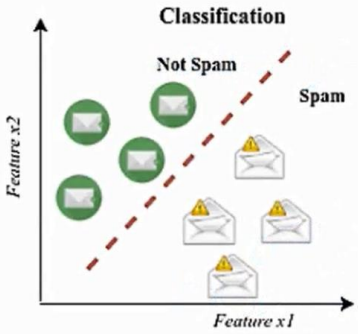
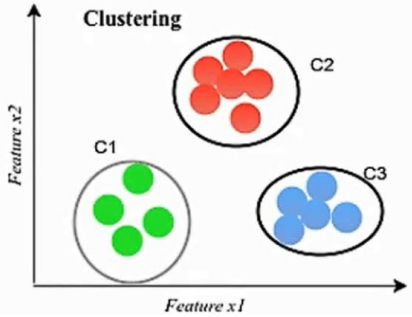
خوشه بندی

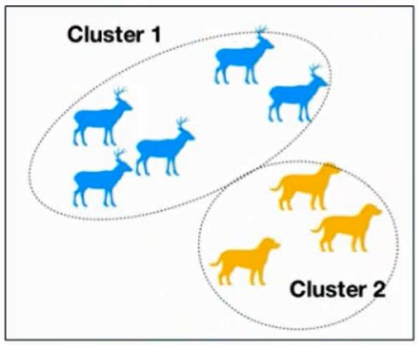
آموزش بدون ناظر

شیوه آموزش: بدون نظارت (unsupervised)





خوشه‌بندی



انتساب اشیاء به خوشه‌ها به‌طوری که اشیاء یک خوشه:

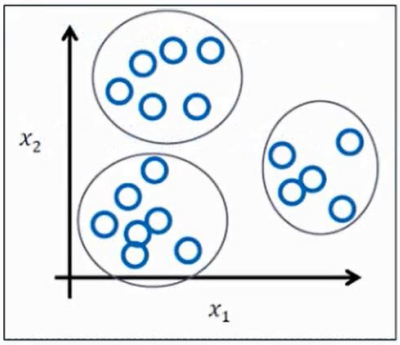
- بیشترین شباهت را با هم داشته باشند.
- بیشترین تفاوت بین خوشه‌های مختلف موجود باشد.

خوشه‌بندی

We have a set of unlabeled data points $\{x^{(i)}\}_{i=1}^{N}$ and we intend to **find groups of similar objects** (based on the observed features)

high intra-cluster similarity: cohesive within clusters

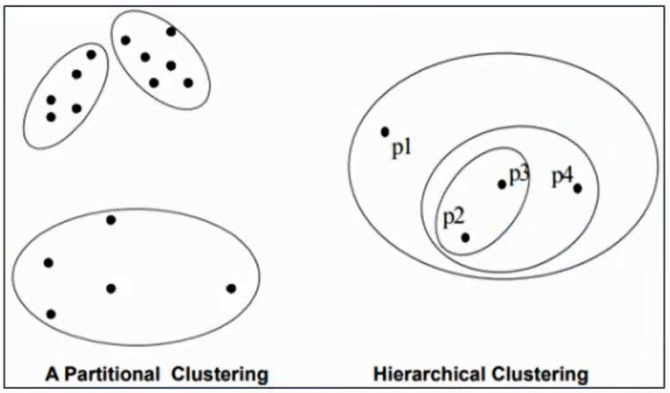
low inter-cluster similarity: distinctive between clusters



کاربردهای خوشه‌بندی

- بازیابی اطلاعات (کلاستر کردن اسناد متنی و تصاویر بر اساس محتویات آن‌ها)
- خوشه‌بندی کاربران شبکه‌های اجتماعی (community detection)
- بیوانفورماتیک (خوشه‌بندی ژن‌های مشابه بر اساس داده‌های میکروآرای)
- بازاریابی (Marketing)
- یافتن الگوهای هواشناسی
- بینایی رایانه‌ای (Computer Vision)
- ...

رویکردهای کلی الگوریتم‌های خوشه‌بندی



۱- پارتیشن‌بندی (Partitioning)

۲- سلسله مراتبی (Hierarchical)

خوشه‌بندی افرازی

Partitional Clustering

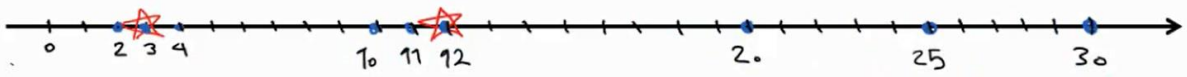
$$\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$$

$$\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K\}$$

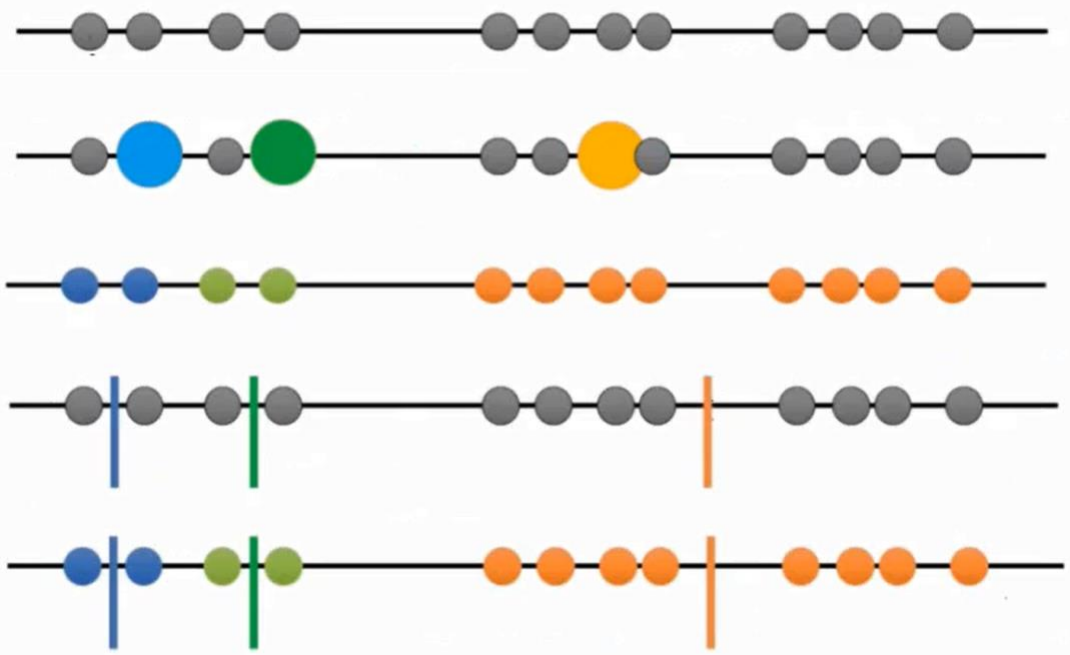
- $\forall j, \mathcal{C}_j \neq \emptyset$
- $\bigcup_{j=1}^{K} \mathcal{C}_j = \mathcal{X}$
- $\forall i, j,\ \mathcal{C}_i \cap \mathcal{C}_j = \emptyset$

مثال

$X = \{ 2 , 3 , 4 , 10 , 11 , 12 , 20 , 25 , 30 \}$

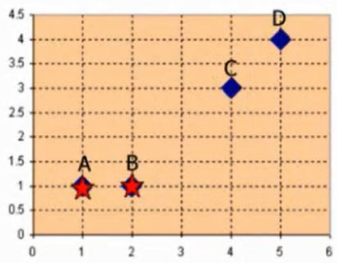


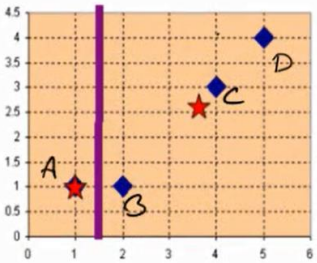
مثال

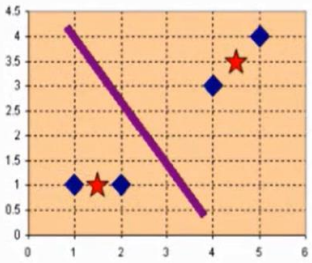


مثال

	X	Y
A	1	1
B	2	1
C	4	3
D	5	4

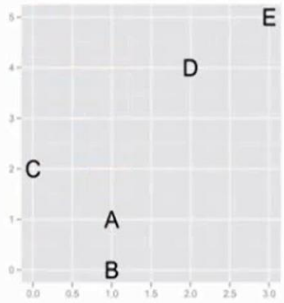


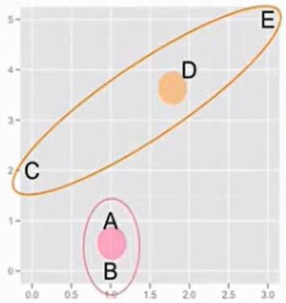


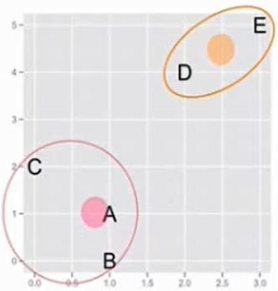


تمرین

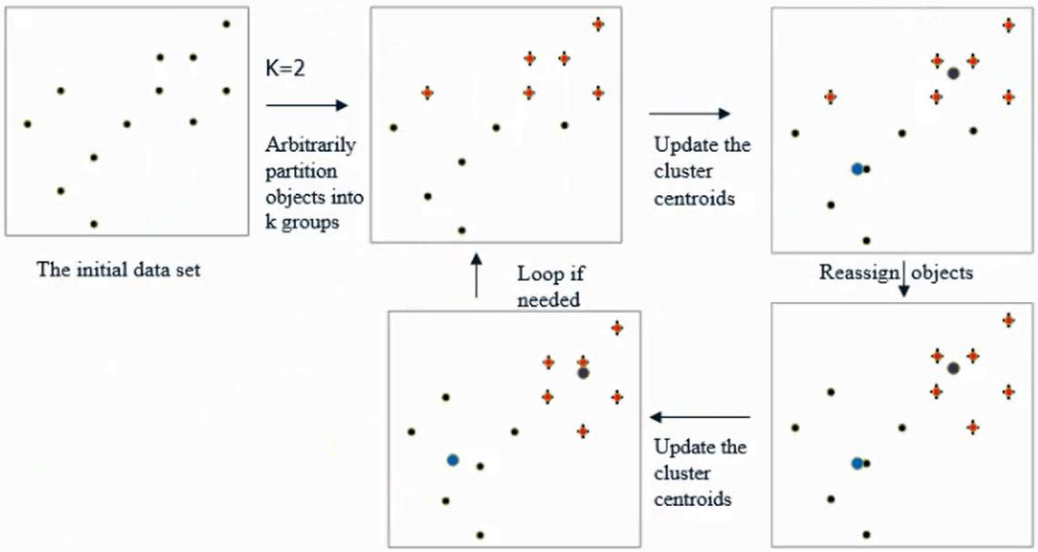
	X	Y
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5







K-Means



K-Means
(The Lloyd's method)

Select k random points $c_1, c_2, \ldots, c_k$ as cluster's initial centroids.

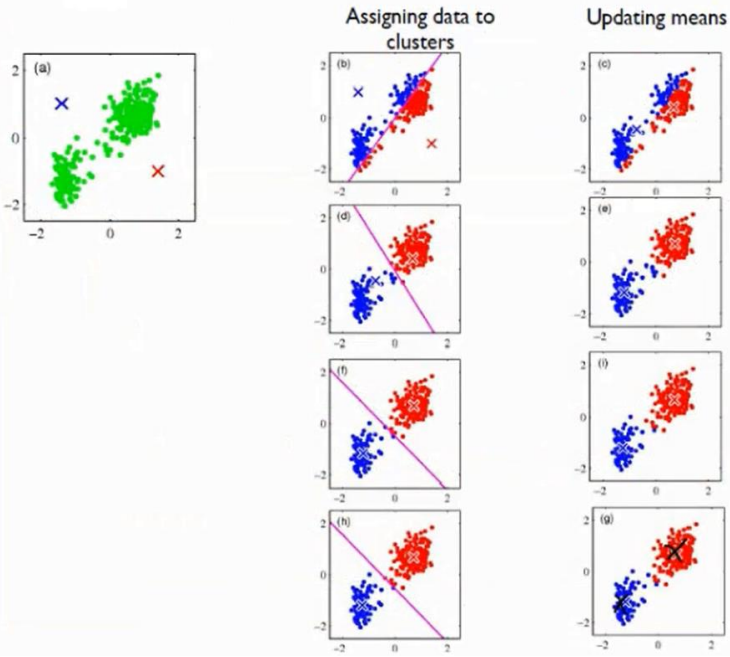
Repeat until converges (or other stopping criterion):

for i=1 to N do:

Assign $x^{(i)}$ to the closet cluster and thus C_j contains all data that are closer to c_j than to any other cluster

for j=1 to k do

$$c_j = \frac{1}{|c_j|}\sum_{x^{(i)} \in c_j} x^{(i)}$$



فرایند K-Means:

K-means Clustering

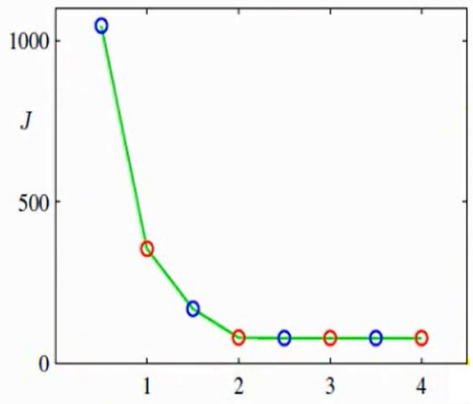
ورودی: a set $x^{(1)}, \ldots, x^{(N)}$ of data points and an integer K (in d-dim feature space)

خروجی: set of K representatives $c_1, c_2, \ldots, c_K \in \mathbb{R}^d$ as the cluster representatives

تابع هدف: choose $c_1, c_2, \ldots, c_K$ to minimize:

$$\sum_{i=1}^{N} \min_{j \in 1,\ldots,K} d^2(x^{(i)}, c_j)$$

همگرایی K-Means



- الگوریتم k-means همواره همگرا است.

در هر دو فاز مقدار تابع هزینه کم می‌شود.

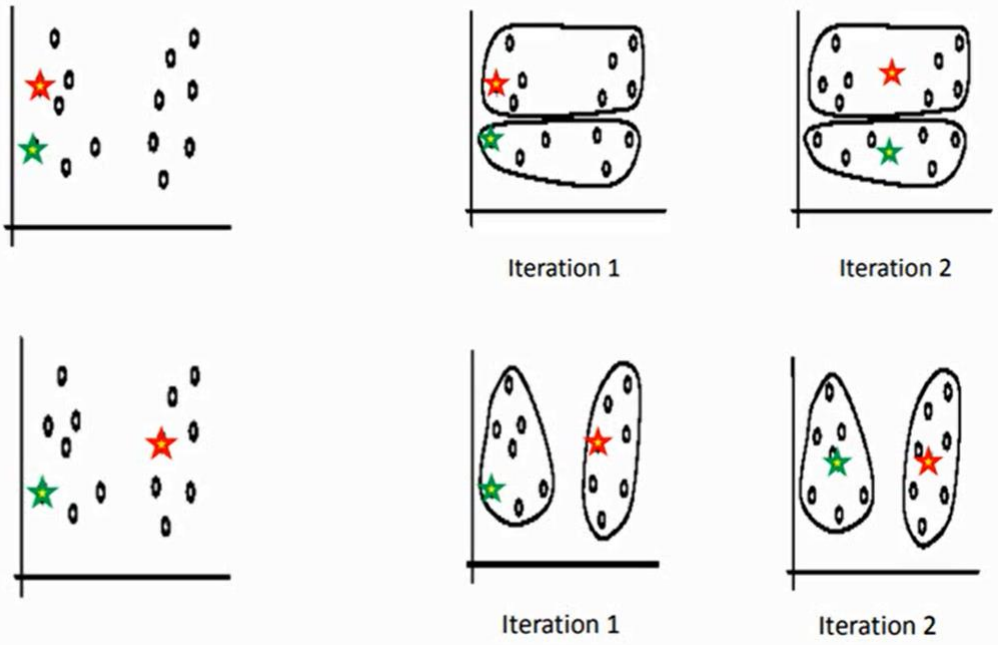
بهینه محلی

الگوریتم k-means ممکن است در بهینه محلی گیر بیفتد.

- ○
- ○
- ○
- ○

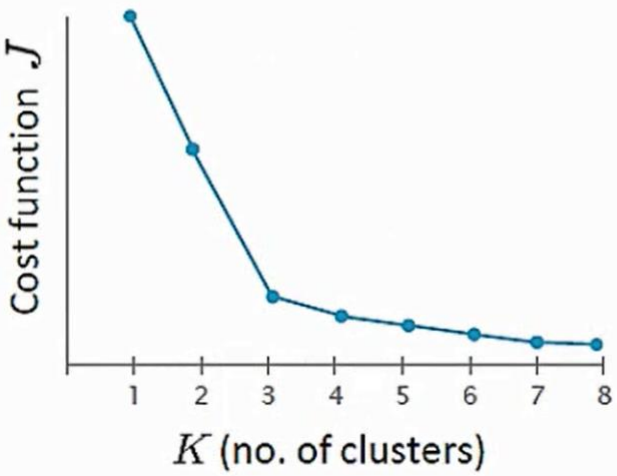
- ●
- ○
- ○
- ●

Sensitivity to initial seeds



انتخاب تعداد خوشه‌ها

elbow



نقاط قوت K-Means

- پیاده‌سازی ساده
- پیچیدگی زمانی الگوریتم: O(nkt)

نقاط ضعف K-Means

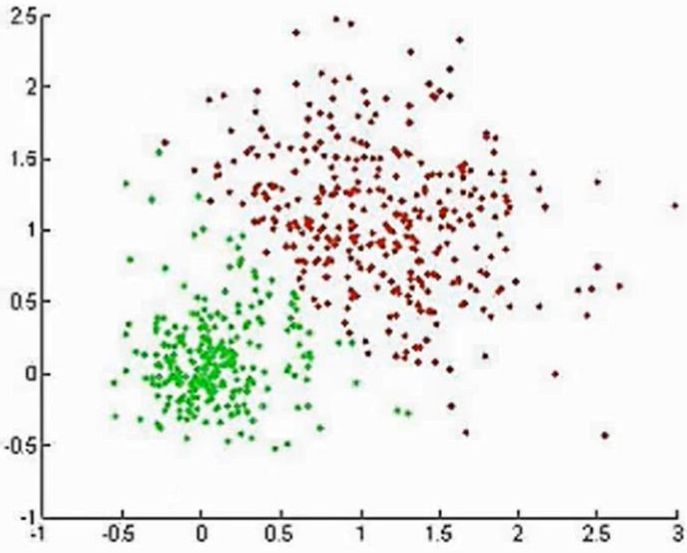
- مقدار k از قبل باید تنظیم شود.
- اغلب در یک بهینه محلی پایان می‌یابد.
- برای کشف کلاسترها با شکل‌های دلخواه مناسب نیست.
- برای داده‌های categorical کار نمی‌کند. (مانند ویژگی رنگ)
- نویز و داده‌های پرت می‌تواند مشکل قابل توجهی برای خوشه‌بندی باشد.
- انتخاب اولیه مرکز خوشه‌ها در نتیجه نهایی تاثیرگذار است.
- در مرحله‌ای از تکرار الگوریتم، ممکن است تعداد اعضای یک خوشه صفر شود.
- ترجیح می‌دهد خوشه‌ها تقریبا هم‌اندازه باشند.

حساس بودن به داده‌های پرت

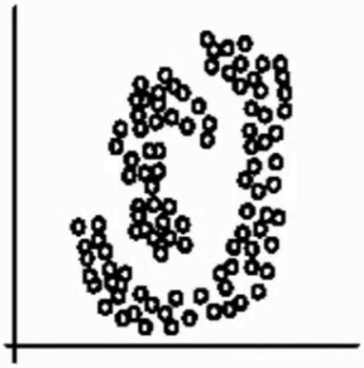
بدون outlier

با outlier

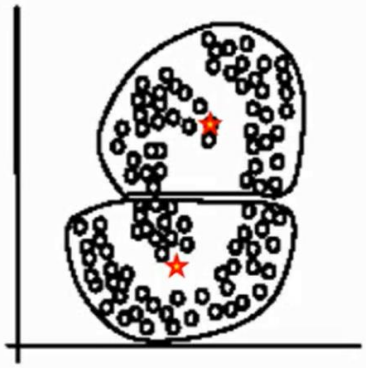
مثال



k-means is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).



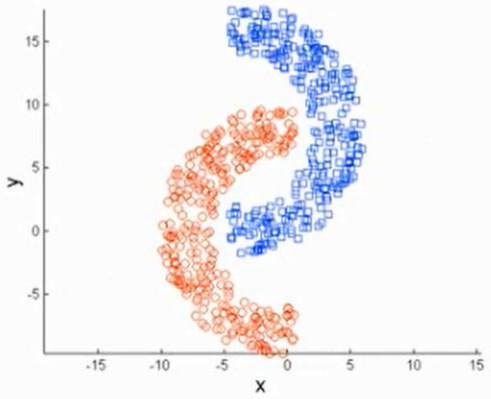
(A): Two natural cluster



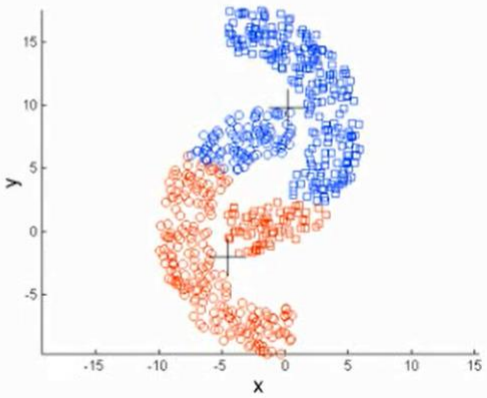
(B): k-means cluster

مثال

Non-globular Shapes

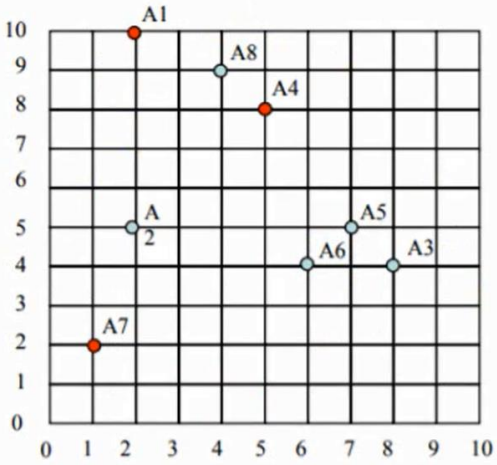


Original Points



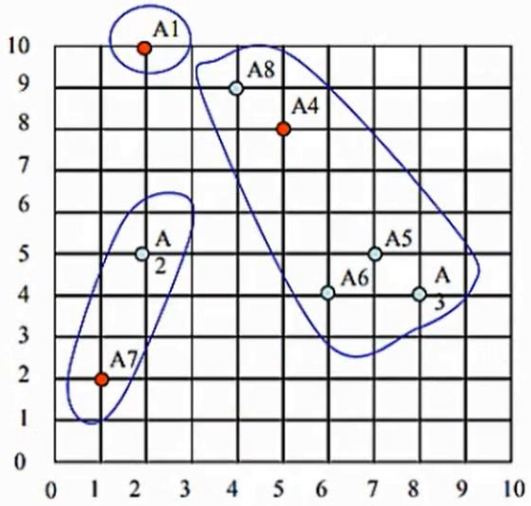
K-means (2 Clusters)

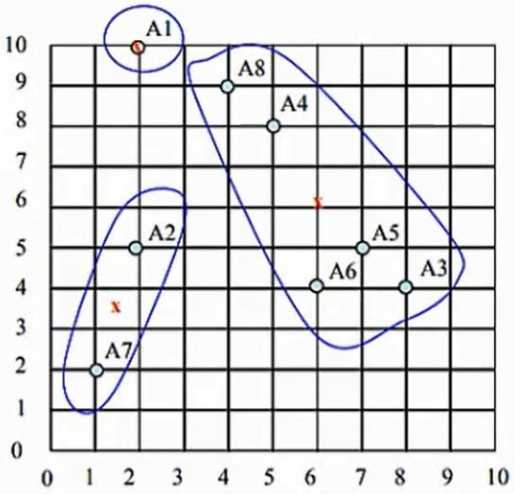
مثال



$$d(a,b)= | x2 - x1 | + | y2 - y1 |$$

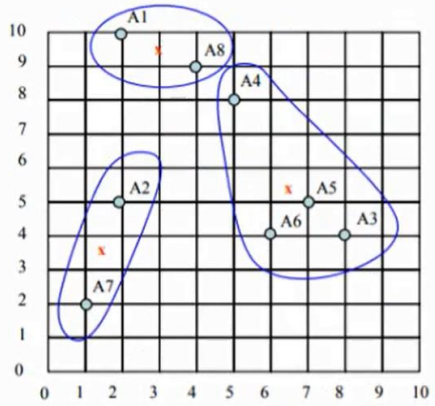
		(2,10)	(5,8)	(1,2)	cluster
A1	(2,10)	0	5	9	1
A2	(2,5)	5	6	4	3
A3	(8,4)	12	7	9	2
A4	(5,8)	5	0	10	2
A5	(7,5)	10	5	9	2
A6	(6,4)	10	5	7	2
A7	(1,2)	9	10	0	3
A8	(4,9)	3	2	10	2





(8+5+7+6+4)/5 = 6 , (4+8+5+4+9)/5) = 6

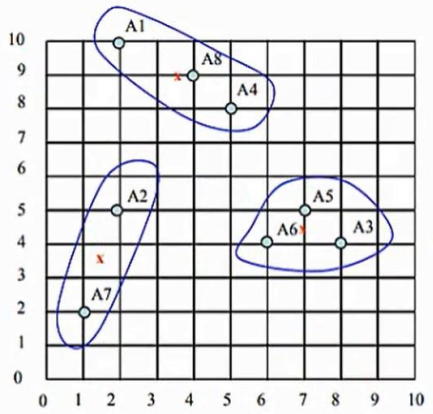
(2+1)/2 = 1.5 , (5+2)/2) = 3.5



((2+4)/2 , (10+9)/2)=(3,9.5)

((8+5+7+6)/4 , (4+8+5+4)/4)=(6.5,5.25)

((2+1)/2 , (5+2)/2)=(1.5,3.5)



((2+5+4)/2 , (10+8+9)/2)=(3.67,9)

((8+7+6)/4 , (4+5+4)/4)=(7,4.3)

((2+1)/2 , (5+2)/2)=(1.5,3.5)