تحليل المعطيات /السنة الرابعة-إحصاء رياضي/

المحاضرات الثلاث الأولى

تحليل العنقدة

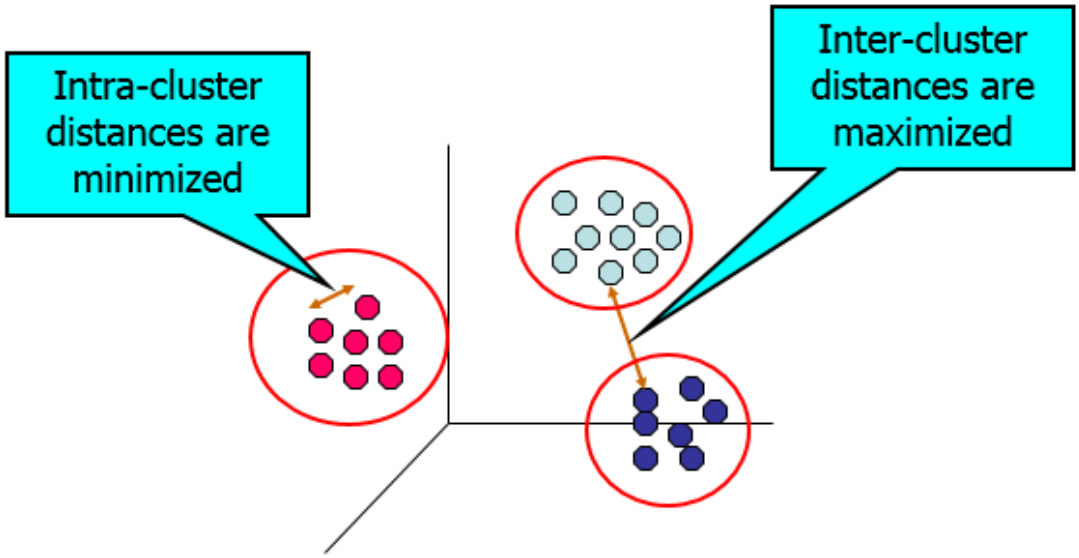
CLUSTER ANALYSIS

للعام الدراسي 2020-2019

تحليل العنقدة Cluster Analysis

ماذا نعني بتحليل العنقدة؟

- لدينا مجموعة من النقاط ونريد بطريقة أتوماتيكية أن نوجد مجموعة من النقاط التي ستكون متشابهة ((أو ذات علاقة مرتبطة)) هذه النقاط تشترك أو تتشابه بخواص معينة، من واحد لآخر ومختلفة عن ((أو غير مرتبطة)) النقاط في المجموعات الأخرى
- العنقدة هي مجموعة من العناقيد.



Inter–cluster: تقيس المسافات الخارجية (أي المسافة بين المجموعات) وتكون المسافات كبيرة.

Intra–cluster: تقيس المسافات الداخلية (أي داخل المجموعة) وتكون المسافة صغيرة.

فوائد العنقدة:

١- استرجاع أو استرداد المعلومات: ((أي تفيد في محركات البحث))

- تجميع نتائج البحث إلى عدد أصغر من العناقيد (كل منها تأخذ سمة معينة من البحث)

- تجميع صـــفحات الويب في فئات (عناقيد) وكل فئة يمكن تقســيمها (أو تجزيئها) إلى فئات جزئية أو ثانوية (عناقيد جزئية أو ثانوية) تنتج تركيباً متسلسلاً هرمياً (ترتيبي).

٢- التحليل النفسي والطبي: ((أي تفيد في تشخيص الأمراض-معرفة الحالة المرضية))

- كثيراً ما يكون للمرض أو لحالة معينة عدد من الاختلافات يمكن اســـتخدام العنقدة لتميز العناقيد الجزئية (الفئات) مختلفة
- تستخدم العنقدة لاكتشاف أنواع مختلفة منخفضة

٣- الأعمال: ((أي تفيد في تحليل سلة التسوق وكشف الغش))

- يمكن أن تستخدم العنقدة لتقسيم الزبائن إلى عدد صغير من المجموعات من أجل أن نحلل بشكل إضافي وفعاليات البيع (أي النشاط التجاري).

أنواع العنقدة:

١- عنقدة منفصلة

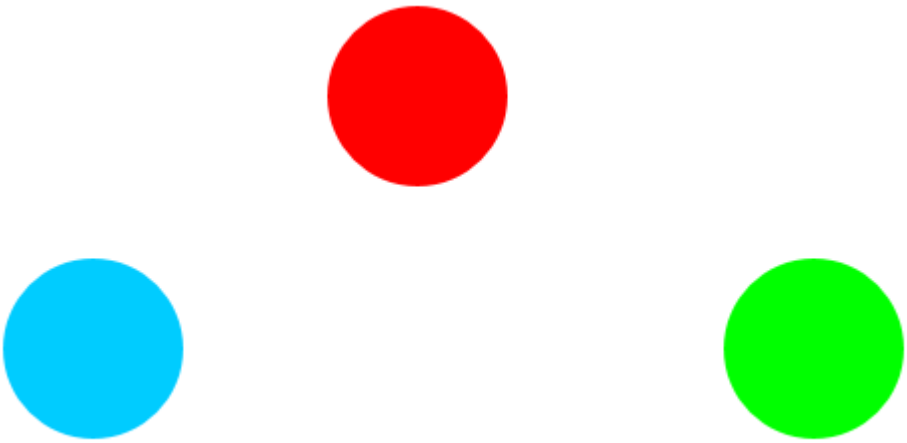
٢- عنقدة مبينة على المركز

٣- عنقدة متلاصقة

٤- عنقدة مبنية على الكثافة

- **العنقدة المنفصلة:**

مجموعة من النقاط بحيث أن أي نقطة في العنقود هي أقرب لكل نقطة أخرى في العنقود من أي نقطة ليست في العنقود.



3 well-separated clusters

- **العنقدة المبنية على المركز:**

العنقدة هي مجموعة النقاط بحيث أن أي نقطة في العنقود هي أقرب إلى مركز العنقود من أي مركز عنقود آخر.

- مركز العنقود غالبا هو متوسط كل نقاط العنقود الواحد

4 center-based clusters

- **العنقدة المتلاصقة (عناقيد متجاورة أو متلاصقة):**

عنقود متلاصق (طريقة الجار الأقرب) ((ينظر إلى أقرب نقطة له))

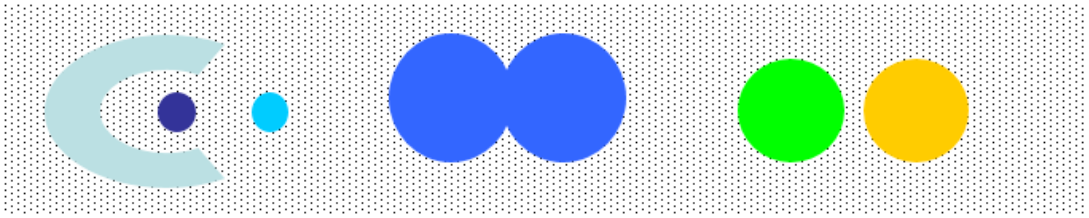
العنقدة هو مجموعة النقاط بحيث أن أي نقطة في العنقود هي أقرب (أو أكثر تشـابه) لنقطة واحدة أو أكثر في العنقود من النقاط الأخرى من أي نقطة ليست في العنقود.

8 contiguous clusters

- **العنقدة المبنية على الكثافة:**

العنقدة هي منطقة كثيفة من النقاط حيث المناطق ذات الكثافة المنخفضـة مفصـولة عن المناطق ذات الكثافة العالية.

يستخدم عندما تكون عناقيد شاذة وعندما تكون العناقيد الحالية يوجد فيها ضجيج وتكون هامشية (شاذة)



6 density-based clusters

طرائق العنقدة:

١- العنقدة الكلاسيكية

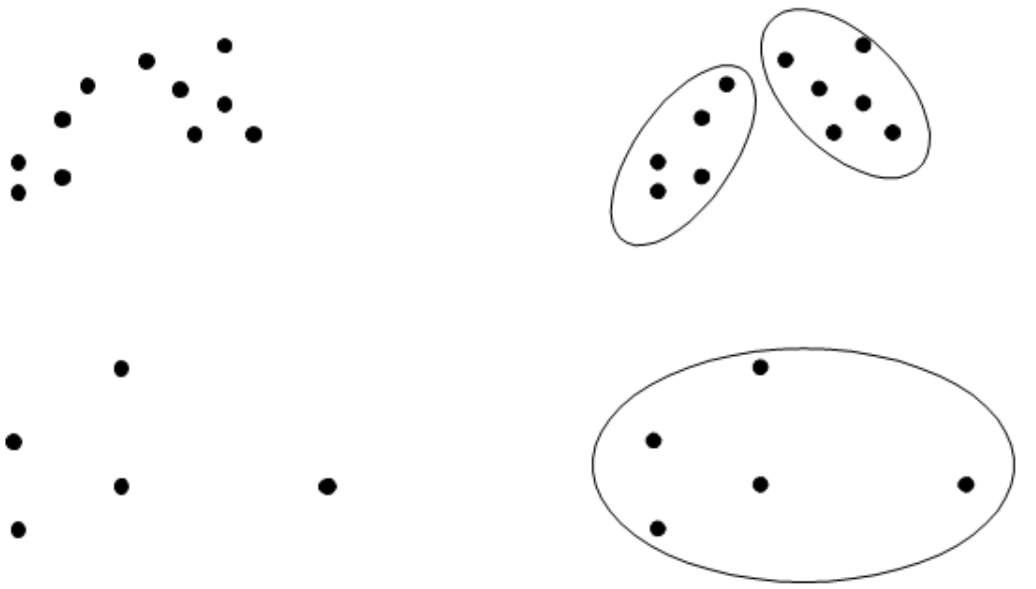
٢- العنقدة العصبية.

- **العنقدة الكلاسيكية:**

١- العنقدة التجزيئية:

تقســـيم نقاط البيانات إلى مجموعات جزئية (عناقيد) غير متقاطعة فيما بينها بحيث أن كل نقطة من النقاط المعطيات هي تماماً مجموعة جزئية واحدة.

Partitional Clustering

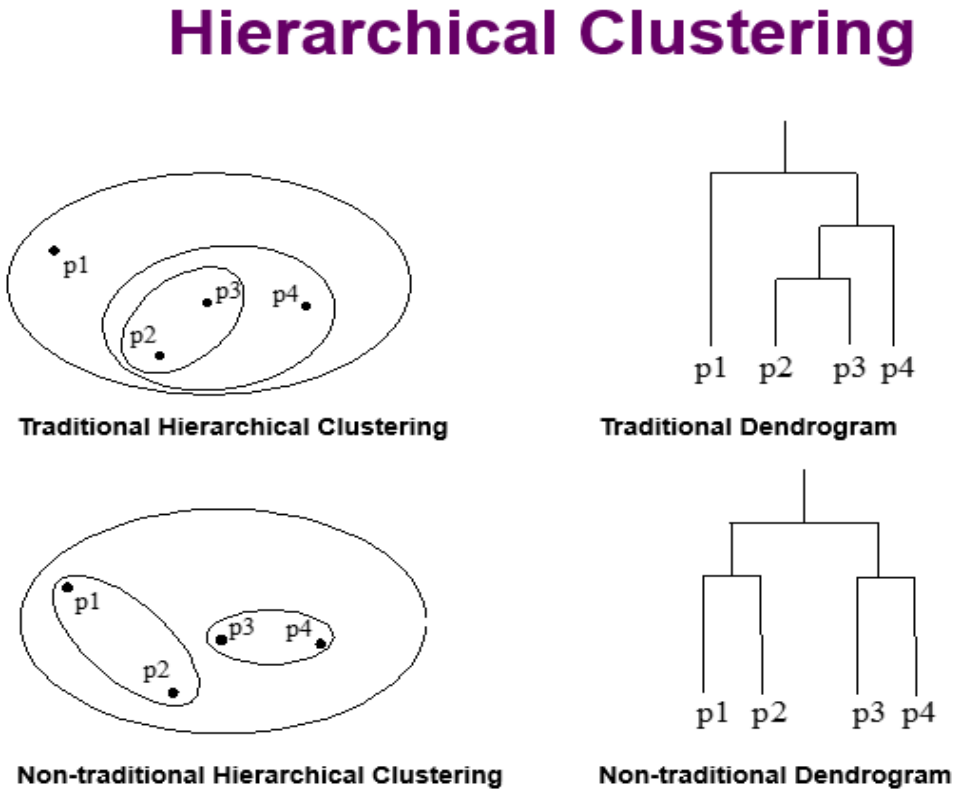


Original Points

A Partitional Clustering

٢- العنقدة الهرمية:

يتشكل لدينا شبكة من العناقيد منظمة بشكل شبكة هرمية.



٣- العنقدة المبنية على الكثافة:

العنقود منطقة كثيفة من النقاط أي أن كل عنقود يمثل منطقة كثيفة من أجل أن تفصل المناطق الأقل كثافة عن المناطق الأكثر كثافة تستخدم عندما تكون العناقيد شاذة أو متشابكة وتساعدنا في تحديد الكثافة الهامشية.

- **خوارزميات العنقدة التجزيئية:**

عنقدة K-means:

- معالجة العنقدة التجزيئية.
- كل عنقود مرتبط بنقطة المركز.

- كل نقطة يجب إلحاقها بعنقود معين (الأقرب لنقطة المركز).
- K عدد العناقيد يجب أن تكون محددة (من قبل المستخدم).

قاعدة الخوارزمية:

١- نختار K نقطة تعتبر مراكز ابتدائية.

٢- كرر.

٣- (نفتح حلقة) من K عنقود مخصصة لكل النقاط لأقرب مركز

٤- نعيد حساب المركز لكل عنقود.

٥- (حتى) لا تتغير المراكز _أي نتوقف عندما يكون المركز نفسه_

<u>تمرين</u>

بفرض أنه لدينا النقاط الست التالية:

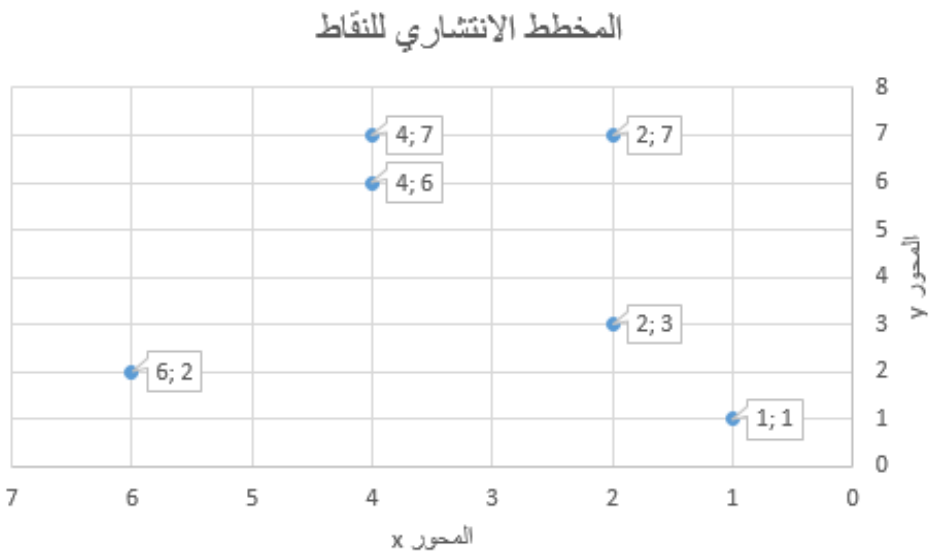
$$p_1(1,1), p_2(2,3), p_3(6,2), p_4(4,6), p_5(4,7), p_6(2,7)$$

والمطلوب: طبق طريقة K_means التجزيئية مستخدماً تسميات

الصفوف $C_1, C_2, \ldots$ باستخدام مسافة منهاتن.

x	y
1	1
2	3
6	2
4	6
4	7
2	7

الشكل الانتشاري للنقاط:



- **أولاً:** نختار K نقطة تعتبر مراكز ابتدائية ولتكن p_1 و p_6 أي أن $K = 2$ أي أننا اخترنا العنقودين $C_1(1,1), C_2(2,7)$ نحسب المسافات (أو الأبعاد) بين C_1, C_2 والنقاط الست فيتشكل لدينا حيث نحسب المسافة من خلال العلاقة:

$$dist = \sum_{k=1}^{n} |p_k - q_k|$$

p_k, q_k: احداثيات النقاط لـ p,q

n: عدد الأبعاد وهنا تساوي 2

$$d(p_1, C_1) = 0, d(C_1, p_2) = |1-2| + |1-3| = 3$$
$$d(C_2, p_4) = |2-4| + |7-6| = 3$$

dist	C_1	C_2
p_1	0	7
p_2	3	4
p_3	6	9
p_4	8	3
p_5	9	2
p_6	7	0

وبالتالي نختار النقاط الأقرب إلى نقطة المركز C_1 ((أي المسافة بين النقاط و C_1 أقل من المسافة بين النقاط و C_2)) ونضعها في عنقود جديد وكذلك الأمر بالنسبة لـــ C_2 فنحصل على:

$$CL_1 = \{p_1, p_2, p_3\} \quad , CL_2 = \{p_4, p_5, p_6\}$$

- **ثانياً**: نعيد حساب المركز لكل عنقود:

- العنقود الأول يحوي ثلاث نقاط فمركزه هو:

$$C_1\left(\frac{x_{p_1}+x_{p_2}+x_{p_3}}{3},\frac{y_{p_1}+y_{p_2}+y_{p_3}}{3}\right)=C_1\left(\frac{1+2+6}{3},\frac{1+3+2}{3}\right)$$
$$C_1(3,2)$$

- العنقود الثاني يحوي ثلاث نقاط فمركزه هو:

$$C_2(\frac{x_{p_4}+x_{p_5}+x_{p_6}}{3},\frac{y_{p_4}+y_{p_5}+y_{p_6}}{3})=C_2\left(\frac{4+4+2}{3},\frac{6+7+7}{3}\right)$$
$$C_1(3.3,6.7)$$

- نحسب المسافات بين المراكز الجديدة وبين النقاط فنحصل على:

نلاحظ أن المسافة الأقل بين المركز الجديد C_1 والنقاط p_1,p_2,p_3

وكذلك المسافة الأقل بين المركز الجديد C_2 والنقاط p_4,p_5,p_6

وبالتالي نحصل على العنقودين:

$$CL_3=\{p_1,p_2,p_3\} \quad ,CL_4=\{p_4,p_5,p_6\}$$

- الآن من جديد نحسب المراكز لكل عنقود فنحصل على نفس المراكز في الخطوة السابقة وبالتالي نتوقف لأن المراكز الجديدة $C_1(3,2)$ و $C_2(3.3,6.7)$ قد كررت.

dist	C_1	C_2
p_1	3	8
p_2	2	5
p_3	3	7.4
p_4	5	1.4
p_5	6	1
p_6	6	1.6

نلخص الخطوات بالشكل:

Step1	
$C_1(1,1)$	$C_2(2,7)$
$\{p_1,p_2,p_3\}$	$\{p_4,p_5,p_6\}$
Step2	
$C_1(3,2)$	$C_2(3.3,6.7)$
$\{p_1,p_2,p_3\}$	$\{p_4,p_5,p_6\}$

Step3	
$C_1(3,2)$	$C_2(3.3,6.7)$
$\{p_1, p_2, p_3\}$	$\{p_4, p_5, p_6\}$
نتوقف	

- نلاحظ أن الحل جيد لأن انتقاء المراكز كان جيداً.

تقييم عنقدة الـ K_means:

المقياس الأكثر شيوعاً هو مجموع مربعات الخطأ (SSE)

- من اجل أي نقطة، الخطأ هو المسافة الأقرب عنقود
- نحصل على SSE من خلال تربيع هذه الأخطاء وأخذ مجموعها:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(C_i, x)$$

$$C_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

- حيث x نقاط المعطيات في العنقود C_i و m_i عدد النقاط في العنقود C_i
- يمكن أن نثبت أن C_i يتقارب الى المركز او المتوسط للعنقود
- اذا كان لدينا عنقودين ونريد الاختيار بينهما فإننا نختار العنقود الأقل خطأً.
- الطريقة الأفضل لتقليل الخطأ هو أن نزيد عدد العناقيد.
- العنقدة الأفضـــل عندما يكون عدد العناقيد قليل ونحصـــل على خطأ SSE أقل أما العنقدة السيئة عندما يكون عدد العناقيد كبير.

ما قبل المعالجة وما بعد المعالجة

ما قبل المعالجة:

١- تطبيع المعطيات (جعل المعطيات طبيعية)

٢- نحذف النقاط الشاذة

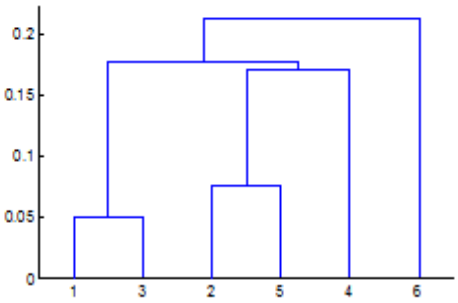
ما بعد المعالجة:

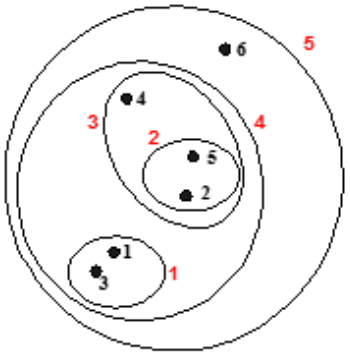
١- نزيل العناقيد الصغيرة التي قد تمثل نقاط شاذة

٢- دمج العناقيد القريبة والتي لها مستوى الخطأ منخفض نسبياً.

العنقدة الهرمية

- هي انتاج مجموعة من العناقيد المتداخلة على شكل شبكة هرمية.
- يمكن تمثيلها على شكل مخطط تفرعي.





مزايا العنقدة الهرمية

- يجب علينا ألا نفترض عدد محدد من العناقيد
- إن أي عدد مطلوب من العناقيد يمكن أن نحصل عليه من خلال قطع المخطط التفرعي في المستوى المناسب.

هناك نوعان رئيسيان من العنقدة الهرمية:

التداخل التجميعي((طريقة هرمية صاعدة)):

- نبدأ بالنقاط ونعتبرها عناقيد فردية.
- في كل خطوة ندمج الزوج الأقرب من العناقيد (أي ندمج أقرب عنصرين) حتى نحصل على عنقود واحد.

التحليل التقسيمي ((طريقة هرمية هابطة)):

- نبدأ بالعنقود الوحيد وهو العنقود الشامل.
- في كل خطوة، نقسم العنقود حتى يصبح لدينا كل عنقود يحتوي على نقطة (أو هناك k عنقود) حتى نعتبر كل نقطة عنقود.
- تستخدم خوارزميات العنقدة الهرمية التشابه أو مصفوفة المسافة.
- نجزء أو ندمج العنقود في كل مرة.

خوارزمية العنقدة الهرمية التجميعية

- هي أكثر العنقدة الهرمية شيوعاً لإنجاز غرض معين.
- الخوارزمية الأساسية هي بشكل مباشر:

١- نحسب مصفوفة التقارب.

٢- يكون كل نقطة معطيات عنقود

٣- كرر

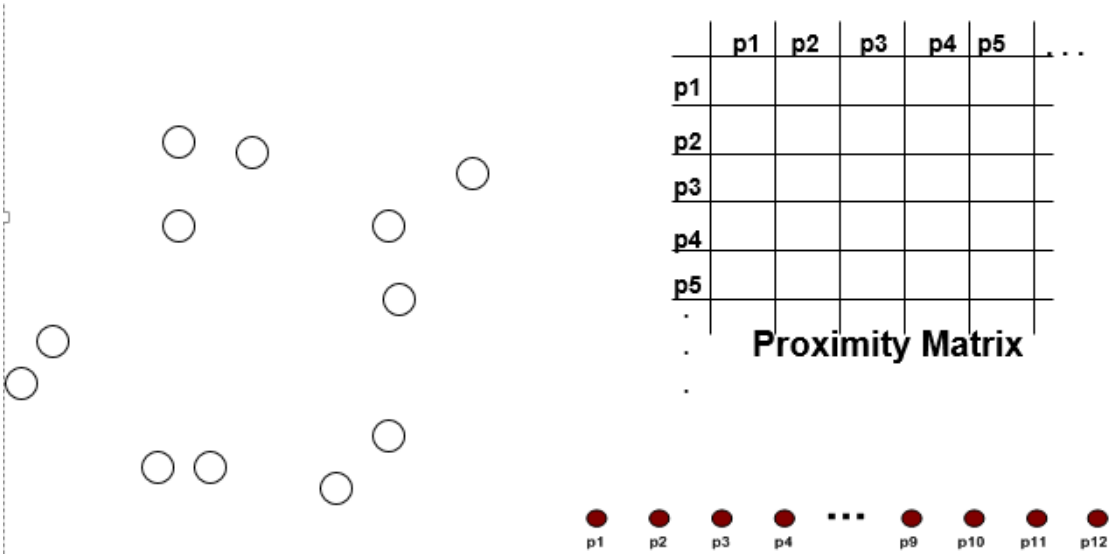
٤- ندمج أقرب عنقودين

٥- نحدث مصفوفة التقارب (أي نعيد الحساب للمسافات أو التشابهات)

٦- حتى يبقى لدينا عنقود وحيد.

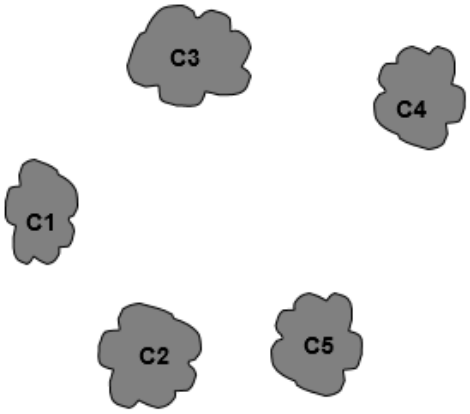
- **العملية الرئيسة هي حساب التقارب لعنقودين.**

- هناك طرق مختلفة لتحديد المسافة بين العناقيد لذلك نصنف أو نميز خوارزميات مختلفة.
- **حالة البدء:** نبدأ بعناقيد فردية ومصفوفة التقارب



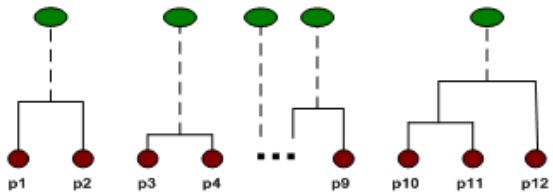
- **الحالة المتوسطة:**

- بعد أن قمنا ببعض خطوات الدمج للنقاط الفردية أصبح لدينا بعض العناقيد
- ندمج العناقيد المتقاربة ونحسب مصفوفة التقارب

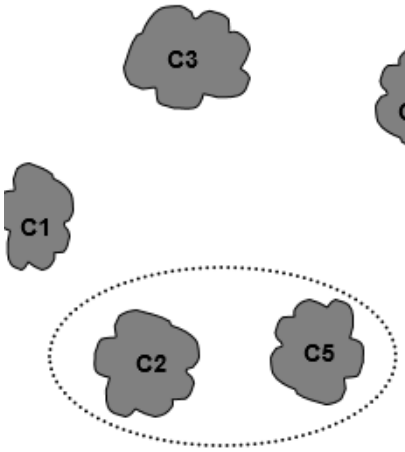


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

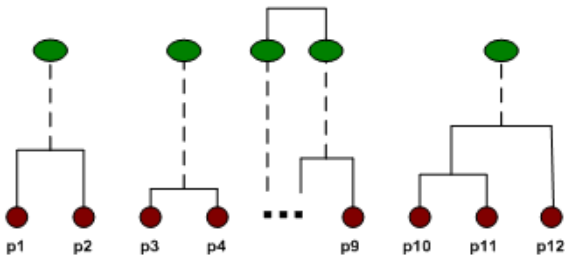


- نريد الآن أن ندمج العنقودين الأقرب (C_1, C_2) ومن ثم نحدث مصفوفة التقارب.



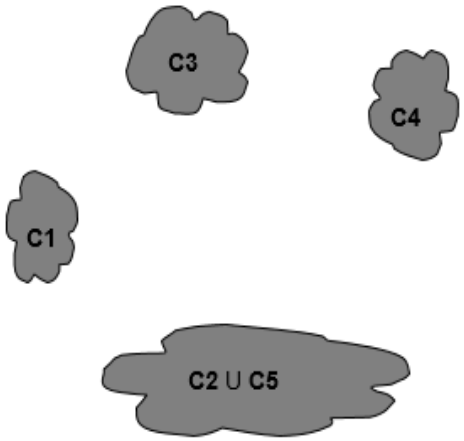
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



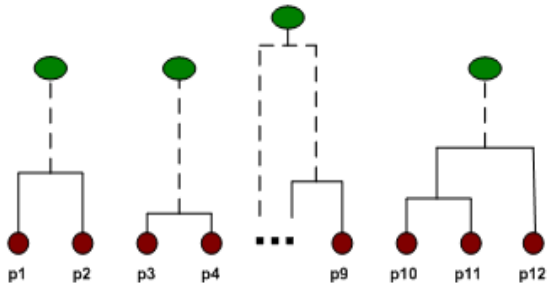
بعد الدمج:

- السؤال كيف نحدث مصفوفة التقارب:



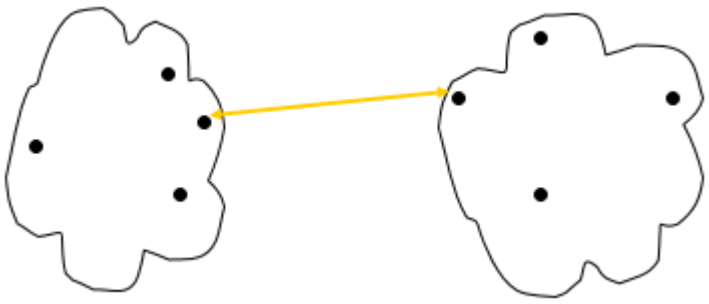
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

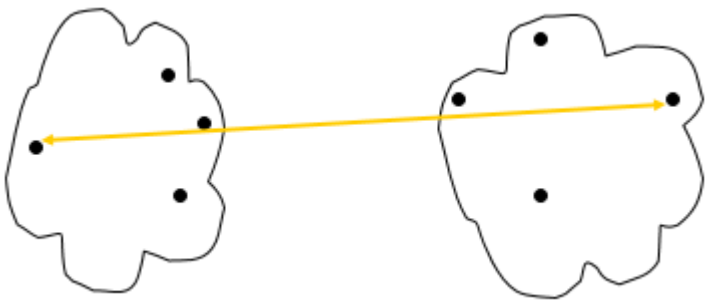


كيف نعرف التشابه بين العناقيد؟

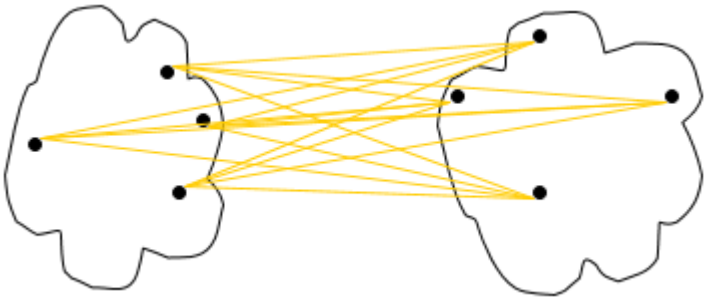
- الـ MIN (single link)



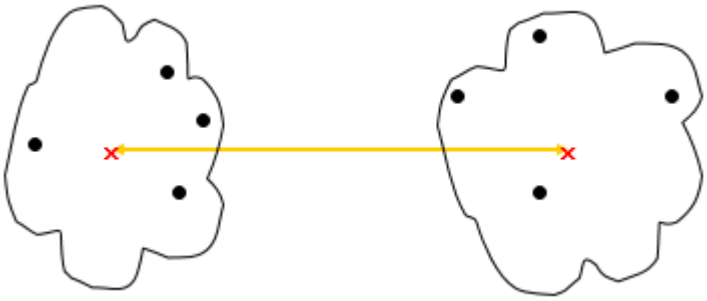
- الـMAX (complete link)



- **متوسط المجموعات.**



- **المسافة بين نقاط المراكز.**



تمرين عن التشابه العنقودي Min:

بفرض أنه لدينا:

point	x	y
p_1	0.4	0.53
p_2	0.22	0.38
p_3	0.35	0.32
p_4	0.26	0.19
p_5	0.08	0.41
p_6	0.45	0.3

المطلوب:

طبق طريقة MIN (single link) الهرمية مستخدماً تسميات الصفوف $C_1, C_2, \ldots$

وباستخدام المسافة الاقليدية.

dist	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0					
p_2	0.23	0				
p_3	0.22	0.14	0			
p_4	0.37	0.2	0.16	0		
p_5	0.034	0.14	0.28	0.29	0	
p_6	0.24	0.24	0.11	0.22	0.39	0

الآن نبحث عن أقل قيمة في مصفوفة المسافة وهي 0.11 وهي المسافة بين النقطتين الثالثة والسادسة ندمجها في عنقود واحد أي أن $C_1 = \{p_3, p_6\}$ ومن ثم نعيد حساب المصفوفة من جديد.

$$dist(p_1, C_1) = \min\{dist(p_1, p_3), dist(p_1, p_6)\} = \min(0.22, 0.24) = 0.22$$

وتكون مصفوفة التقارب المحدثة:

dist	p_1	p_2	C_1	p_4	p_5
p_1	0				
p_2	0.23	0			
C_1	0.22	0.14	0		
p_4	0.37	0.2	0.16	0	
p_5	0.034	0.14	0.28	0.29	0

والآن نبحث عن أقل قيمة وهي 0.14 وتكررت مرتين نفضل النقطة التي ليس لها عنقود أي نأخذ p_2 مع p_5 فيتشكل لدينا العنقود الثاني $C_2 = \{p_2, p_5\}$

وتصبح مصفوفة التقارب بالشكل التالي:

$$dist(C_2, C_1) = \min\{dist(C_1, p_2), dist(C_1, p_5)\} = \min(0.14, 0.28) = 0.14$$

وذلك بالاعتماد على الجدول السابق:

dist	p_1	C_2	C_1	p_4
p_1	0			
C_2	0.23	0		
C_1	0.22	0.14	0	
p_4	0.37	0.2	0.16	0

والان نختار أقل قيمة وهي 0.14 وهي المسافة بين العنقودين C_1 و C_2 فيتشكل لدينا عنقود جديد $C_3 = \{C_1, C_2\}$

ونحسب مصفوفة التقارب من جديد(مصفوفة المسافة):

dist	p_1	C_3	p_4
p_1	0		
C_3	0.22	0	
p_4	0.37	0.16	0

نأخذ أقل مسافة التي هي 0.16 بين C_3 و p_4 فيتشكل لدينا العنقود الرابع وهو:

$$C_4 = \{p_4, C_3\}$$

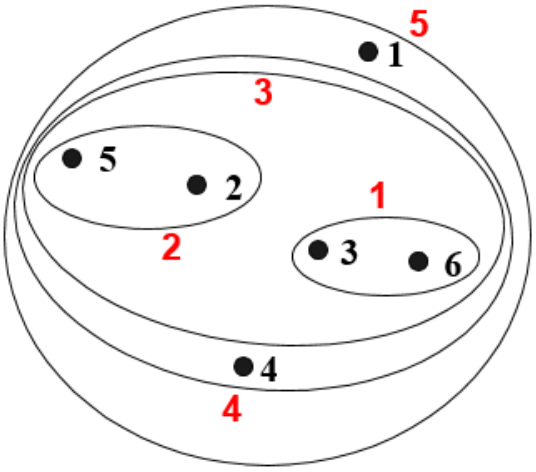
ومصفوفة التقارب:

dist	p_1	C_4
p_1	0	0.22
C_4	0.22	0

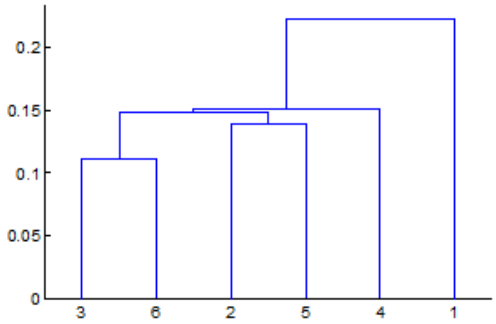
ويتشكل لدينا أخيراً آخر عنقود وهو:

$$C_5 = \{p_1, C_4\}$$

ويمكن أن نصيغ الطريقة السابقة على شكل مخطط هرمي أو رسم بياني:



Nested Clusters



Dendrogram

- **التشابه العنقودي Max أو الرابط التام(.MAX (complete link)):**
- بالعودة للتمرين السابق طبق طريقة الـMAX (complete link) الهرمية على النقاط الست مستخدماً المسافة الاقليدية.
- بالعودة إلى مصفوفة المسافة نأخذ أصغر قيمة وهي 0.11 من النقطتين p_3, p_6 فيتشكل لدينا العمود الأول:

$$C_1 = \{p_3, p_6\}$$

والآن نحسب المسافات من جديد:

$$dist(p_1, C_1) = \max\{dist(p_1, p_3), dist(p_1, p_6)\} = \max(0.22,0.24) = 0.24$$
$$dist(p_2, C_1) = \max\{dist(p_2, p_3), dist(p_2, p_6)\} = \max(0.14,0.24) = 0.24$$

وتكون مصفوفة التقارب المحدثة:

dist	p_1	p_2	C_1	p_4	p_5
p_1	0				
p_2	0.23	0			
C_1	0.24	0.24	0		
p_4	0.37	0.2	0.22	0	
p_5	0.034	0.14	0.39	0.29	0

والآن نبحث عن أقل قيمة وهي 0.14 فنأخذ p_2 مع p_5 فيتشكل لدينا العنقود الثاني

$$C_2 = \{p_2, p_5\}$$

وتصبح مصفوفة التقارب بالشكل التالي:

$$dist(C_2, C_1) = \max\{dist(C_1, p_2), dist(C_1, p_5)\} = \max(0.24,0.39) = 0.39$$

وذلك بالاعتماد على الجدول السابق:

dist	p_1	C_2	C_1	p_4
p_1	0			
C_2	0.34	0		
C_1	0.24	0.39	0	
p_4	0.37	0.28	0.22	0

والان نختار أقل قيمة وهي 0.22 وهي المسافة بين العنقودين C_1 و p_4 فيتشكل لدينا عنقود جديد $C_3 = \{C_1, p_4\}$ وتكون مصفوفة التقارب المحدثة:

$$dist(p_1, C_3) = \max\{dist(p_1, C_1), dist(p_1, p_4)\} = \max(0.24,0.37) = 0.37$$

$$dist(C_2, C_3) = \max\{dist(C_2, C_1), dist(C_2, p_4)\} = \max(0.39,0.28) = 0.39$$

dist	p_1	C_2	C_3
p_1	0		
C_2	0.34	0	
C_3	0.37	0.39	0

نأخذ أقل مسافة التي هي 0.34 بين C_2 و p_1 فيتشكل لدينا العنقود الرابع وهو:

$$C_4 = \{p_1, C_2\}$$

ومصفوفة التقارب:

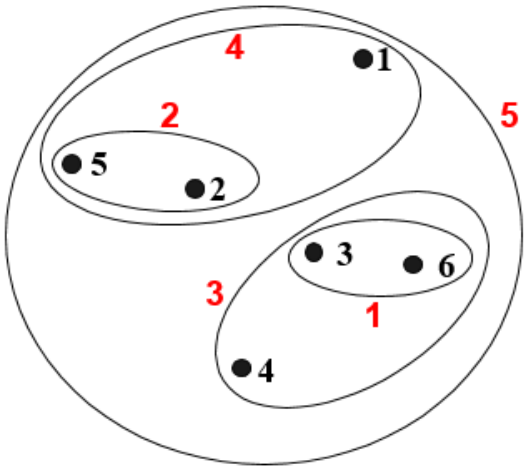
$$dist(C_3, C_4) = \max\{dist(C_3, p_1), dist(C_3, C_2)\} = \max(0.39,0.37) = 0.39$$

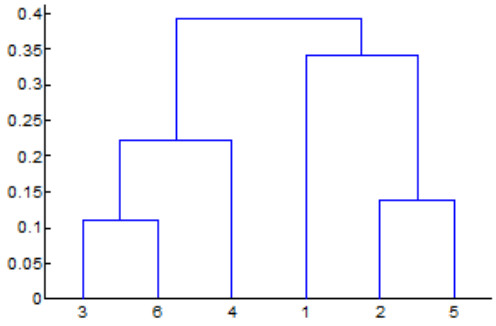
dist	C_4	C_3
C_4	0	0.39
C_3	0.39	0

ويتشكل لدينا أخيراً آخر عنقود وهو:

$$C_5 = \{C_4, C_3\}$$

ويمكن أن نصيغ الطريقة السابقة على شكل مخطط هرمي أو رسم بياني:





- **التشابه العنقودي (متوسط المجموعات):**

التقارب العنقودي يكون متوسط تقارب زوج بين النقاط في العنقودين:

ويكون حسابه بالعلاقة التالية:

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

تمرين:

بالعودة للمثال السابق

- طبق متوسط المجموعات الهرمية بين النقاط باستخدام المسافة الإقليدية.

الحل:

- أولا نوجد مصفوفة المسافات او التقارب المحسوبة في المثال السابق

بالعودة إلى مصفوفة المسافة نأخذ أصغر قيمة وهي 0.11 من النقطتين p_3, p_6 فيتشكل لدينا العمود الأول:

$$C_1 = \{p_3, p_6\}$$

لحساب المسافات نطبق الدستور:

$$d(Cli, Clj) = \frac{\sum_{i,j} d(p_i, p_j)}{|Cli| * |Clj|}$$

يسمى $|Cli|$ الكاردنيك الذي يعبر عن عدد النقاط الموجودة ضمن العنقود.

نوجد الآن مصفوفة المسافة:

$$d(p_1, C_1) = \frac{d(p_1, p_3) + d(p_1, p_6)}{(1) * (2)} = \frac{0.22 + 0.24}{2} = 0.23$$

$$d(p_2, C_1) = \frac{d(p_2, p_3) + d(p_2, p_6)}{(2) * (1)} = \frac{0.14 + 0.24}{2} = 0.19$$

$$d(p_4, C_1) = \frac{d(p_4, p_3) + d(p_4, p_6)}{2} = \frac{0.14 + 0.24}{2} = 0.19$$

$$d(p_5, C_1) = \frac{d(p_5, p_3) + d(p_5, p_6)}{2} = \frac{0.28 + 0.39}{2} = 0.34$$

dist	p_1	p_2	C_1	p_4	p_5
p_1	0				
p_2	0.23	0			
C_1	0.23	0.19	0		
p_4	0.37	0.2	0.19	0	
p_5	0.034	0.14	0.34	0.29	0

والآن نبحث عن أقل قيمة وهي 0.14 فنأخذ p_2 مع p_5 فيتشكل لدينا العنقود الثاني

$$C_2 = \{p_2, p_5\}$$

وتصبح مصفوفة التقارب بالشكل التالي:

$$d(p_1, C_2) = \frac{d(p_1, p_2) + d(p_1, p_5)}{(2) * (1)} = \frac{0.23 + 0.34}{2} = 0.29$$

$$d(p_4, C_2) = \frac{d(p_4, p_2) + d(p_4, p_5)}{(2) * (1)} = \frac{0.20 + 0.29}{2} = 0.25$$

$$d(C_1, C_2) = \frac{d(p_3,p_2)+d(p_3,p_5)+d(p_6,p_2)+d(p_6,p_5)}{(2)*(2)} = \frac{0.14+0.24+0.28+0.39}{4} = 0.26$$

dist	p_1	C_2	C_1	p_4
p_1	0			
C_2	0.29	0		

C_1	0.37	0.26	0	
p_4	0.37	0.25	0.19	0

والآن نختار أقل قيمة وهي 0.19 وهي المسافة بين العنقودين C_1 و p_4 فيتشكل لدينا عنقود جديد $C_3 = \{C_1, p_4\}$ وتكون مصفوفة التقارب المحدثة:

$$d(C_2, C_3) = \frac{d(p_2,p_4)+d(p_2,p_3)+d(p_2,p_6)+d(p_5,p_4)+d(p_5,p_3)+d(p_5,p_6)}{(2)*(3)} = \frac{0.14+0.20+0.24+0.29+0.28+0.39}{6} = 0.26$$

$$d(p_1, C_3) = \frac{d(p_1,p_4)+d(p_1,p_3)+d(p_1,p_6)}{(1)*(3)} = \frac{0.37+0.22+0.24}{3} = 0.28$$

نأخذ أقل مسافة التي هي 0.26 بين C_2 و C_3 فيتشكل لدينا العنقود الرابع وهو:

$$C_4 = \{C_3, C_2\}$$

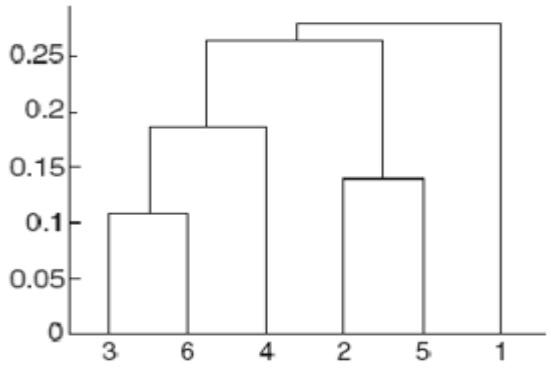
dist	p_1	C_4
p_1	0	0.28
C_4	0.28	0

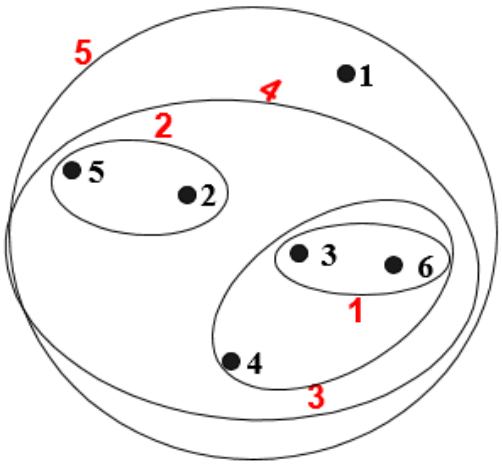
$$d(p_1, C_4) = \frac{d(p_1,p_2)+d(p_1,p_3)+d(p_1,p_4)+d(p_1,p_5)+d(p_1,p_6)}{(1)*(5)} = \frac{0.23+0.22+0.37+0.34+0.24}{5} = 0.28$$

ويتشكل العنقود الأخير:

$$C_5 = \{p_1, C_4\}$$

ويمكن أن نصيغ الطريقة السابقة على شكل مخطط هرمي أو رسم بياني:





مثال:

بفرض أنه لدينا النقاط الست التالية:

$$p_1(1,1), p_2(2,3), p_3(6,2), p_4(4,6), p_5(4,7), p_6(2,7)$$

والمطلوب:

طبق طريقة الـ(complete link) MAX الهرمية على النقاط الست مستخدماً تسميات الصفوف $C_1, C_2, \ldots$ باستخدام مسافة منهاتن.

باستخدام مسافة منهاتن تكون مصفوفة المسافة بالشكل التالي:

dist	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0					
p_2	3	0				
p_3	6	5	0			
p_4	8	5	6	0		
p_5	9	6	7	1	0	
p_6	7	4	9	3	2	0

نبحث عن أقل قيمة وهي 1 فنأخذ p_4 مع p_5 فيتشكل لدينا العنقود الأول $C_1 = \{p_4, p_5\}$

والآن نحسب المسافات:

$$dist(p_1, C_1) = \max\{dist(p_1, p_4), dist(p_1, p_5)\} = \max(8,9) = 9$$

$$dist(p_2, C_1) = \max\{dist(p_2, p_4), dist(p_2, p_5)\} = \max(5,6) = 6$$

وتكون مصفوفة التقارب المحدثة:

dist	p_1	p_2	p_3	C_1	p_6
p_1	0				
p_2	3	0			
p_3	6	5	0		
C_1	9	6	7	0	
p_6	7	4	9	3	0

والآن نبحث عن أقل قيمة وهي3 فنأخذ p_1 مع p_2 فيتشكل لدينا العنقود الثاني

$$C_2 = \{p_1, p_2\}$$

وتصبح مصفوفة المسافة بالشكل التالي:

$$dist(C_2, C_1) = \max\{dist(C_1, p_1), dist(C_1, p_2)\} = \max(9,6) = 9$$

وذلك بالاعتماد على الجدول السابق:

dist	C_2	p_3	C_1	p_6
C_2	0			
p_3	6	0		
C_1	9	7	0	
p_6	7	9	3	0

والان نختار أقل قيمة وهي 3 وهي المسافة بين C_1 و p_6 فيتشكل لدينا عنقود جديد

$C_3 = \{C_1, p_6\}$ وتكون مصفوفة التقارب المحدثة:

$$dist(C_2, C_3) = \max\{dist(p_1, C_1), dist(p_1, p_6)\} = \max(9,7) = 9$$

$$dist(p_3, C_3) = \max\{dist(p_3, C_1), dist(p_3, p_6)\} = \max(9,7) = 9$$

dist	p_1	C_2	C_3
C_2	0		
p_3	6	0	
C_3	9	9	0

نأخذ أقل مسافة التي هي 6 بين C_2 و p_3 فيتشكل لدينا العنقود الرابع وهو:

$$C_4 = \{p_3, C_2\}$$

ومصفوفة التقارب:

$$dist(C_3, C_4) = \max\{dist(C_3, p_3), dist(C_3, C_2)\} = \max(6,9) = 9$$

dist	C_4	C_3
C_4	0	9
C_3	9	0

ويتشكل لدينا أخيراً آخر عنقود وهو:

$$C_5 = \{C_4, C_3\}$$