جامعة دمشق

كلية الهندسة المعلوماتية

قسم الذكاء الصنعي

K_means

مهند الطباع محمد سامي العش

محمد علاء خير الله

● k-means

أحد أنواع خوارزميات التجميع الغير خاضع للإشراف، حيث يتم تقسيم مجموعة من البيانات D، والمكونة من n عنصراً (غرضاً) إلى k قسم (مجموعة أو صنف أو صف أو عنقود)، بحيث تكون العناصر متشابهة ضمن القسم الواحد، ومختلفة عن العناصر في باقي الأقسام.

تعتمد هذه الخوارزمية على جعل تابع الخطأ أصغر ما يمكن، والذي نحصل عليه من مجموع مربعات المسافات بين العناصر في كل قسم ومركز هذا القسم، وذلك وفق العلاقة التالية:

$$SSE(S) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - X_k||^2$$

ألية عمل الخوارزمية:

دخل الخوارزمية: 1-عدد العناقيد k. 2- الأغراض.

خرج الخوارزمية: 1- توزع الأغراض على هذه العناقيد. 2- مراكز العناقيد.

آلية العمل: 1- اختيار k نقطة بشكل عشوائي واعتبارها مراكز العناقيد.

2- تشكيل العناقيد بحساب بُعد كل نقطة من نقاط البيانات عن هذه المراكز وإسنادها إلى العنقود ذو المركز الأقرب.

3- إعادة حساب مراكز العناقيد الجديدة (متوسطات النقاط المسندة إليها).

4- يتم تكرار الخطوتان 2، 3 حتى يصبح التغير في مراكز العناقيد معدوماً أو الوصول إلى عدد معين (محدد مسبقاً) من التكرارات أو الوصول إلى قيمة خطأ مقبولة (تكون محددة مسبقاً).

إيجابيات الخوارزمية:

فعالة: إن تعقيدها $O(tKn)$ حيث t عدد التكرارات، ولا يمكن ل K و t أن تكون أكبر من n.

سلبيات الخوارزمية:

1- تتأثر النتائج بالمراكز البدائية.
2- تحتاج لتحديد عدد العناقيد.
3- تتأثر بقيم الضجيج وبقيم النقاط الشاذة.
4- لا يمكنها إيجاد عناقيد ذات أشكال غير محدبة.

استعمال الخوارزمية:

- تستخدم بشكل مستقل للتصنيف ودراسة خصائص المجموعات.
- ضغط / تلخيص البيانات أو تقليص الأبعاد: عند معالجة الصور أو الكلام يمكن ضغط الإشارات الصوتية أو الصورية بشعاع واحد.
- في أنظمة الاقتراح: في الخوارزميات التعاونية يمكن تجميع المستخدمين المتشابهين مع بعضهم.

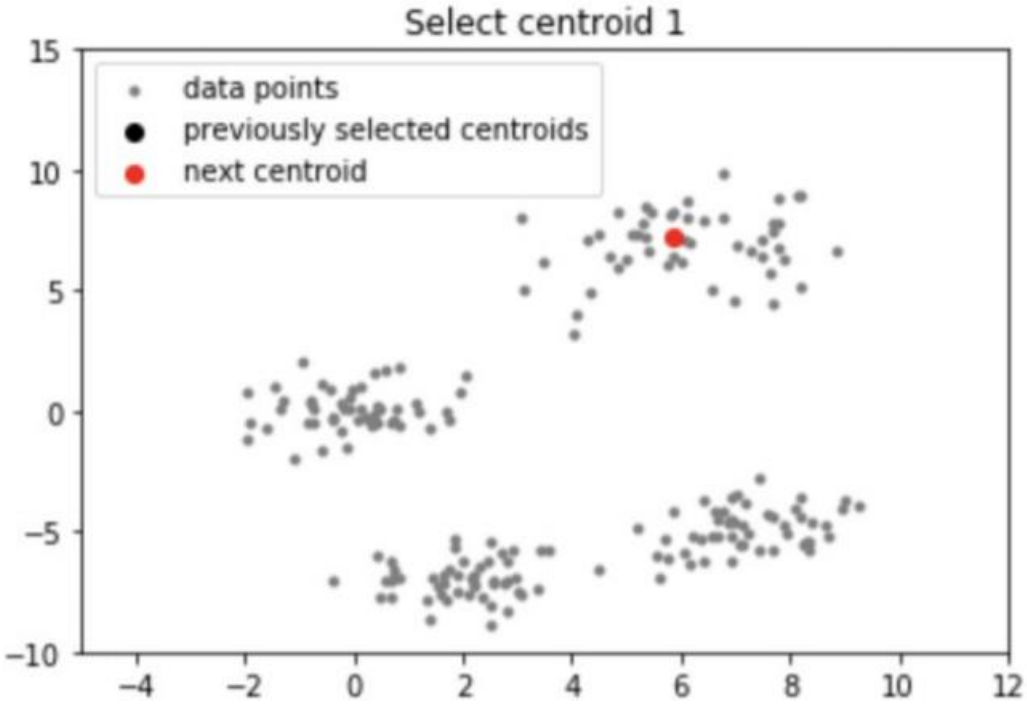
يمكن تطبيق الخوارزمية باستعمال البيانات الموجودة في هذا الرابط:

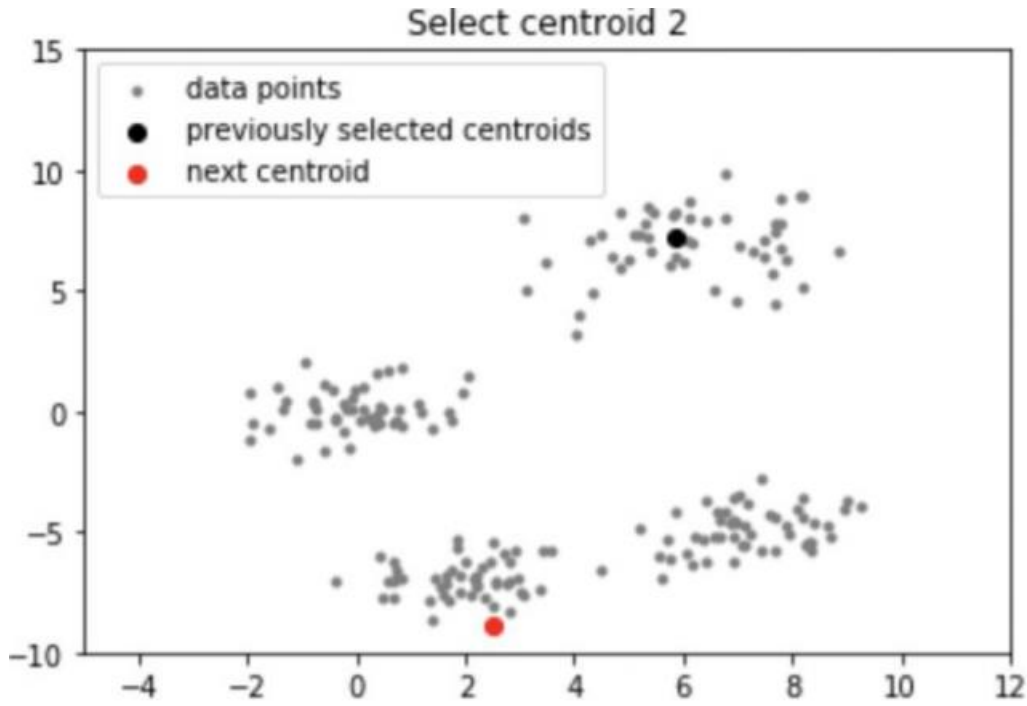
https://www.kaggle.com/jchen2186/machine-learning-with-iris-dataset

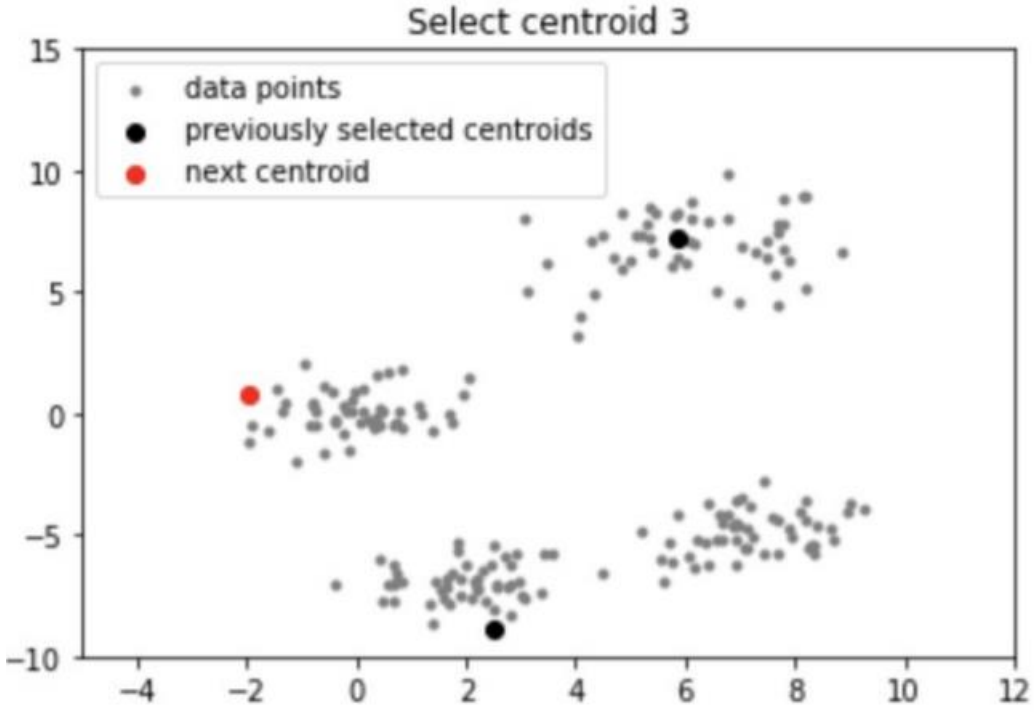
++k-means: ●

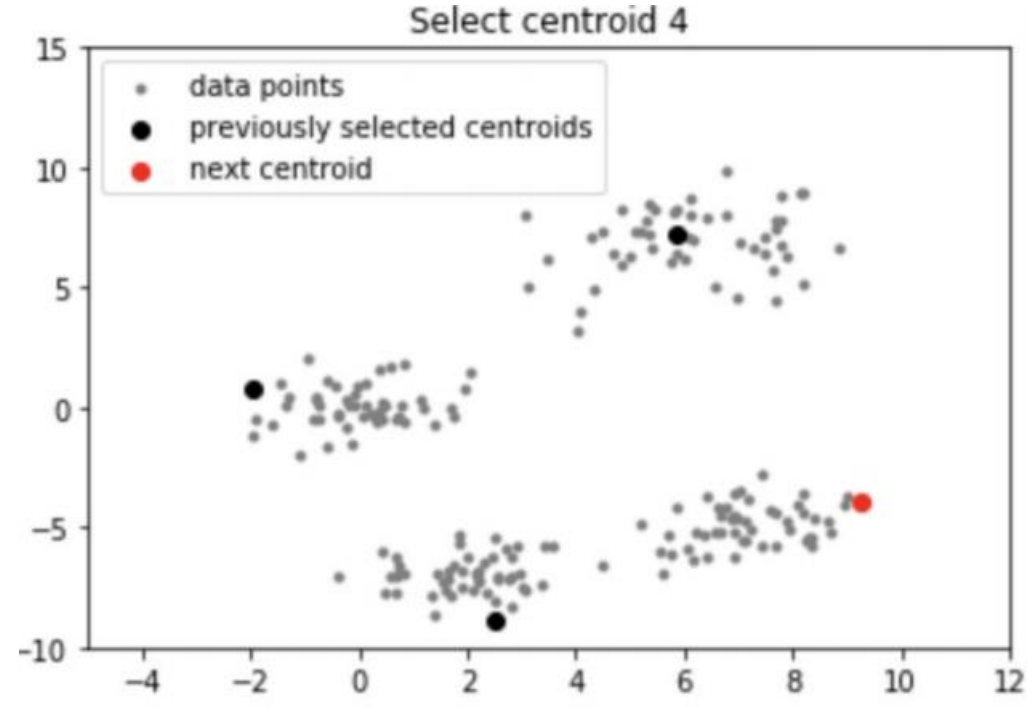
نسخة مطورة من خوارزمية ال k-means لحل مشكلة التهيئة العشوائية للمراكز.

تعمل هذه الخوارزمية بنفس خطوات عمل خوارزمية ال k-means مع اختلاف الخطوة الأولى، فيتم في هذه الخوارزمية اختيار أول مركز بشكل عشوائي، ومن ثم يتم اختيار باقي المراكز بحيث تكون أبعد ما يمكن عن بعضها.






Select centroid 3
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12


Select centroid 4
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12

● K-Medoids:

نسخة مطورة من خوارزمية ال k-means لحل مشكلة التأثر بالقيم الشاذة، حيث أنَّ خوارزمية k-means تتأثر بالقيم الشاذة فتحدد مركز العنقود بمتوسط القيم المسندة لهذا العنقود، وحساب المتوسط يتأثر بكل النقاط داخل العنقود. حلت خوارزمية ال K-Medoids هذه المشكلة باعتبار النقاط الوسطى في كل عنقود هي المراكز.

ألية عمل الخوارزمية:

دخل الخوارزمية: 1-عدد العناقيد k. 2- الأغراض.

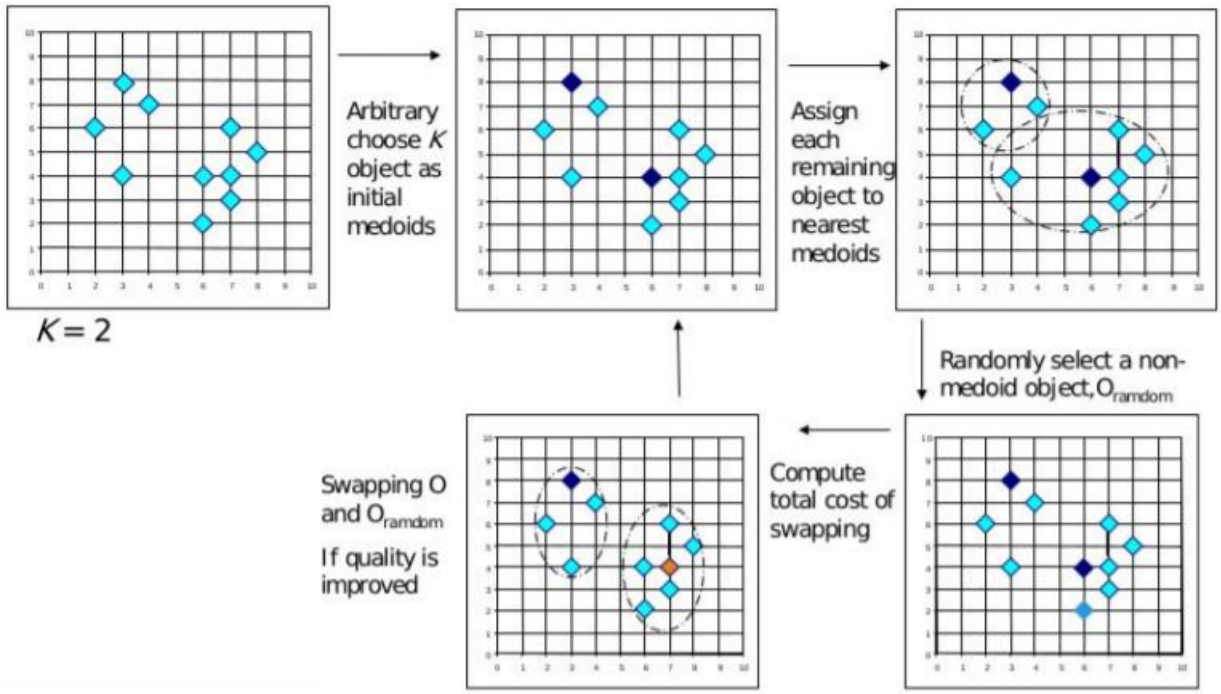
خرج الخوارزمية: 1- توزع الأغراض على هذه العناقيد. 2- مراكز العناقيد.

آلية العمل: 1- اختيار k نقطة بشكل عشوائي واعتبارها مراكز العناقيد.

2- تشكيل العناقيد بحساب بُعد كل نقطة من نقاط البيانات عن هذه المراكز وإسنادها إلى العنقود ذو المركز الأقرب.

3- إعادة حساب مراكز العناقيد الجديدة (اختيار نقطة عشوائية O_i حساب كلفة استبدال المركز الحالي للعنقود التي تنتمي إليه النقطة O_i بالنقطة O_i. إذا كانت كلفة الاستبدال أكبر من الصفر يتم اختيار نقطة أخرى وإلا تصبح O_i المركز الجديد).

4- يتم تكرار الخطوتان 2، 3 حتى يصبح التغير في مراكز العناقيد معدوماً أو الوصول إلى عدد معين (محدد مسبقاً) من التكرارات أو الوصول إلى قيمة خطأ مقبولة (تكون محددة مسبقاً).



• K-Medians:

نسخة مطورة من خوارزمية ال k-means لحل مشكلة التأثر بالقيم الشاذة، حيث أنَّ خوارزمية k-means تتأثر بالقيم الشاذة فتحدد مركز العنقود بمتوسط القيم المسندة لهذا العنقود، وحساب المتوسط يتأثر بكل النقاط داخل العنقود، أما القيم المتوسطة فهي أقل حساسية للقيم الشاذة (المتطرفة).

تعتمد هذه الخوارزمية على جعل تابع الخطأ أصغر ما يمكن، والذي نحصل عليه من مجموع القيم المطلقة للمسافات بين العناصر في كل قسم ومركز هذا القسم (القيم الوسطى)، وذلك وفق العلاقة التالية:

$$SSE(S) = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_{ij} - med_{kj}|$$

آلية عمل الخوارزمية:

دخل الخوارزمية: 1-عدد العناقيد k. 2- الأغراض.

خرج الخوارزمية: 1- توزع الأغراض على هذه العناقيد. 2- مراكز العناقيد.

آلية العمل: 1- اختيار k نقطة بشكل عشوائي واعتبارها مراكز العناقيد.

2- تشكيل العناقيد بإسناد كل نقطة للعنقود الذي يملك أقرب وسيط.

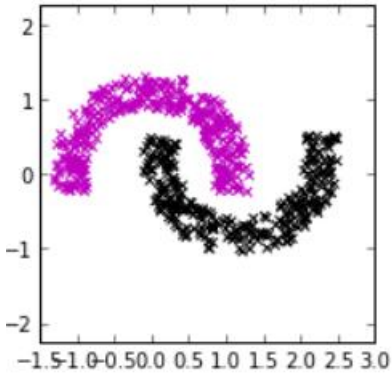
3- إعادة حساب مراكز العناقيد الجديدة (القيم الوسطى للنقاط المسندة إليها).

4- يتم تكرار الخطوتان 2، 3 حتى يصبح التغير في مراكز العناقيد معدوماً أو الوصول إلى عدد معين (محدد مسبقاً) من التكرارات أو الوصول إلى قيمة خطأ مقبولة (تكون محددة مسبقاً).

• Kernel K-Means:

نسخة مطورة من خوارزمية ال k-means لإيجاد عناقيد ذات أشكال غير محدبة.

يتم إسقاط نقاط البيانات إلى فضاء أبعاده أكبر ولكنه قابل للفصل خطياً، ثم تطبيق k-means على النقاط في الفضاء الجديد.



ألية عمل الخوارزمية:

دخل الخوارزمية: 1-عدد العناقيد k. 2- الأغراض 3- النواة.

خرج الخوارزمية: 1- توزع الأغراض على هذه العناقيد. 2- مراكز العناقيد.

آلية العمل: 1- اختيار k نقطة بشكل عشوائي واعتبارها مراكز العناقيد.

2- يتم حساب المسافة أو التشابه $d(X_n, \mu_k) = || \Phi(X_n) - \Phi(\mu_k) ||$ بعد إسقاط النقاط إلى الفضاء الجديد عن طريق تطبيق النواة

3- إعادة حساب مراكز العناقيد الجديدة

4- يتم تكرار الخطوتان 2، 3 حتى يصبح التغير في مراكز العناقيد معدوماً أو الوصول إلى عدد معين (محدد مسبقاً) من التكرارات أو الوصول إلى قيمة خطأ مقبولة (تكون محددة مسبقاً).

إيجابيات الخوارزمية:

تجميع البيانات التي شكل البيانات ضمنها غير محدب.

سلبيات الخوارزمية:

تعقيدها كبير لأنها بحاجة إلى تطبيق مصفوفة طي على كل نقطتين من البيانات للتحويل إلى الفضاء الجديد، ويصبح تابع الخطأ:

$$SSE(S) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left|\emptyset(x_i) - X_k\right|\right|^2$$

يمكن تطبيق الخوارزمية باستعمال البيانات الموجودة في هذا الرابط:

https://pafnuty.wordpress.com/2013/08/14/non-convex-sets-with-k-means-and-hierarchical-clustering/

● تقيم جودة العناقيد:

- يستخدم مصطلح التحقق العنقودي لتصميم إجراء تقييم جودة نتائج خوارزمية التجميع وكذلك من اجل المقارنة بين خوارزميتين للتجميع.

-بشكل عام يمكن تصنيف اجرائيات التحقق من صحة العنقود إلى فئتين:

1. التحقق من صحة العنقود الداخلي:

اخترع هذا النهج عام 2004، حيث يتم تقييم جودة بنية التجميع باستخدام المعلومات الداخلية لعملية التجميع دون الرجوع الى المعلومات الخارجية، ويمكن استخدامه لتقدير عدد المجموعات وخوارزمية التجميع المناسبة دون أي بيانات خارجية.

يعتمد هذا النهج على حساب معاملين:

1. انضغاط (تماسك) العنقود: يقيس مدى تشابه الأغراض مع بعضها داخل نفس العنقود، حيث أنّ التشابه الأكبر يعني تماسكاً أكبر، أي جودة أفضل.
2. تمايز(فصل) العنقود: يقيس مدى اختلاف عناصر العنقود عن عناصر العناقيد الأخرى، وتتم من خلال حساب ما يلي:
   - المسافات بين مراكز العناقيد
   - المسافة بين أقرب غرضين من أجل كل عنقودين من العناقيد المختلفة.

فتصبح العلاقة النهائية كما يلي:

$$\square\square\square\square\square\square\square\ \square\square\square\square\square\square\square\ \square\square\square = \frac{(\alpha\ *\ \square\square\square\square\square\square\square\ \square\square\square\square\square)}{(\beta\ *\ \square\square\square\square\square\square\square\ \square\square\square\square\square)}$$

حيث: α وβ هي عبارة عن أوزان لإعطاء أفضلية لأحد المعاملين.

2. التحقق من صحة العنقود الخارجي:

اخترع هذا النهج عام 2008، يتم مقارنة نتائج تحليل العنقود بنتيجة معروفة خارجياً، أي يتم مقارنة نتائج خوارزمية التجميع مع النتائج الصحيحة، يتم استخدام هذا النهج بشكل أساسي لاختيار خوارزمية التجميع المناسبة لمجموعة بيانات محددة.

- # k-means

A type of unsupervised clustering algorithm, suppose a data set, D, contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters. the objects within a cluster are similar to one another but dissimilar to objects in other clusters.

This algorithm is based on making minimize for error function, which is the sum of squared error between all objects in C_i and the centroid C, defined as

$$SSE(S) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left||x_i - X_k|\right|^2$$

Algorithm: k-means:

Input: k: the number of clusters, D: a data set containing n objects.

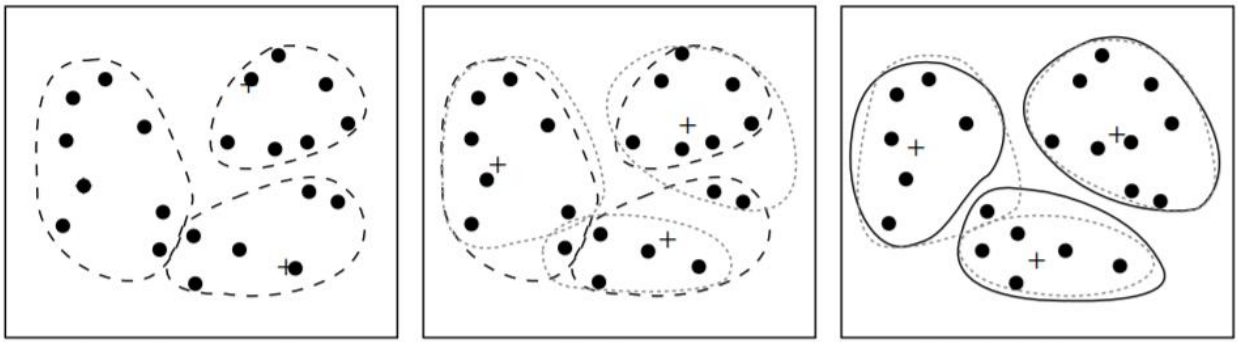
Output: A set of k clusters.

Method: (1) arbitrarily choose k objects from D as the initial cluster centers;

(2) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(3) update the cluster means, that is, calculate the mean value of the objects for each cluster;

(4) repeat step 2 & step 3 until the positions of the centers become fixed and do not change..



(a) Initial clustering (b) Iterate (c) Final clustering

Advantages of the algorithm:

Efficiency: because it has linear complexity $O(tkn)$, where t is the number of iterations.

disadvantages of the algorithm:

1- Results are affected by primitive centers.
2- Need to specify the number of clusters.
3- It is affected by noise values and outliers.
4- It cannot find clusters of non-convex shapes.

Algorithm usage:

• Used independently for classification and study of the characteristics of groups.

• Compress/summarize data or reduce dimensions: When processing images or speech, audio or visual signals can be compressed with a single vector.

• In recommendation systems: in collaborative algorithms, similar users can be grouped together.

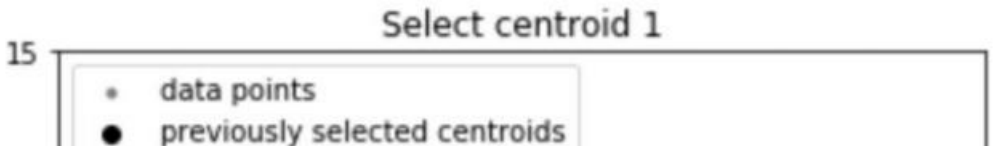
The algorithm can be implemented using the data in this link:

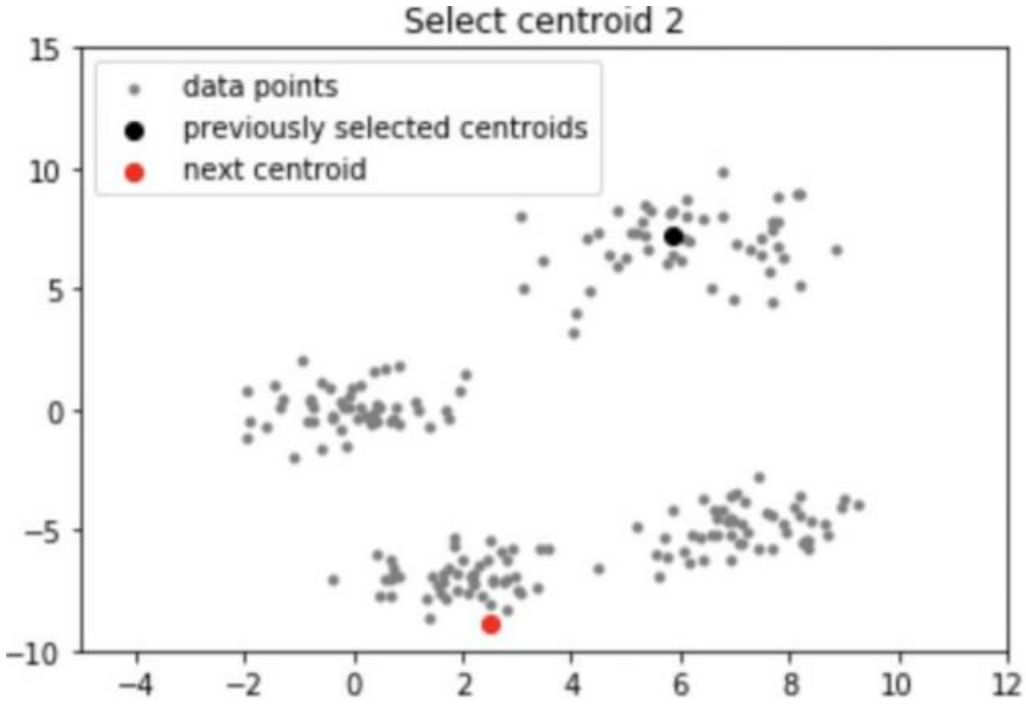
https://www.kaggle.com/jchen2186/machine-learning-with-iris-dataset

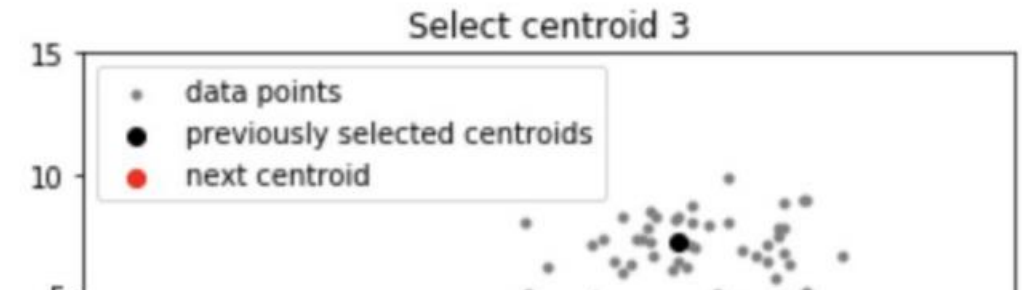
• k-means++:

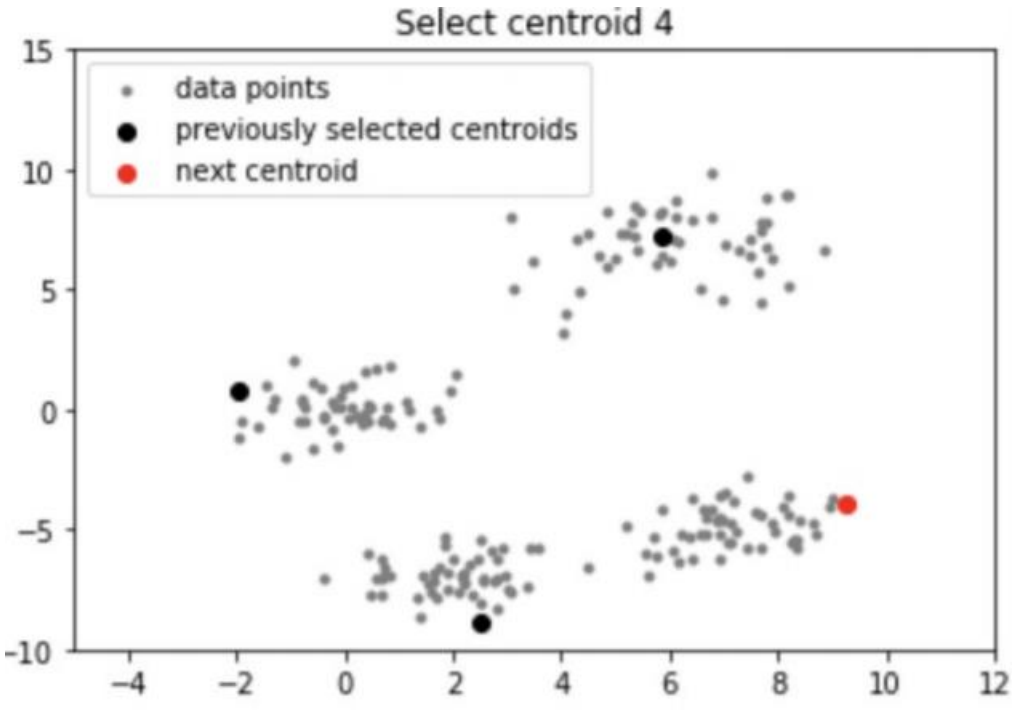
An upgraded version of the k-means algorithm to solve the problem of random initialization of centers.

This algorithm works in the same steps as the k-means algorithm, with the difference in the first step. In this algorithm, the first center is chosen randomly, and then the rest of the centers are chosen so that they are as far from each other as possible.




Select centroid 2
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12


Select centroid 3
data points
previously selected centroids
next centroid
15
10


Select centroid 4
data points
previously selected centroids
next centroid
15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12

- **K-Medoids:**

An upgraded version of the k-means algorithm to solve the problem of being affected by outliers.

The k-means algorithm is sensitive to outliers because such objects are far away from the majority of the data, and thus, when assigned to a cluster, they can dramatically distort the mean value of the cluster. This inadvertently affects the assignment of other objects to clusters. This effect is particularly exacerbated due to the use of the squared-error function.

Algorithm: k-medoids:

Input: k: the number of clusters, D: a data set containing n objects.

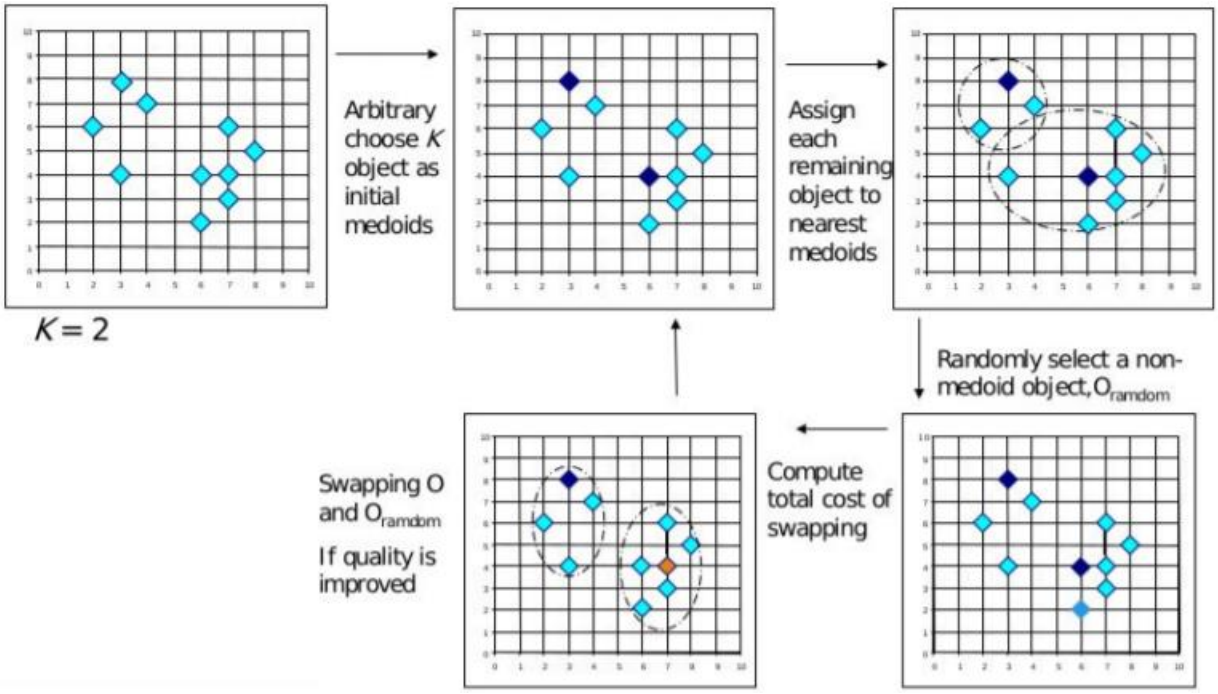
Output: A set of k clusters.

Method: (1) arbitrarily choose k objects in D as the initial representative objects or seeds;

(2) assign each remaining object to the cluster with the nearest representative object;

(3) randomly select a nonrepresentative object, o_{random}; compute the total cost, S, of swapping representative object, o_j , with o_{random}; if S < 0 then swap o_j with $o_{random}$to form the new set of k representative objects;

(4) repeat 2 and 3 until no change;



● K-Medians:

An upgraded version of the k-means algorithm to solve the problem of being affected by outliers.

The k-means algorithm is sensitive to outliers because such objects are far away from the majority of the data, and thus, when assigned to a cluster, they can dramatically distort the mean value of the cluster. This inadvertently affects the assignment of other objects to clusters. This effect is particularly exacerbated due to the use of the squared-error function.

this algorithm uses Medians values because it less sensitive to outliers. This algorithm is based on reducing the error function, which we get from the sum of the absolute values of the distances between the objects in each cluster and the center of this cluster (the values of the averages), according to the following equation:

$$SSE(S) = \sum_{k=1}^{K} \sum_{x_i \in C_k} |x_{ij} - med_{kj}|$$

Algorithm: k- Medians:

Input: k: the number of clusters, D: a data set containing n objects.

Output: A set of k clusters.

Method: (1) arbitrarily choose k objects from D as the initial cluster centers;

(2) (re)assign each object to the cluster to which the object is the most similar, based on the medians value of the objects in the cluster;

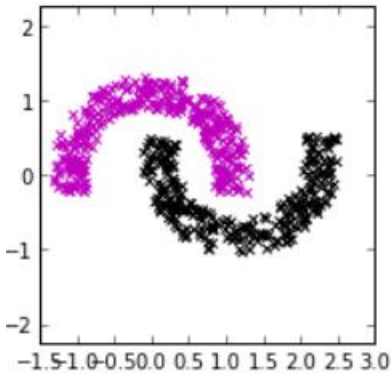
(3) update the cluster means, that is, calculate the medians value of the objects for each cluster;

(4) repeat 2 and 3 until no change, or reach a certain number of iterations.

● Kernel K-Means:

An upgraded version of the k-means algorithm for finding clusters of non-convex shapes.

The data points are dropped into a larger dimensional space, but linearly separable, and then k-means are applied to the points in the new space.



Algorithm: kernel k-means:

Input: k: the number of clusters, D: a data set containing n objects, kernel.

Output: A set of k clusters.

Method: (1) arbitrarily choose k objects from D as the initial cluster centers;

(2) calculating distance / similarity $d(X_n, \mu_k) = || \Phi(X_n) - \Phi(\mu_k) ||$ between points after projecting them into new space by applying the kernel

(3) update the cluster means, that is, calculate the mean value of the objects for each cluster;

(4) repeat step 2 & step 3 until the positions of the centers become fixed and do not change..

Advantages of the algorithm:

clustering data in which has non convex shape.

disadvantages of the algorithm:

clustering data in which has non convex shape.

It has large complexity because it needs to apply a kernel matrix to every two data points to convert to the new space, and it becomes the error function:

$$SSE(S) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|| \emptyset(x_i) - X_k \right||^2$$

The algorithm can be implemented using the data in this link:

https://pafnuty.wordpress.com/2013/08/14/non-convex-sets-with-k-means-and-hierarchical-clustering/

• Evaluate the quality of clusters:

The term cluster validation is used to design the procedure of evaluating the goodness of clustering algorithm results. This is important to compare the two clustering algorithms.

Generally, clustering validation statistics can be categorized into 2 classes:

1- **Internal cluster validation:** which uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropriate clustering algorithm without any external data.

This approach is based on the calculation of two parameters:

1. **Compactness** or cluster cohesion: Measures how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e., a good clustering). The different indices for evaluating the compactness of clusters are based on distance measures such as the cluster-wise within average/median distances between observations.

2. **Separation**: Measures how well-separated a cluster is from other clusters. The indices used as separation measures include:
   - distances between cluster centers
   - the pairwise minimum distances between objects in different cluster.

2- **External cluster validation:** which consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match externally supplied class labels. Since we know the "true" cluster number in advance, this approach is mainly used for selecting the right clustering algorithm for a specific data set.

Generally, most of the indices used for internal clustering validation combine compactness and separation measures as follow:

$$Index = \frac{(\alpha \ * \ Separation)}{(\beta * \ Compactness\ )}$$

Where αα and α, β are weights.

References:

Jure Leskovec, Anad Rajaraman, Jeffery D.Ukkman. Mining of massing datasets. Chapter 7. 2014.

Micheline Kamber, Jian Pei. Data mining concepts techniques. Chapter 10.