

بيانات التعدين

محاضرات لدرجة البكالوريوس طلاب

أقسام تكنولوجيا المعلومات وعلوم الكمبيوتر واللجنة العليا

كلية دراسات الحاسوب والإحصاء

جامعة كردفان

إعداد : المحاضر : محمد عثمان بشار

قسم تكنولوجيا المعلومات-كلية دراسات الحاسوب والإحصاء

يونيو:
2022

المحاضرة 9

التحليل العنقودي

الخطوط العريضة للمحاضرة:

- التحليل العنقودي Clustering
- K-Means

ما هو التحليل العنقودي؟

- الكتلة: مجموعة من كائنات البيانات
 - متشابهة مع بعضها البعض داخل نفس المجموعة
 - تختلف عن الكائنات الموجودة في مجموعات أخرى
- التحليل العنقودي
- إيجاد أوجه التشابه بين البيانات حسب الخصائص الموجودة في البيانات وتجميع كائنات البيانات المماثلة في مجموعات
- التعلم غير الخاضع للرقابة: لا توجد فصول محددة مسبقا
- التطبيقات النموذجية
- كأداة قائمة بذاتها للحصول على نظرة ثاقبة لتوزيع البيانات • خطوة المعالجة المسبقة للخوارزميات الأخرى

أمثلة على تطبيقات التجميع

• التسويق: ساعد المسوقين على اكتشاف مجموعات متميزة في قواعد عملائهم، ثم استخدم هذه المعرفة

لتطوير برامج التسويق المستهدفة

استخدام الأراضي: تحديد المناطق ذات الاستخدام المماثل للأراضي في قاعدة بيانات رصد الأرض

التأمين: تحديد مجموعات حاملي وثائق تأمين المركبات ذوي متوسط تكلفة المطالبات المرتفع

تخطيط المدن: تحديد مجموعات من المنازل حسب نوع المنزل وقيمتها وجغرافيته

موقع

دراسات الزلازل الأرضية: ينبغي أن تتجمع مراكز الزلازل الأرضية المرصودة على طول الصدوع القارية

الجودة: ما هو التجميع الجيد؟

• طريقة التجميع الجيدة ستنتج مجموعات عالية الجودة

• التشابه العالي بين الطبقات

• انخفاض التشابه بين الطبقات

• تعتمد جودة نتائج التجميع على كل من مقياس التشابه المستخدم من قبل

الطريقة وتنفيذها

• يتم قياس جودة أسلوب التجميع أيضًا من خلال قدرته على اكتشاف بعض أو كل الأشياء

من الأنماط المخفية

قياس جودة التجميع

- **مقياس الاختلاف/التشابه:** يتم التعبير عن التشابه من حيث دالة المسافة،
متري عادة: $d(i, j)$
- **توجد وظيفة "جودة"** منفصلة تقيس "جودة" المجموعة.
- عادة ما تكون تعريفات **وظائف المسافة** مختلفة جدًا بالنسبة إلى النطاق الزمني،
المتغيرات المنطقية والفتوية والترتيبية والمتغيرات المتوجهة.
- يجب أن ترتبط الأوزان بمتغيرات مختلفة بناء على التطبيقات والبيانات
دللات.
- من الصعب تحديد عبارة "مشابه بما فيه الكفاية" أو "جيد بما فيه الكفاية". عادة
ما تكون الإجابة ذاتية إلى حد كبير.

متطلبات التجميع في استخراج البيانات

• قابلية التوسيع

• القدرة على التعامل مع أنواع مختلفة من السمات

• القدرة على التعامل مع البيانات الديناميكية

• اكتشاف التجمعات ذات الشكل العشوائي

• الحد الأدنى من متطلبات المعرفة بالمجال لتحديد معلمات الإدخال

• القدرة على التعامل مع الضوضاء والقيم المتطرفة

• غير حساس لترتيب سجلات الإدخال

• أبعاد عالية

• دمج القيود المحددة من قبل المستخدم

• إمكانية التفسير وسهولة الاستخدام

نوع البيانات في تحليل المجموعات

• المتغيرات ذات الفواصل الزمنية

• المتغيرات الثنائية

• المتغيرات الاسمية والترتيبية والنسبية

• المتغيرات ذات الأنواع المختلطة

المتغيرات ذات القيمة الفاصلة

- توحيد البيانات

- حساب متوسط الانحراف المطلق:

$$\text{متوسط} = \frac{\sum |x_i - \bar{x}|}{n}$$

—

$$\text{متوسط} = \frac{\sum |x_i - \bar{x}|}{n}$$

—

أي رقم يليها

- حساب القياس الموحد (z-score)

$$z = \frac{x - \bar{x}}{s}$$

إذا و

- استخدام متوسط الانحراف المطلق أقوى من استخدامه

الانحراف المعياري

التشابه والاختلاف بين الكائنات

• تستخدم المسافات عادة لقياس التشابه أو

الاختلاف بين كائنين البيانات

• بعض منها شعبية تشمل: مسافة مين柯فسكي:

$$\text{مسافة مين柯فسكي} = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \dots + |x_n - y_n|^p}$$

الحادي عشر (ط، ي) \Rightarrow مسافة مين柯فسكي

حيث $(x_i^1, x_i^2, \dots, x_i^p)$ و $(y_j^1, y_j^2, \dots, y_j^p)$ زهما كائنان بيانات ذات أبعاد ،

و عدد صحيح موجب

إذا كانت $d_{ij} = 1$ فإن d_{ij} هي مسافة مانهاتن

$$d_{ij} = |x_i^1 - y_j^1| + |x_i^2 - y_j^2| + \dots + |x_i^n - y_j^n|$$

التشابه والاختلاف بين الكائنات (تابع)

• إذا كانت $d = q^2$ هي المسافة الإقلية:

$$\sqrt{d_{ij}} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

• ملكيات

• $d(t, y) = 0$

• $d(t, t) = 0$

• $d(t, y) = d(y, t)$

• $d(t, y) \leq d(t, k) + d(k, y)$

• أيضاً، يمكن استخدام المسافة الموزونة، بيرسون البارامتري

الارتباط لحظة المنتج، أو تدابير التباين الأخرى

المتغيرات الثنائية

		كائن ي
	1 0	مجموع
	جدول طوارئ للثنائي	نات
ج	كائنات	dc@
	a@bdp	مجموع

• قياس المسافة المتماثلة

المتغير الثنائي:

قياس المسافة للمتغيرات الثنائية غير المتماثلة:

د) ج

فَبِلِ الْمِيلَادِ

معامل الجاكار (التشابه)

فياس المتغيرات الثنائية غير المتماثلة ():

سیم

جاکارڈ

۱۰
ای (عای) ی

الاختلاف بين المتغيرات الثنائية

• مثال

الاسم الجنس اختبار السعال والحمى 1- الاختبار-2 الاختبار-3 الاختبار-4	جاك مينين
	Y N P N P N لاري فـ جـيم

• الجنس هو سمة متتماثلة

• أما السمات المتبقية فهي ثنائية غير متتماثلة

• دع القيمتين Y و P مضبوطتين على 1، وقيمة N مضبوطة على 0

$$\begin{array}{r} 0 \ 1 \\ 2 \ 0 \ 1 \\ \hline (د جاك ماري) \end{array} \quad 0.3 \ 3$$

$$\begin{array}{r} 1 \ 1 \\ 1 \ 1 \ 1 \\ \hline (د جاك جيم) \end{array} \quad 0.6 \ 7$$

$$\begin{array}{r} 1 \ 0 \\ 1 \ 1 \ 0 \\ \hline (د جيم ماري) \end{array} \quad 0.7 \ 5$$

استخراج البيانات: المفاهيم والتقنيات

المتغيرات الاسمية

• تعميم المتغير الثنائي من حيث أنه يمكن أن يستغرق أكثر من ذلك من ولايتين، على سبيل المثال، الأحمر والأصفر والأزرق والأخضر

• الطريقة الأولى: المطابقة البسيطة

• م: عدد التطابقات، ع: إجمالي عدد المتغيرات

com.pmdjj

ص

• الطريقة الثانية: استخدام عدد كبير من المتغيرات الثنائية

• إنشاء متغير ثنائي جديد لكل من M الاسمية

تنص على

المتغيرات الترتيبية

• يمكن أن يكون المتغير الترتيبی منفصلأً أو مستمراً

• الترتيب مهم، على سبيل المثال، الرتبة

• يمكن التعامل معها على أنها متدرجة على فترات

• استبدل $i-th$ بترتيبهم

قم بتعيين نطاق كل متغير على $[0, 1]$ عن طريق استبدال $i-th$ كائن في المتغير $f-th$ بواسطة

ص	لو
إذا	م

• حساب الاختلاف باستخدام طرق القياس الفاصل

المتغيرات

مناهج التجميع الرئيسية (I)

• أسلوب التقسيم:

▪ إنشاء أقسام مختلفة ومن ثم تقييمها ببعض المعايير، على سبيل المثال، تقليل المجموع

أخطاء مربعة

• الطرق النموذجية: k-means, k-medoids, CLRANS

• النهج الهرمي:

▪ إنشاء تحليل هرمي لمجموعة البيانات (أو الكائنات) باستخدام بعض المعايير

• الطرق النموذجية: ديانا، أغنيس، بيرش، روك، كاميليون

• النهج القائم على الكثافة:

▪ بناء على وظائف الاتصال والكثافة

• الطرق النموذجية: DBSCAN, OPTICS, DenClue

مناهج التجميع الرئيسية (II)

• النهج القائم على الشبكة:

• يعتمد على بنية تفصيلية متعددة المستويات

• الطرق النموذجية: STING, WaveCluster, CLIQUE

• على أساس النموذج:

• يتم افتراض نموذج لكل مجموعة من المجموعات ومحاولة العثور على أفضل نموذج يناسب كل مجموعة آخر

• الطرق النموذجية: EM, SOM, COBWEB

• النمط المتكرر:

• بناء على تحليل الأنماط المتكررة

• الطرق النموذجية: pCluster

• موجه للمستخدم أو قائم على القيود:

• التجميع من خلال النظر في القيود المحددة من قبل المستخدم أو القيود الخاصة بالتطبيق

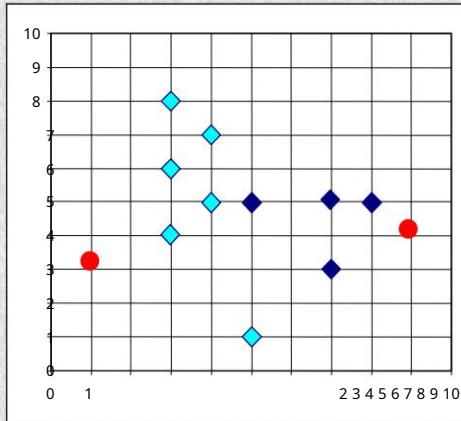
• الطرق النموذجية: COD(العوائق)، التجميع المقيد

طريقة التجميع K-Means

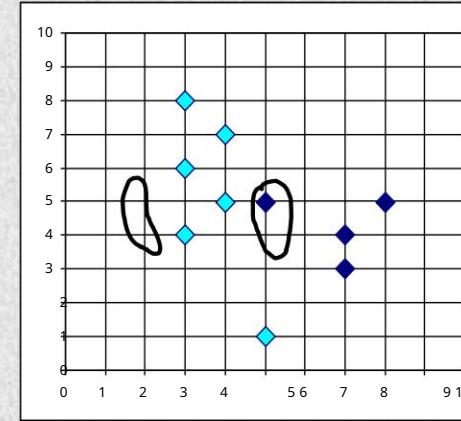
- بالنظر إلى ، يتم تنفيذ خوارزمية k-means في أربع خطوات:
- تقسيم الكائنات إلى مجموعات فرعية غير فارغة
- حساب نقاط البدور باعتبارها النقط الوسطى من مجموعات
القسم الحالي (النقطة الوسطى هي المركز، أي **النقطة المتوسطة** للمجموعة)
- قم بتعيين كل كائن إلى المجموعة بأقرب نقطة أساسية
- ارجع إلى الخطوة ، وتوقف عندما لا يكون هناك أي مهمة جديدة

K-Means التجميع طريقة

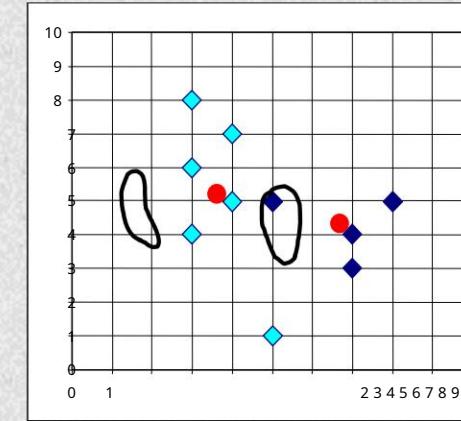
• مثال



قم بتعيين كل
كائن إلى المركز
الأكثر تشابهًا



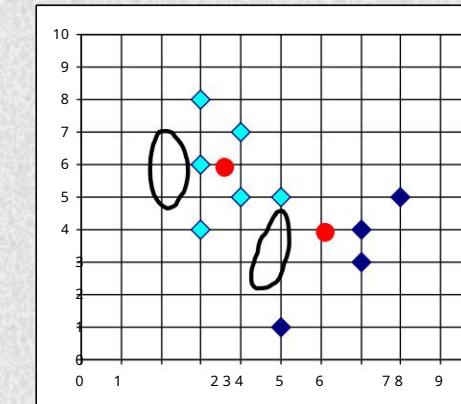
تحديث
الكتلة
وسائل



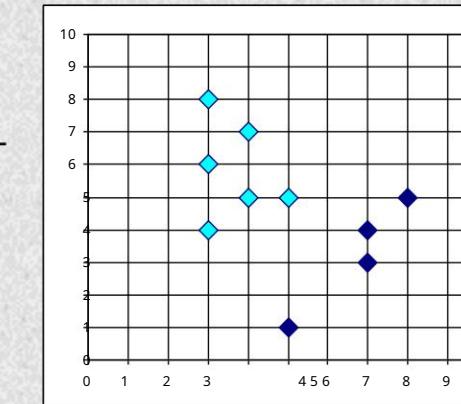
إعادة تعيين

$k = 2$

اختر كائن K بشكل تعسفي كمركز
الكتلة الأولى



تحديث
الكتلة
وسائل



K-Means على طريقة تعليقات

• القوة: كفاءة نسبية: $O(tkn)$ حيث n تمثل #كائنات، و k تمثل #مجموعات، و t تمثل #تكرارات.

عادة، $k < n$.

• المقارنة: PAM: $O(k(nk)^2)$ ، CLARA: $O(ks^2 + k(nk))$

• التعليق: غالباً ما ينتهي عند المستوى الأمثل المحلي. يمكن العثور على الأمثل العالمي باستخدام تقنيات مثل: التلدين الحتمي والخوارزميات الجينية

• ضعف

لا ينطبق إلا عندما يتم تعريف المتوسط ، فماذا عن البيانات الفئوية؟

• تحتاج إلى تحديد k عدد المجموعات ، مقدما

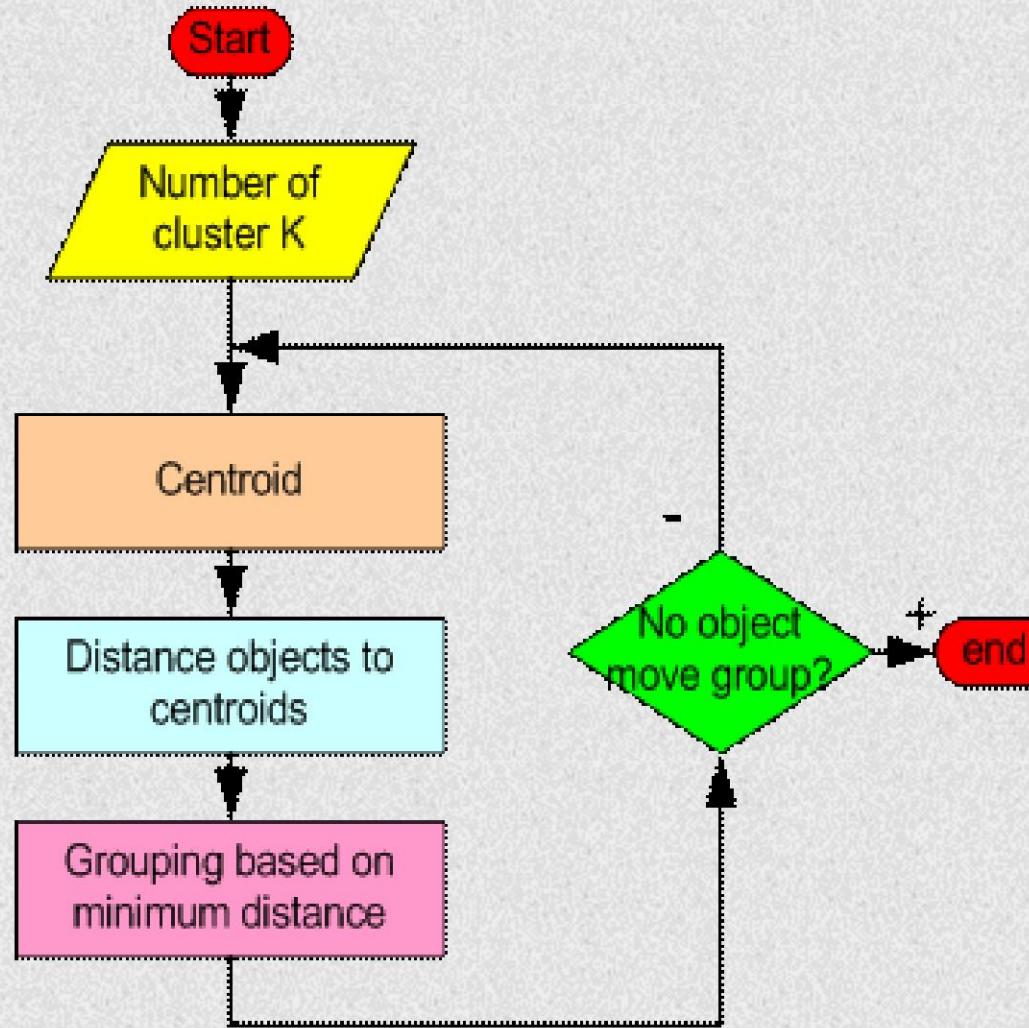
• غير قادر على التعامل مع البيانات المزعجة والقيم المتطرفة

• غير مناسب لكتشاف التجمعات ذات الأشكال غير المحدبة

K-Means في طريقة الاختلافات

- هناك عدد قليل من المتغيرات للوسائل k التي تختلف في اختيار الوسائل k الأولية
- حسابات الاختلاف
- استراتيجيات لحساب وسائل الكتلة
- التعامل مع البيانات الفئوية: أوضاع k (Huang'98)
- استبدال وسائل العنقود بالأوضاع
- استخدام مقاييس الاختلاف الجديدة للتعامل مع الأشياء الفئوية
- استخدام أسلوب يعتمد على التردد لتحديث أوضاع المجموعات
- خليط من البيانات الفئوية والعددية: طريقة النموذج k

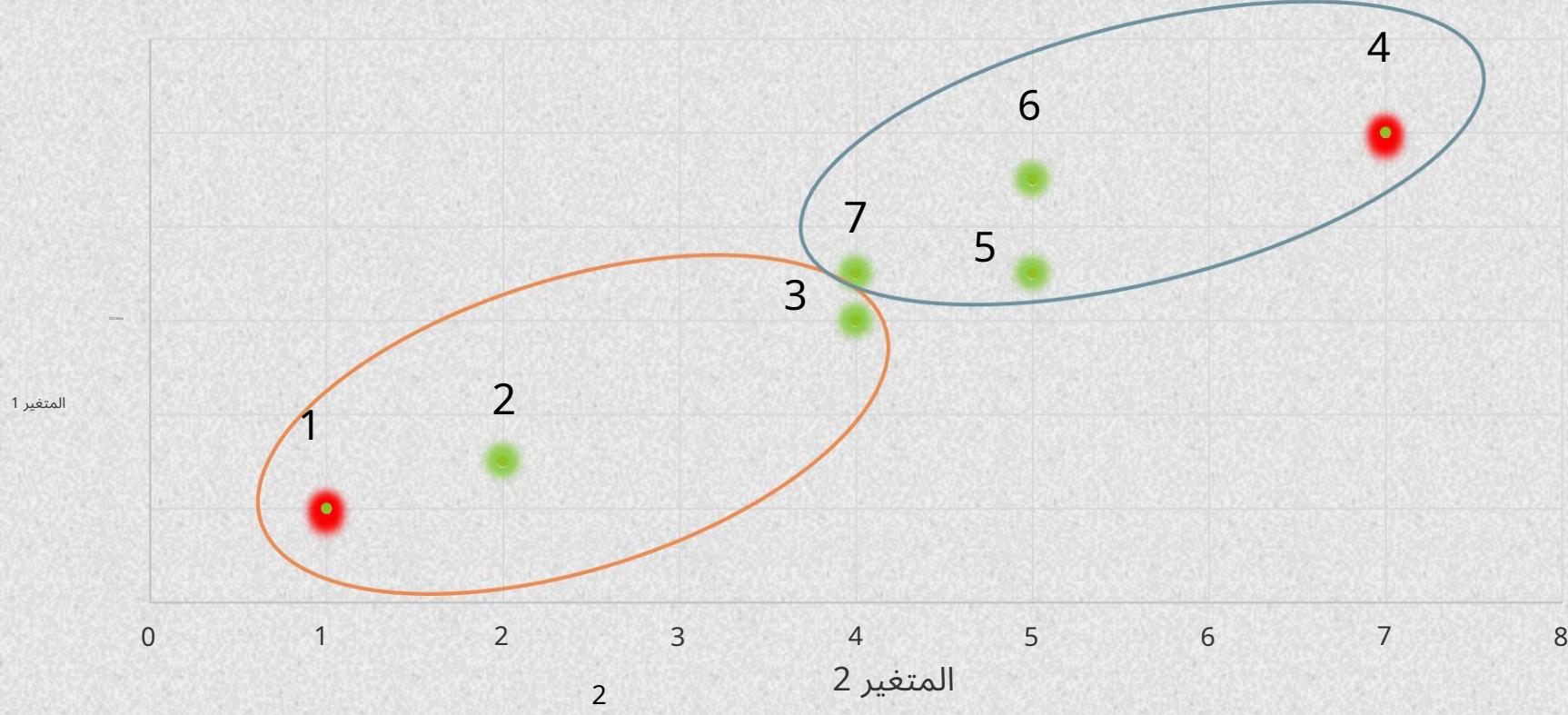
كيف تعمل خوارزمية التجميع ؟ K-



مثال بسيط يوضح تنفيذ خوارزمية k-Means (باستخدام K=2)

Individual	Variable 1	Variable 2
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

$$2 = \kappa$$



الخطوة: 1

التهيئة: نختار بشكل عشوائي اتباع النقطتين المركزيتين ($k = 2$) لمجموعتين. في هذه الحالة النقطتان المركزيتان هما: $(1.0, 1.0) = m_1$ و $(0.7, 0.5) = m_2$.

الخطوة: 1

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

الخطوة: 2

	النقطه الوسطى 1	النقطه الوسطى 2
1	$(1 - 1)^2 + (1 - 1)^2 = 0$	$(5 - 1.5)^2 + (7 - 2)^2 = 5.21$
2	$\sqrt{(1 - 1.5)^2 + (1 - 2)^2} = 1.12$	$\sqrt{(5 - 1.5)^2 + (7 - 2)^2} = 6.10$
3	$\sqrt{(1 - 3)^2 + (1 - 4)^2} = 3.61$	$\sqrt{(5 - 3)^2 + (7 - 4)^2} = 3.61$
4	$\sqrt{(1 - 5)^2 + (1 - 7)^2} = 7.21$	$\sqrt{(5 - 5)^2 + (7 - 7)^2} = 0$
5	$\sqrt{(1 - 3.5)^2 + (1 - 5)^2} = 4.72$	$\sqrt{(5 - 3.5)^2 + (7 - 5)^2} = 2.5$
6	$\sqrt{(1 - 4.5)^2 + (1 - 5)^2} = 5.31$	$\sqrt{(5 - 4.5)^2 + (7 - 5)^2} = 2.06$
7	$\sqrt{(1 - 3.5)^2 + (1 - 4.5)^2} = 4.30$	$\sqrt{(5 - 3.5)^2 + (7 - 4.5)^2} = 2.92$

الخطوة: 2

- Thus, we obtain two clusters containing:
 $\{1,2,3\}$ and $\{4,5,6,7\}$.
- Their new centroids are:

$$\text{Group 1} = \left(\frac{1+1.5+3}{3}, \frac{1+2+4}{3} \right) = (1.83, 2.33)$$

$$\text{Group 2} = \left(\frac{5+3.5+4.5+3.5}{4}, \frac{7+5+5+4.5}{4} \right) = (4.12, 5.38)$$

الخطوه: 3:

	النقطه الوسطى 1	النقطه الوسطى 2
1	$(1.83 - 2)^2 + (2.33 - 5.38)^2 = 1.83^2 + 2.33^2 - 2 \cdot 1.83 \cdot 5.38 = 2.57$	$(4.12 - 5)^2 + (5.38 - 3.8)^2 = 4.12^2 + 5.38^2 - 2 \cdot 4.12 \cdot 5.38 = 3.38$
2	$\square(1.83 - 1.5)^2 + (2.33 - 2)^2 = 0.47$	$\square(4.12 - 1.5)^2 + (5.38 - 2)^2 = 4.29$
3	$\square(1.83 - 3)^2 + (2.33 - 4)^2 = 2.04$	$\square(4.12 - 3)^2 + (5.38 - 4)^2 = 1.78$
4	$\square(1.83 - 5)^2 + (2.33 - 7)^2 = 5.64$	$\square(4.12 - 5)^2 + (5.38 - 7)^2 = 1.84$
5	$\square(1.83 - 3.5)^2 + (2.33 - 5)^2 = 3.15$	$\square(4.12 - 3.5)^2 + (5.38 - 5)^2 = 0.73$
6	$\square(1.83 - 4.5)^2 + (2.33 - 5)^2 = 3.78$	$\square(4.12 - 4.5)^2 + (5.38 - 5)^2 = 0.54$
7	$\square(1.83 - 3.5)^2 + (2.33 - 4.5)^2 = 2.74$	$\square(4.12 - 3.5)^2 + (5.38 - 4.5)^2 = 1.08$

وبالتالي فإن المجموعات الجديدة هي:

$$\{3,4,5,6,7\} \text{ و } \{1,2\}$$

$$1 = \overline{\left(\begin{matrix} \text{المجموع} \\ \text{المجموع} \end{matrix} \right)}^{\pm} = (1.25, 1.5)$$

$$+ \overline{\left(\begin{matrix} \text{المجموع} \\ \text{المجموع} \end{matrix} \right)}^{\mp} = (3.9, 5.1)$$

الخطوة: 4:

	النقطه الوسطى 1	النقطه الوسطى 2
1	$(1.25 - 1.5)^2 + (1.5 - 2)^2 = 0.56$	$(3.9 - 1.5)^2 + (5.1 - 2)^2 = 5.901$
2	$(1.25 - 3)^2 + (1.5 - 4)^2 = 3.05$	$(3.9 - 3)^2 + (5.1 - 4)^2 = 1.42$
3	$(1.25 - 5)^2 + (1.5 - 7)^2 = 6.66$	$(3.9 - 5)^2 + (5.1 - 7)^2 = 2.20$
4	$(1.25 - 3.5)^2 + (1.5 - 5)^2 = 4.16$	$(3.9 - 3.5)^2 + (5.1 - 5)^2 = 0.41$
5	$(1.25 - 4.5)^2 + (1.5 - 5)^2 = 4.78$	$(3.9 - 4.5)^2 + (5.1 - 5)^2 = 0.61$
6	$(1.25 - 3.5)^2 + (1.5 - 4.5)^2 = 3.75$	$(3.9 - 3.5)^2 + (5.1 - 4.5)^2 = 0.72$

□ لذلك، لا يوجد أي تغيير في الكتلة.
□ وبالتالي، تتوقف الخوارزمية هنا وت تكون النتيجة النهائية من مجموعتين {1,2} و {3,4,5,6,7}.