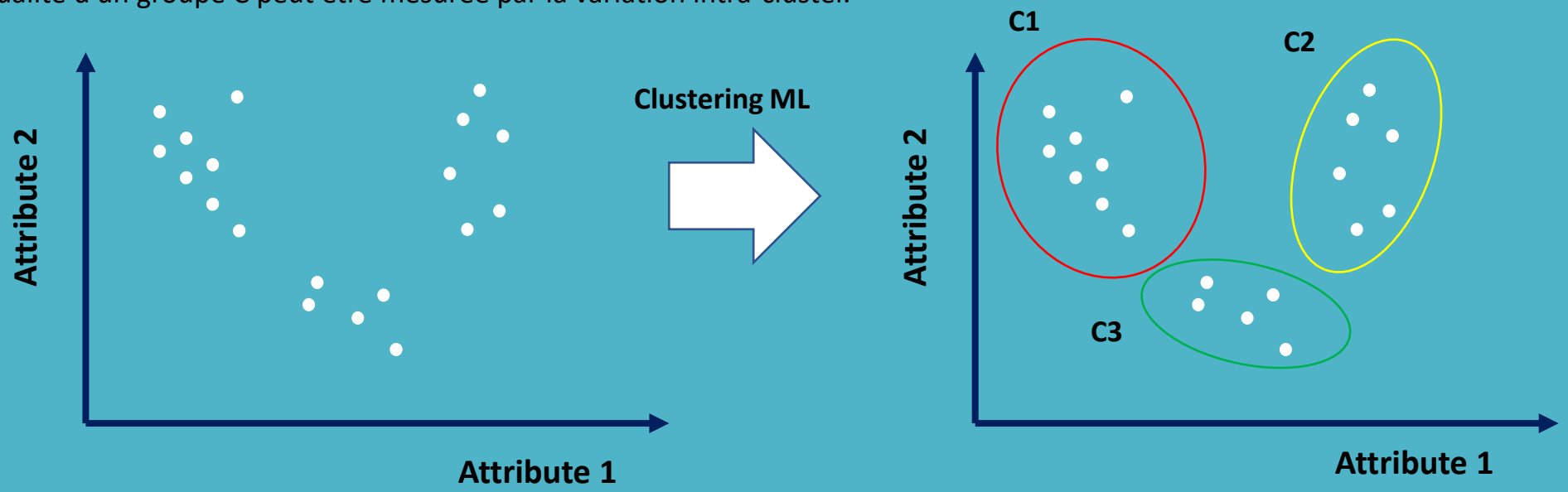
- Exploration de données non supervisée
- Énoncé du problème de clustering
- Partitionnement vs clustering hiérarchique
- Algorithme k-Means
- piège k-Means
- k-Évaluation de la performance des moyennes
- Méthode du coude pour sélectionner k
- Clustering hiérarchique
- Regroupement conflictuel et agglomératif
- Dendrogrammes
- Applications pour le clustering

Exploration de données non supervisée

- Les problèmes d'exploration de données non supervisés sont un type de problèmes d'apprentissage automatique où les données sont entraînées sans utiliser d'étiquettes/classes.
- Dans l'apprentissage non supervisé, la classification des données est généralement inconnue.
- La performance d'un algorithme non supervisé dépend d'une distribution de probabilité inconnue
- Il est possible d'utiliser l'exploration de données non supervisée comme première étape dans l'attribution d'étiquettes pour le futur algorithme d'apprentissage supervisé

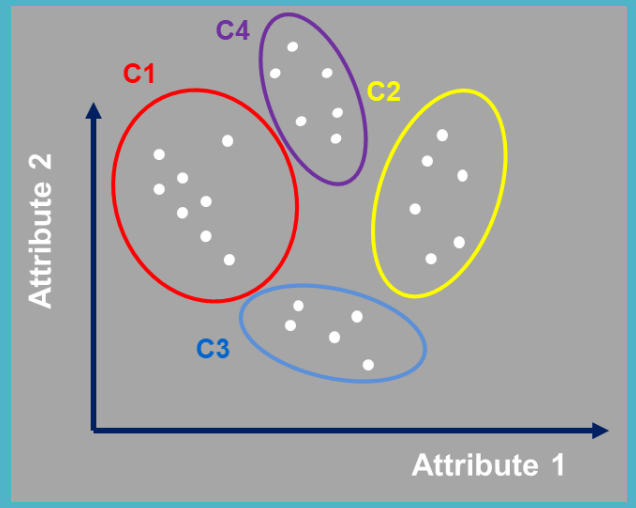
Énoncé du problème de clustering

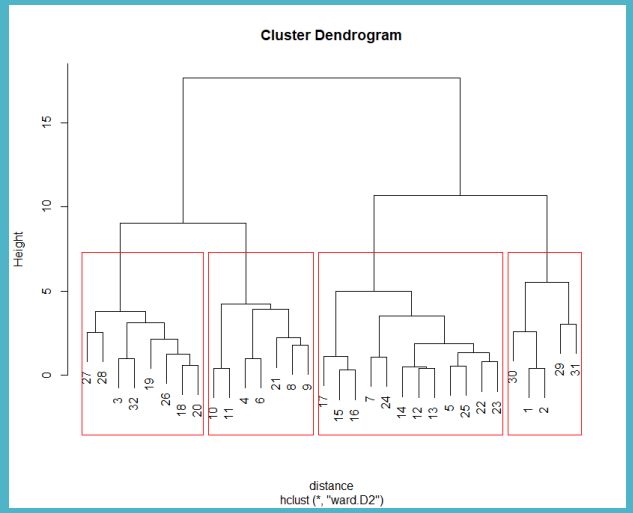
- Dans les problèmes de clustering non supervisé, nous essayons de regrouper les données en clusters avec des attributs (fonctionnalités) similaires.
- Le clustering est un problème d'apprentissage de l'attribution d'étiquettes à des exemples en tirant parti d'un jeu de données non étiqueté.
- Parce que l'ensemble de données est complètement non étiqueté, décider si le modèle appris est optimal est beaucoup plus difficile, et souvent ne vaut pas la peine d'être poursuivi que dans l'apprentissage supervisé.
- Le problème de clustering 2D et parfois 3D peut souvent être visualisé et compris intuitivement par les humains. Cependant, pour plus de quatre attributs, il est impossible de visualiser les grappes à l'aide de tracés cartésiens réguliers.
- La qualité d'un groupe C peut être mesurée par la variation intra-cluster.



Partitionnement vs clustering hiérarchique

- Bien que l'accent soit mis sur le clustering k-means une méthode de partitionnement de jeux de données, nous discuterons brièvement d'un autre type de clustering appelé clustering hiérarchique
- Partitionnement : regroupement des données en fonction de la similitude avec les centroïdes de données
- Hiérarchique : regroupement descendant ou ascendant des données en fonction de la similarité





Digital Bridge RDC

Algorithme k-Means (1)

- **Entrée**: un jeu de données D de n instances et f entités, et une cible de k clusters
- **Sortie**: k clusters
- **Étape 1:**
- Attribuer arbitrairement k centroïdes de cluster avec des coordonnées dans l'espace attributs
- **Étape 2:**
- (ré)affecter chaque instance à un cluster en fonction de sa similitude (distance) avec le centroïde du cluster
- **Étape 3:**
- Mettre à jour les centroïdes du cluster en calculant la moyenne des instances affectées au cluster
- **Étape 4:**
- Répétez les étapes 2 et 3 jusqu'à ce qu'aucune instance ne modifie l'appartenance au cluster

Digital Bridge
R D C

Algorithme K-Means (2)

K-means algorithm

The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters	$\mu_i =$ some value $, i = 1, \ldots, k$
2. Attribute the closest cluster to each data point	$\mathbf{c}_i = \{j : d(\mathbf{x}_j, \mu_i) \leq d(\mathbf{x}_j, \mu_l), l \neq i, j = 1, \ldots, n\}$
3. Set the position of each cluster to the mean of all data points belonging to that cluster	$\mu_i = \frac{1}{\lvert c_i \rvert} \sum_{j \in c_i} \mathbf{x}_j, \forall i$
4. Repeat steps 2-3 until convergence	
Notation	$\lvert \mathbf{c} \rvert =$ number of elements in $\mathbf{c}$

The algorithm eventually converges to a point, although it is not necessarily the minimum of the sum of squares. That is because the problem is non-convex and the algorithm is just a heuristic, converging to a local minimum. The algorithm stops when the assignments do not change from one iteration to the next.

Source: http://www.onmyphd.com/?p=k-means.clustering

Piège k-Means

- La sélection des centroïdes initiaux est importante dans l'algorithme k-means
- Une stratégie courante consiste à sélectionner les clusters initiaux au hasard. Une autre moins courante consiste à choisir les centres de cluster initiaux parmi les instances de jeu de données existantes.
- La sélection aléatoire des centroïdes entraînera probablement une légère différence entre les centroïdes finaux chaque fois que nous exécuterons l'algorithme k-moyennes pour un k donné, mais cela donne généralement de bons résultats lorsque le k optimal est sélectionné.
- Cependant, il est tout à fait possible de sélectionner un ensemble de centroïdes initiaux qui conduisent à une convergence très lente ou à un jeu de données très mal partitionné. C'est ce qu'on appelle le piège k-means.
- k-means++ est une méthode qui évite le piège en sélectionnant les centroïdes le plus loin possible les uns des autres lors de l'initialisation.

Évaluation de la performance k-Means: Liaison

- **Liaison simple:** la proximité de deux grappes est définie par la proximité de leurs deux points les plus proches
- **Couplage moyen:** la proximité de deux grappes est définie par la proximité de leurs centroïdes
- **Liaison complète:** la proximité de deux grappes est définie par la proximité de leurs deux points les plus éloignés

Évaluation de la performance k-Means : au sein du cluster Somme des carrés

- Dans d'autres cas, pour évaluer la performance des k-moyennes, une méthode populaire consiste à calculer le pourcentage de variance de l'ensemble de données qui s'explique par la variance au sein des grappes.
- Ce pourcentage est calculé par: (entre cluster_SS)/(Total_SS)
- La somme entre le carré pour chaque cluster est la variance des membres du cluster par rapport au centroïde du cluster

4 clusters

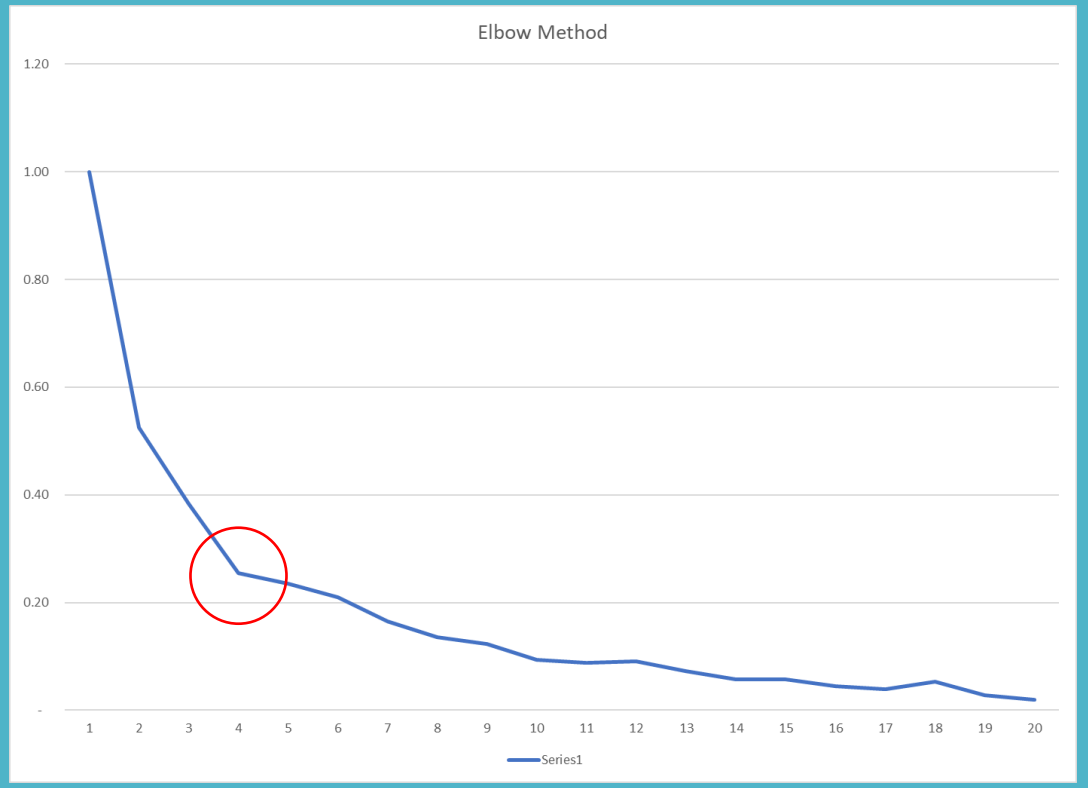
Sum of square of each cluster

Performance metric

```
> km[4]
[[1]]
K-means clustering with 4 clusters of sizes 7, 8, 5, 12

Cluster means:
         mpg        cyl        disp         hp        drat          wt
1  0.1082193 -0.5849321 -0.44867013 -0.6496905 -0.04967936 -0.02346989
2  1.3247791 -1.2248578 -1.10626771 -0.9453003  1.09820619 -1.20086981
3 -0.2639188  0.3429602 -0.05907659  0.7600688  0.44781564 -0.22101115
4 -0.8363478  1.0148821  1.02385129  0.6924910 -0.88974768  0.90635862
        qsec         vs         am       gear       carb
1  1.1854841  1.1160357 -0.8141431 -0.1573201 -0.4145882
2  0.3364684  0.8680278  1.1899014  0.7623975 -0.8125929
3 -1.2494801 -0.8680278  1.1899014  1.2367782  1.4781451
4 -0.3952280 -0.8680278 -0.8141431 -0.9318192  0.1676779

Clustering vector:
 [1] 3 3 2 1 4 1 4 1 1 1 1 4 4 4 4 4 4 2 2 2 1 4 4 4 4 2 2 2 3 3 3 2

Within cluster sum of squares by cluster:
[1] 21.28798 19.04480 23.40276 23.08349
 (between_SS / total_SS =  74.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"

> _
```

Méthode du coude pour sélectionner k

- Les méthodes du coude que nous avons utilisées dans k-NN peuvent être appliquées en k-moyennes pour déterminer la taille optimale du cluster pour un ensemble de données donné.
- Il s'agit de calculer la métrique de performance pour une plage de k valeurs, de tracer la courbe pour déterminer le k optimal où l'augmentation de k ne donne pas de performances significativement plus élevées.



Digital Bridge RDC

Clustering hiérarchique

- Clustering ascendant aka agglomératif : commence avec chaque point de données en tant que cluster, puis combine récursivement des paires de clusters en plus grands jusqu'à ce qu'il y ait un grand cluster.
- Clustering descendant aka divisive: commence à partir d'un grand ensemble de données et se divise récursivement en plus petits

Mise en cluster d’applications

- Vision par ordinateur : segmentation d’images
- Détection de fraude à l’assurance
- cyber-profilage des criminels
- Segmentation du marché
- Génétique
- Optimisation de l’emplacement des magasins de livraison
- Identification des localités criminelles
- Biologie de l'évolution