# داده‌کاوی

# —خوشه‌بندی—

امینه امینی

دانشگاه آزاد اسلامی واحد کرج

# فهرست مطالب

# انگیزه

قرار است مشتریان AllElectronics به پنج بخش گروه‌بندی شوند و هر یک از این گروه‌ها توسط یکی از پنج مدیر این شرکت، مدیریت شوند.

قاعدتا مشتریان هر یک از این گروه‌ها باید تا حد ممکن به یکدیگر شبیه باشند (شباهت از لحاظ چگونگی خریدهایشان). از سوی دیگر، دو مشتری که الگوی تجاری متفاوتی دارند نباید در یک گروه قرار گیرند.

هدف، تشکیل کمپین‌هایی است که هر گروه را بر اساس ویژگی‌های مشترک مشتریان آنها هدف قرار دهند (و مواردی مانند میزان فروش، سوددهی و رضایت مشتریان را بهبود بخشند).

چگونه و با چه تکنیک داده کاوی باید این تقسیم‌بندی را انجام داد؟

# انگیزه، ادامه

گروه‌های مشتریانی که به این ترتیب باید تشکیل شوند، از قبل نامی برایشان وجود ندارد. این گروه‌بندی‌ها در واقع باید ***کشف*** شوند.

اگر تعداد مشتریان زیاد باشند و تعداد مقادیری (یعنی ویژگی‌هایی) که برای آنها نگهداری می‌شوند نیز زیاد باشند، برای یک انسان، انجام این تقسیم‌بندی هزینه‌بر و یا شاید غیرممکن باشد.

نیاز به ابزاری برای ***خوشه‌بندی*** خواهد بود.

# خوشه‌بندی، تعریف

## **خوشه‌بندی** Clustering

خوشه‌بندی، فرایند گروه‌بندی مجموعه‌ای از اشیاء داده‌ای به گروه‌ها است.

این گروه‌ها، خوشه نامیده می‌شوند.

خوشه‌بندی باید به‌گونه‌ای باشد که اشیاء درون یک خوشه شباهت زیادی به هم داشته باشند و از اشیاء درون خوشه‌های دیگر متفاوت باشند.

خوشه‌بندی یا تحلیل خوشه‌ها (Cluster Analysis)، تقسیم‌بندی یک مجموعه از داده‌ها به چند زیرمجموعه است

# خوشه‌بندی، ادامه

هر روش خوشه‌بندی می‌تواند نتایج متفاوتی داشته باشد.

یعنی زیر مجموعه‌های متفاوتی از مجموعه اصلی بسازند.

نتیجه خوشه‌بندی می‌تواند منجر به کشف گروه‌هایی از داده‌ها شود که قبلا تشخیص داده نشده بودند.

# خوشه‌بندی، کاربردها

## Business Intelligence (BI)

گروه‌بندی مشتریان

- مشتریان هر گروه شبیه یکدیگرند
  - مدیریت بهتر مشتریان

## Image Recognition

- Character Recognition

تشخیص انواع نوشتاری ارقام

## Web Search

دسته‌بندی نتایج یک جستجو برای دسترسی آسان‌تر

# خوشه‌بندی، کاربرد دیگر

## تشخیص Outlierها

- مقادیری که از بقیه خوشه‌ها دور هستند
- استفاده: استثناها در تراکنش‌های کارت اعتباری
  - مثلا خرید خیلی گران قیمت و غیرعادی

# خوشه‌بندی، ادامه

خوشه‌بندی، «یادگیری بدون نظارت» شناخته می‌شود

- زیرا اطلاعات «برچسب کلاس» موجود نیست
  - برای خوشه‌ها «نام از قبل تعیین شده‌ای» وجود ندارد

به همین دلیل، «یادگیری بوسیله مشاهده» عبارت مناسب‌تری برای خوشه‌بندی، نسبت به «یادگیری بوسیله مثال» است.

# نیازمندی‌هایی برای آنالیز خوشه‌بندی

نیازمندی‌های معمول برای خوشه‌بندی در داده‌کاوی

- مقیاس‌پذیری
  - الگوریتم خوشه‌بندی باید برای مجموعه داده بسیار بزرگ نیز بتواند خوب عمل کند
- عملکرد مناسب برای انواع مختلف ویژگی‌ها
  - Numeric, Binary, Nominal (Categorical), Ordinal and their mixture
- کشف خوشه‌هایی با شکل اختیاری (غیر دایره‌ای)
  - کشف خط حرکت آتش در جنگل
- نیاز به دانش دامنه موضوع برای تشخیص مقادیر ورودی
  - تعداد مطلوب خوشه‌ها
- رسیدگی به داده‌های نویزدار (دارای اختلال)
  - داده خطادار یا مقادیر ناموجود

# نیازمندی‌هایی برای آنالیز خوشه‌بندی، ادامه

## نیازمندی‌های معمول برای خوشه‌بندی در داده‌کاوی، ادامه

- خوشه‌بندی افزایشی (یا تدریجی) و عدم حساسیت به ترتیب مقادیر ورودی
  - امکان ورود داده‌های جدید وجود دارد: خوشه‌بندی نباید دوباره از ابتدا شروع کند
  - اگر ترتیب ورود مقادیر تغییر کند، نتیجه خوشه‌بندی باید یکسان باشد
- قابلیت خوشه‌بندی داده‌های با ابعاد زیاد
  - امکان خوشه‌بندی با کیفیت مناسب داده‌های با تعداد زیادی ویژگی برای مقادیر داده‌ای
- خوشه‌بندی با درنظرگرفتن محدودیت‌های خاص
  - مثال: تشخیص محل قرارگیری دستگاه‌های عابربانک جدید: درنظرگیری محل رودخانه‌ها و بزرگراه‌ها
- قابلیت تفسیر نتایج خوشه‌بندی

# مقایسه روش‌های خوشه‌بندی، ۱

از جنبه‌های زیر روش‌های خوشه‌بندی با یکدیگر مقایسه می‌شوند:

۱. معیار بخش‌بندی (افراز) Partitioning

- در بعضی از متدها، همه اشیا به گونه‌ای افراز می‌شوند که ساختار سلسله مراتبی بین خوشه‌ها وجود ندارد و همه خوشه‌ها در یک سطح مفهومی قرار گیرند.
  - مثال: بخش بندی کردن مشتریان به گروه‌هایی که هر کدام مدیر خاص خود را داشته باشند
- در حالیکه بعضی از متدها به صورت سلسله مراتبی افراز می‌کنند که خوشه‌ها در سطوح مختلف معنایی قرار می‌گیرند.
  - مثال: در متن کاوی بخواهیم مطالب را بر اساس چندین موضوع عمومی تقسیم‌بندی کنیم مانند سیاست و ورزش.
  - هر کدام زیر موضوعی داشته باشند: مانند فوتبال، بسکتبال، بیس بال که در سطح پایین‌تری از موضوع ورزش قرار می‌گیرند.

# مقایسه روش‌های خوشه‌بندی، ۲

۲. جداسازی خوشه‌ها

- بعضی از روش‌های خوشه‌بندی اشیاء یا اقلام داده‌ای را به خوشه‌های دوبه‌دو جدا (متمایز) از هم تقسیم می‌کنند.
- در بعضی خوشه‌بندی‌ها یک داده ممکن است، متعلق به چند خوشه باشد. مثلا وقتی اسنادی را براساس موضوع آنها خوشه‌بندی می‌کنیم، ممکن است، یک سند مربوط به چند موضوع باشد. بنابراین خوشه‌ها انحصاری نیستند

# مقایسه روش‌های خوشه‌بندی، ۳

۳. معیار شباهت

- بعضی از روش‌ها شباهت بین اشیاء داده‌ای را براساس فاصله بین اشیاء داده‌ای مشخص می‌کنند. در متدهای دیگر ممکن است بر اساس اتصال در چگالی باشد. معیارهای شباهت، اساس طراحی روش‌های خوشه‌بندی هستند
- در حالیکه روش‌های براساس فاصله از مزایای روش‌های بهینه‌سازی استفاده می‌کنند، روشهای براساس چگالی می‌توانند خوشه‌هایی با شکل دلخواه (غیردایره‌ای) تشخیص دهند.

# مقایسه روش‌های خوشه‌بندی، ۴

## ۴. فضای خوشه‌بندی

- بیشتر روش‌های خوشه‌بندی همه فضای داده‌ای را برای پیدا کردن خوشه‌ها جستجو می‌کنند. این روش‌ها برای داده‌های با تعداد ابعاد کم کاربرد دارند.
- برای داده‌های با تعداد ابعاد بالا، ممکن است صفت‌های غیرمهم زیادی باشند که باعث شوند دیگر معیار شباهت استفاده شده، قابل اطمینان نباشند.
  بنابراین خوشه‌هایی که در همه فضا پیدا شوند معنی‌دار نباشند. باید خوشه‌ها در زیرفضایی از همان داده پیدا شوند.
- خوشه‌بندی زیرفضا، خوشه‌ها و زیرفضاهایی را کشف می‌کند که نشان‌دهنده شباهت اشیاست

# مروری بر روش‌های خوشه‌بندی

## روش‌های افراز Partitioning

- در این روش‌ها، مجموعه داده‌ای با n عضو، به k بخش مختلف تقسیم می‌شود
  - هر بخش نشان‌دهنده یک خوشه است k<=n
  - خوشه‌ها، داده‌ها را به k گروه تقسیم می‌کنند که هر یک حداقل یک عضو دارند

## روش‌های سلسله‌مراتبی Hierarchical

- اشیاء داده‌ای به ساختار سلسله‌مراتبی تجزیه می‌شوند

## روش‌های مبتنی بر چگالی Density

- رشد خوشه تا جایی که تعداد اشیاء داده‌ای همسایه از یک «حد آستانه» بیشتر باشد

## روش‌های مبتنی بر گرید Grid

- فضای اشیاء داده‌ای به تعدادی سلول تقسیم می‌شود و سلول‌ها خوشه‌بندی می‌شوند

# روش‌های خوشه‌بندی

| روش | ویژگی‌های کلی |
|---|---|
| افراز | پیدا کردن خوشه‌های متمایز به شکل کروی<br>مبتنی بر فاصله<br>ممکن است از میانگین یا medoid برای نمایش مرکز خوشه استفاده کند<br>مناسب برای داده‌های با اندازه (تعداد مقادیر) کم و متوسط |
| سلسله مراتبی | خوشه‌بندی یک تجزیه سلسله مراتبی است<br>عدم توانایی تصحیح خطاهای ادغام یا جداسازی |
| مبتنی بر چگالی | پیداکردن خوشه‌های با شکل‌های مختلف (غیرکروی)<br>خوشه‌ها ناحیه‌های متراکم اشیاء هستند<br>چگالی خوشه: هر نقطه باید حداقل تعدادی نقطه دیگر در همسایگی داشته باشد<br>می‌تواند outlierها را جداسازی کند |
| مبتنی بر گرید | ساختار گرید با رزولوشن مختلف<br>سرعت بالای پردازش (مستقل از تعداد اشیاء داده‌ای و وابسته به اندازه گرید) |

# روش‌های خوشه‌بندی

**روش‌های افراز Partitioning**

روش‌های سلسله‌مراتبی Hierarchical

روش‌های مبتنی بر چگالی Density

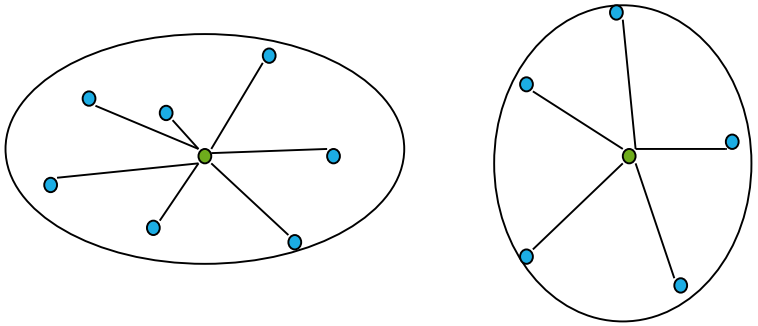
روش‌های مبتنی بر گرید Grid

# روش‌های افراز

روش افراز:

افراز یک پایگاه داده D با n شیء داده‌ای به مجموعه‌ای از k خوشه یا کلاستر، بطوریکه مجموع مربعات فاصله‌ها از مرکز خوشه‌ها در آن حداقل شود ($c_i$ میانگین یا medoid از کلاستر $C_i$ است)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2$$



دو الگوریتم در افراز عبارتند از:

k-Means: روشی مبتنی بر مرکز

k-Medoids: روشی مبتنی بر اشیاء داده‌ای

# k-Means

مجموعه داده‌ای در فضای اقلیدسی وجود دارد

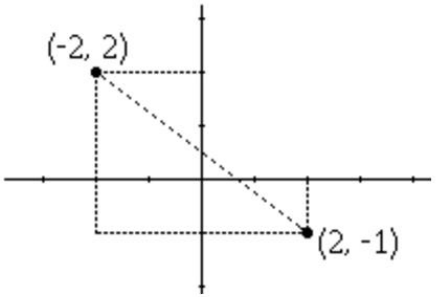
- تعداد اشیاء داده‌ای: n

نتیجه اجرای افراز k-Means:

- k خوشه:

$$C_1, \ldots, C_k, \text{ that is, } C_i \subset D \text{ and } C_i \cap C_j = \emptyset \text{ for } (1 \leq i, j \leq k)$$

- مرکز (centroid) هر خوشه: مرکز یا وسط آن خوشه
- $dist(p,c_i)$: فاصله اقلیدسی p تا مرکز خوشه $c_i$

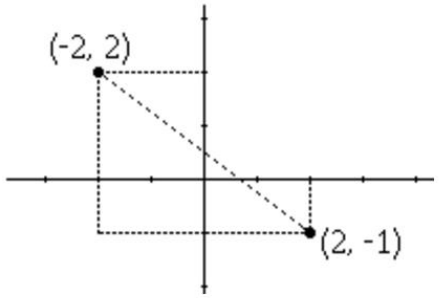



$$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

# k-Means

محاسبه فاصله اقلیدسی

$$\begin{aligned}
\text{dist}((2, -1), (-2, 2)) &= \sqrt{(2 - (-2))^2 + ((-1) - 2)^2} \\
&= \sqrt{(2 + 2)^2 + (-1 - 2)^2} \\
&= \sqrt{(4)^2 + (-3)^2} \\
&= \sqrt{16 + 9} \\
&= \sqrt{25} \\
&= 5.
\end{aligned}$$



# k-Means: ماتریس فاصله

ماتریس داده

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

ماتریس فاصله‌ها

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

# الگوریتم k-Means

**Algorithm: *k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

**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) **repeat**
(3) (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;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

# الگوریتم k-Means

مرکز هر خوشه: میانگین (mean value) اشیاء داده‌ای در آن خوشه

- ورودی‌ها: k: تعداد خوشه‌ها، D: مجموعه داده‌ای شامل n شیء داده‌ای
- خروجی: k خوشه

۱. بصورت تصادفی k شیء از D را به عنوان مراکز اولیه خوشه‌ها انتخاب کن

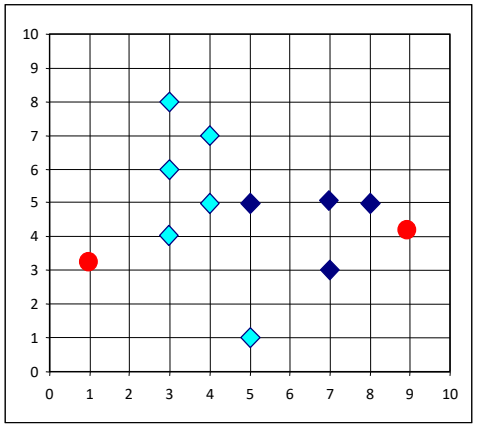
۲. شروع حلقه

۳. تخصیص (یا تخصیص مجدد) هر شیء داده‌ای به خوشه‌ای که بیشترین میزان شباهت را دارد (معیار: میانگین (mean value) اشیاء داده‌ای در آن خوشه)

۴. بروزرسانی مرکز خوشه (محاسبه میانگین (mean value) اشیاء داده‌ای هر خوشه

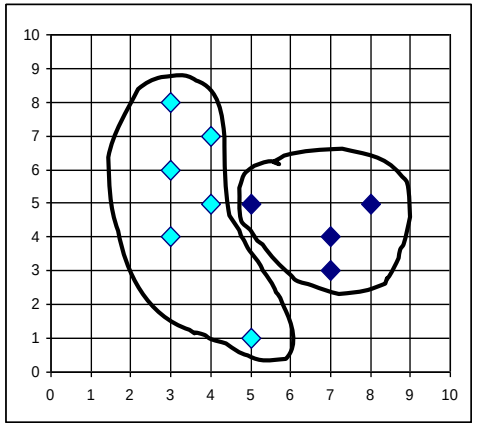
۵. پایان حلقه: تکرار تا زمانی که دیگر تغییر جدیدی اتفاق نیفتاده باشد

# The *k-Means* Clustering Method

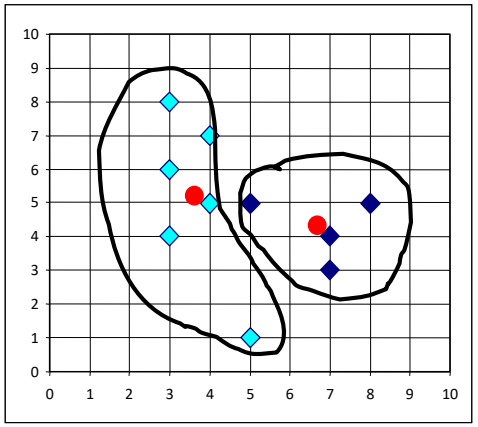


K=2

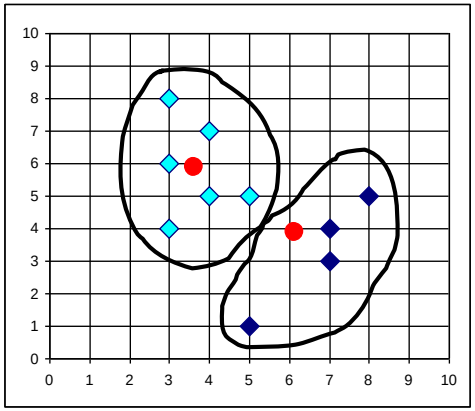
Arbitrarily choose K object as initial cluster center



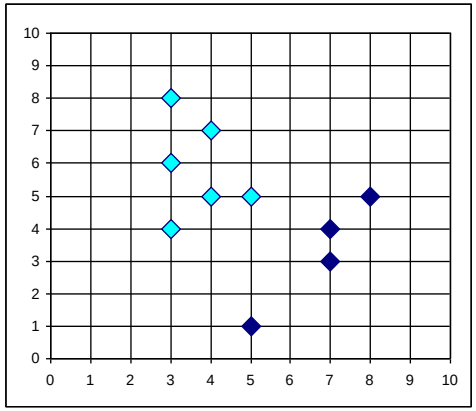
Assign each objects to most similar center



Update the cluster means

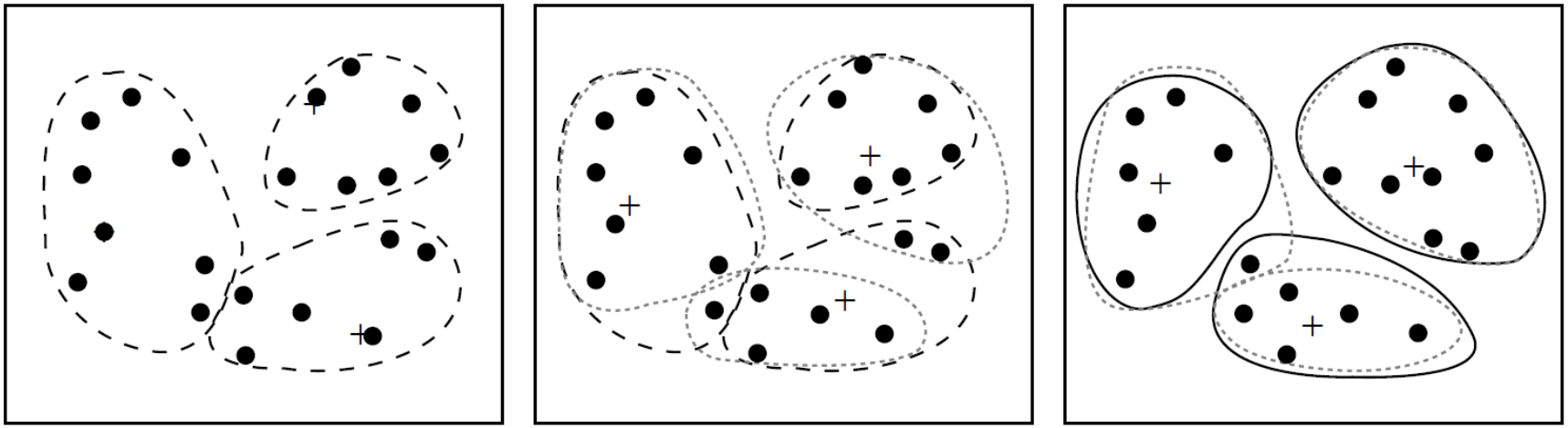


reassign



Update the cluster means

# K-Means: خوشه‌بندی به سه خوشه



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

# k-Means :مثال

***Exercise 1.* K-means clustering**

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters:
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
The distance matrix based on the Euclidean distance is given below:

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| A1 | 0 | $\sqrt{25}$ | $\sqrt{36}$ | $\sqrt{13}$ | $\sqrt{50}$ | $\sqrt{52}$ | $\sqrt{65}$ | $\sqrt{5}$ |
| A2 | | 0 | $\sqrt{37}$ | $\sqrt{18}$ | $\sqrt{25}$ | $\sqrt{17}$ | $\sqrt{10}$ | $\sqrt{20}$ |
| A3 | | | 0 | $\sqrt{25}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{53}$ | $\sqrt{41}$ |
| A4 | | | | 0 | $\sqrt{13}$ | $\sqrt{17}$ | $\sqrt{52}$ | $\sqrt{2}$ |
| A5 | | | | | 0 | $\sqrt{2}$ | $\sqrt{45}$ | $\sqrt{25}$ |
| A6 | | | | | | 0 | $\sqrt{29}$ | $\sqrt{29}$ |
| A7 | | | | | | | 0 | $\sqrt{58}$ |
| A8 | | | | | | | | 0 |

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:
a) The new clusters (i.e. the examples belonging to each cluster)
b) The centers of the new clusters
c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
d) How many more iterations are needed to converge? Draw the result for each epoch.

# k-Means: مثال

***Solution:***

a)

d(a,b) denotes the Eucledian distance between a and b. It is obtained directly from the distance matrix or calculated as follows: $d(a,b)=sqrt((x_b-x_a)^2+(y_b-y_a)^2))$

seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

epoch1 – start:

A1:

d(A1, seed1)=0 as A1 is seed1

d(A1, seed2)= $\sqrt{13}$ >0

d(A1, seed3)= $\sqrt{65}$ >0

➔A1 ∈ cluster1

A2:

d(A2,seed1)= $\sqrt{25}$ = 5

d(A2, seed2)= $\sqrt{18}$ = 4.24

d(A2, seed3)= $\sqrt{10}$ = 3.16 ← smaller

➔ A2 ∈ cluster3

A3:

d(A3, seed1)= $\sqrt{36}$ = 6

d(A3, seed2)= $\sqrt{25}$ = 5 ← smaller

d(A3, seed3)= $\sqrt{53}$ = 7.28

➔ A3 ∈ cluster2

A4:

d(A4, seed1)= $\sqrt{13}$

d(A4, seed2)=0 as A4 is seed2

d(A4, seed3)= $\sqrt{52}$ >0

➔ A4 ∈ cluster2

A5:

d(A5, seed1)= $\sqrt{50}$ = 7.07

A6:

d(A6, seed1)= $\sqrt{52}$ = 7.21

# k-Means: مثال

d(A5, seed2)= $\sqrt{13}$ = 3.60 ← smaller
d(A5, seed3)= $\sqrt{45}$ = 6.70
➔ A5 ∈ cluster2

d(A6, seed2)= $\sqrt{17}$ = 4.12 ← smaller
d(A6, seed3)= $\sqrt{29}$ = 5.38
➔ A6 ∈ cluster2

A7:
d(A7, seed1)= $\sqrt{65}$ >0
d(A7, seed2)= $\sqrt{52}$ >0
d(A7, seed3)=0 as A7 is seed3
➔ A7 ∈ cluster3

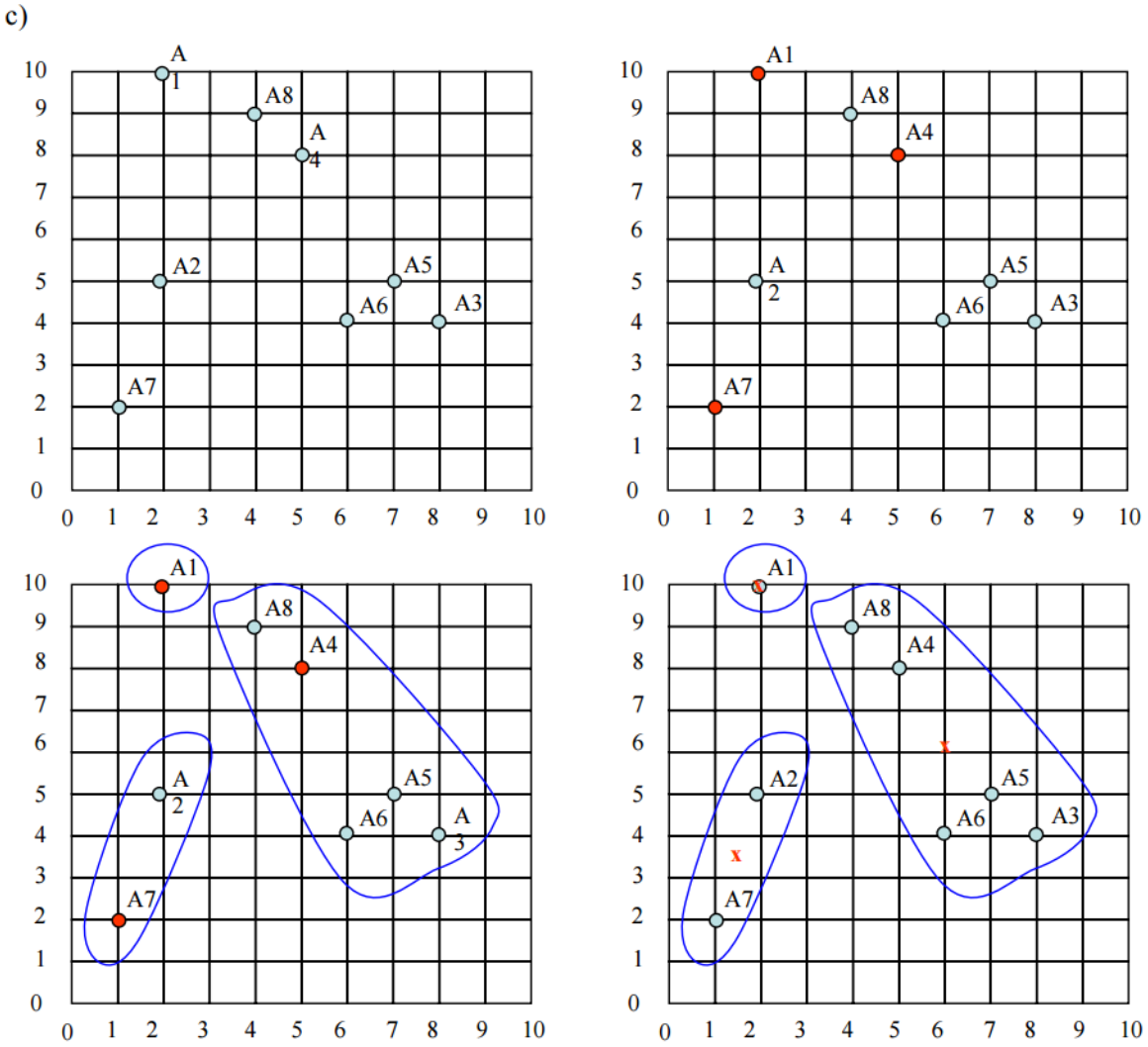
A8:
d(A8, seed1)= $\sqrt{5}$
d(A8, seed2)= $\sqrt{2}$ ← smaller
d(A8, seed3)= $\sqrt{58}$
➔ A8 ∈ cluster2

end of epoch1

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

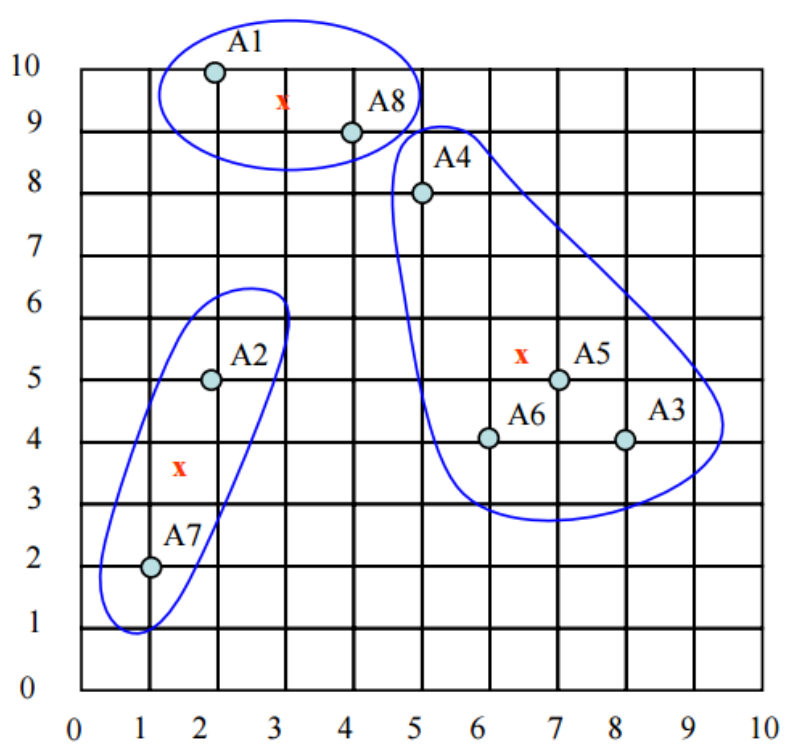
b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

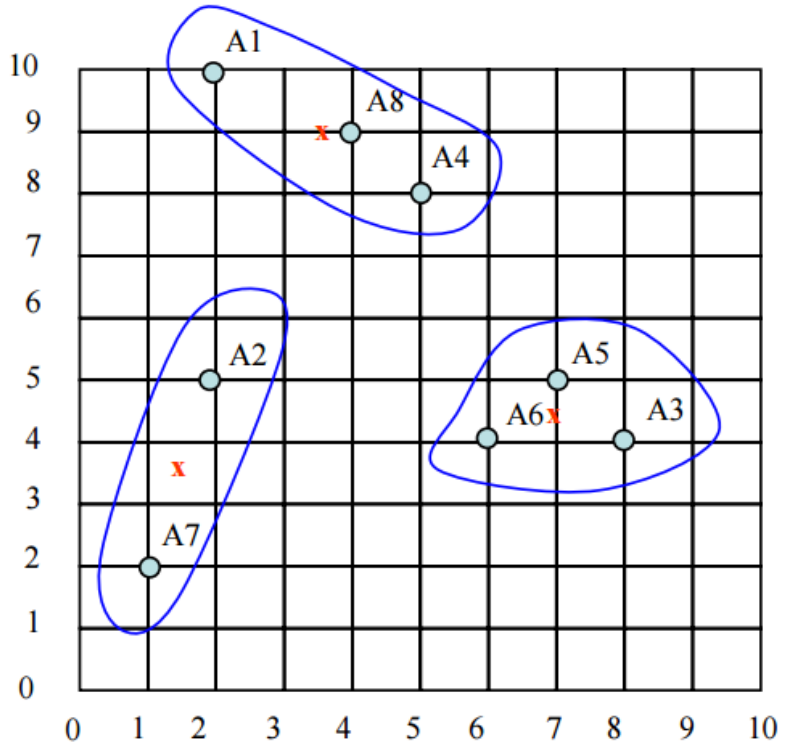
# k-Means: مثال



# k-Means: مثال

d)
We would need two more epochs. After the 2nd epoch the results would be:
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}
with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5).
After the 3rd epoch, the results would be:
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}
with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).



# k-Means: کیفیت نتیجه خوشه‌بندی

مجموع توان دوم فواصل نقاط تا مرکز خوشه

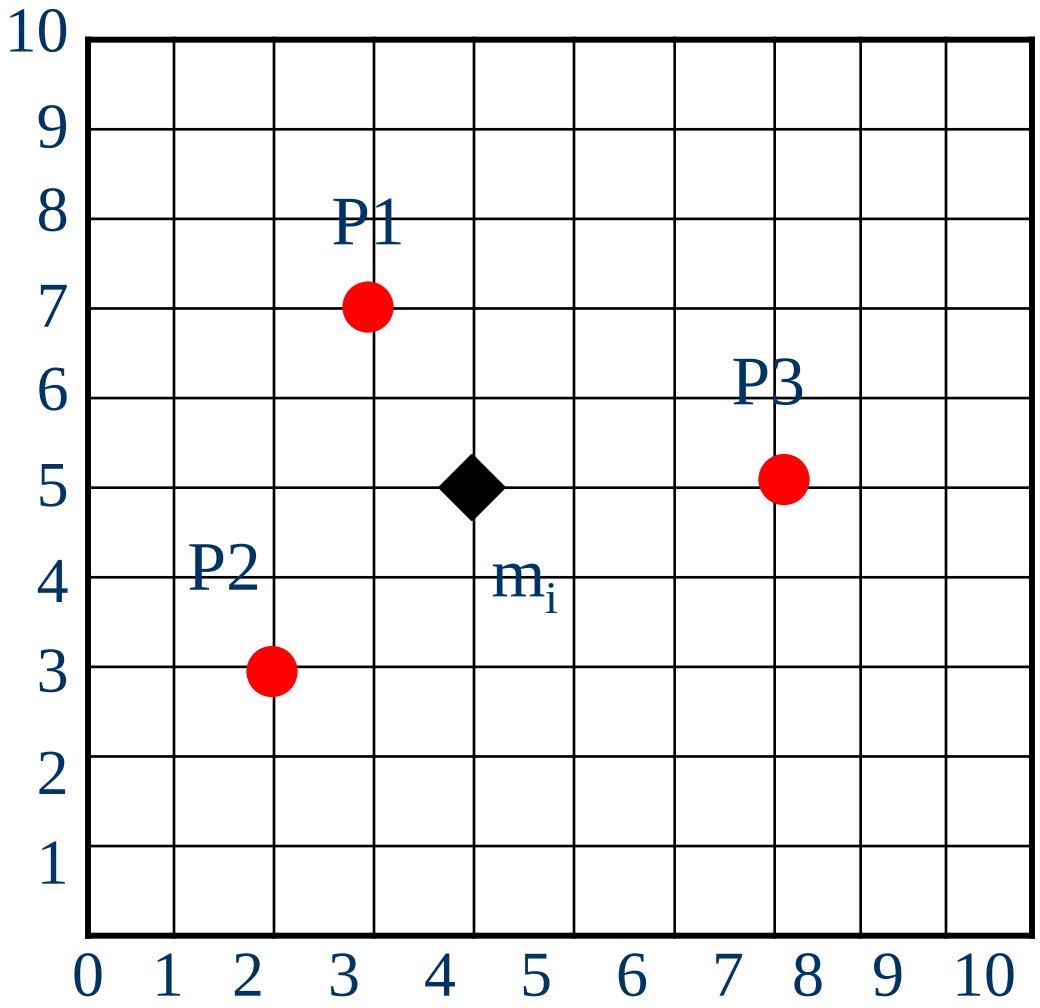
$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2$$

افراز بر اساس اصل به حداقل رساندن تفاوت بین داده p و داده نماینده مربوطه است، که به آن معیار خطای کامل absolute-error گفته می‌شود

در این فرمول، E مجموع خطاها برای همه اشیاء داده‌ای موجود در مجموعه داده است

هر چه خوشه فشرده‌تر و از سایر خوشه‌ها متمایزتر (جداشده‌تر) باشد، بهتر است

# مثال: محاسبه کیفیت خوشه‌بندی




$C_i$={P1, P2, P3}

P1 = (3, 7)

P2 = (2, 3)

P3 = (7, 5)

$m_i$ = (4, 5)

$|d(P1, m_i)|^2$

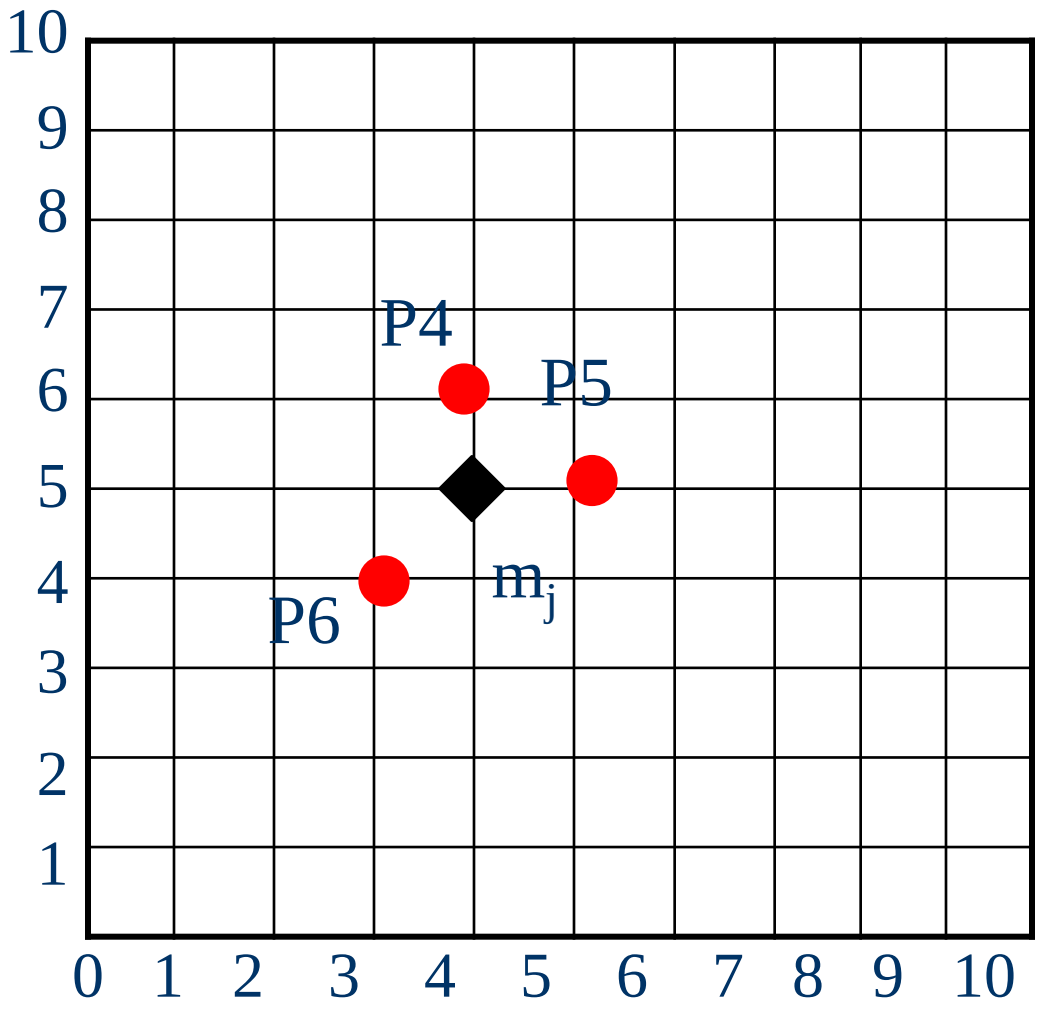
$=(3-4)^2+(7-5)^2=5$

$|d(P2, m_i)|^2=8$

$|d(P3, m_i)|^2=9$

Error ($C_i$)=5+8+9=22

# مثال: محاسبه کیفیت خوشه‌بندی




$C_j$={P4, P5, P6}

P4 = (4, 6)

P5 = (5, 5)

P6 = (3, 4)

$m_j$ = (4, 5)

$|d(P4, m_j)|^2$

$=(4-4)^2+(6-5)^2=1$

$|d(P5, m_j)|^2=1$

$|d(P6, m_j)|^2=2$

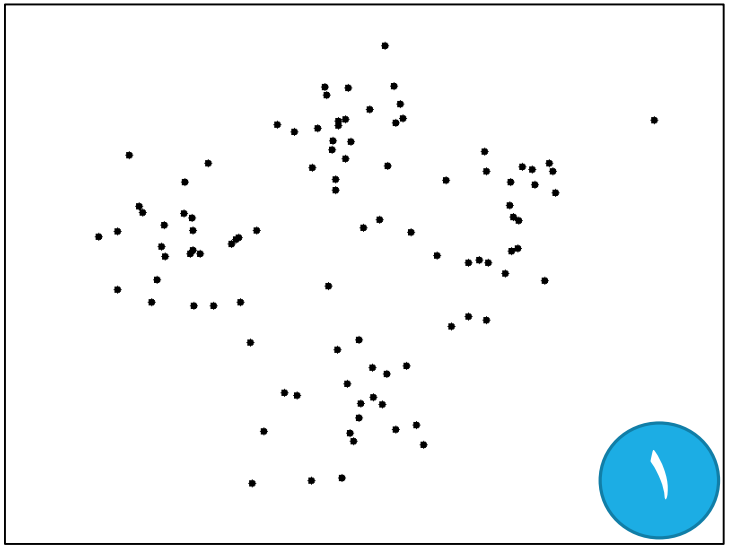
Error ($C_j$)=1+1+2=4

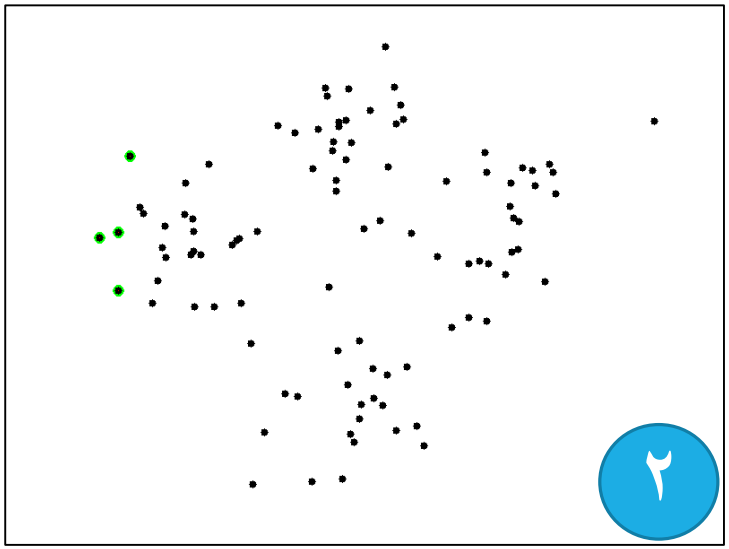
# k-Means: بهینه محلی

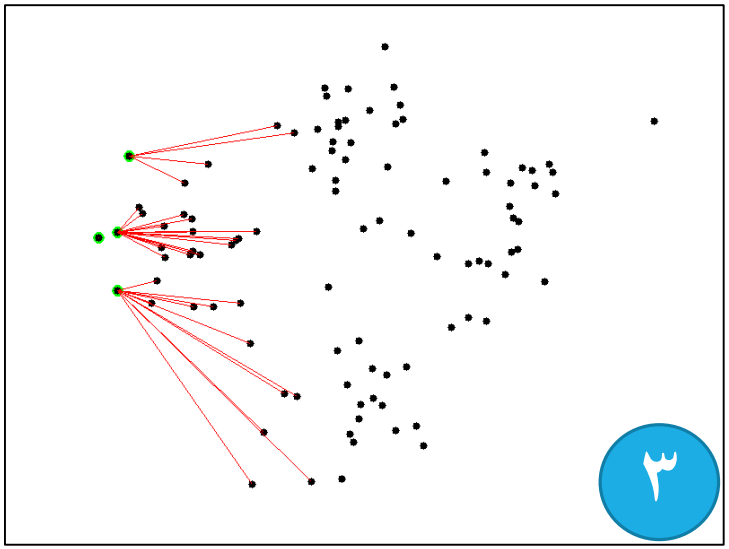
تضمینی وجود ندارد که k-Means نتیجه بهینه عمومی تولید نماید.

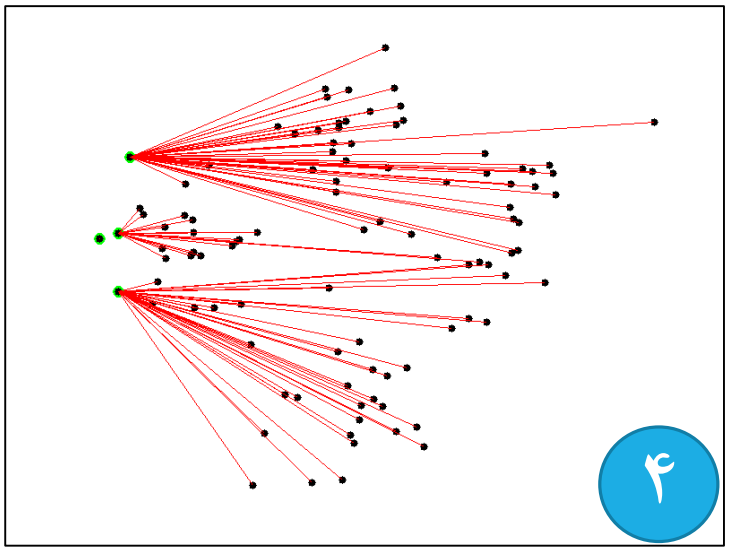
- نتیجه، معمولا بهینه محلی است.
- نتیجه، وابسته به انتخاب نقاط اولیه است.

# K-Means, 4 left-most points

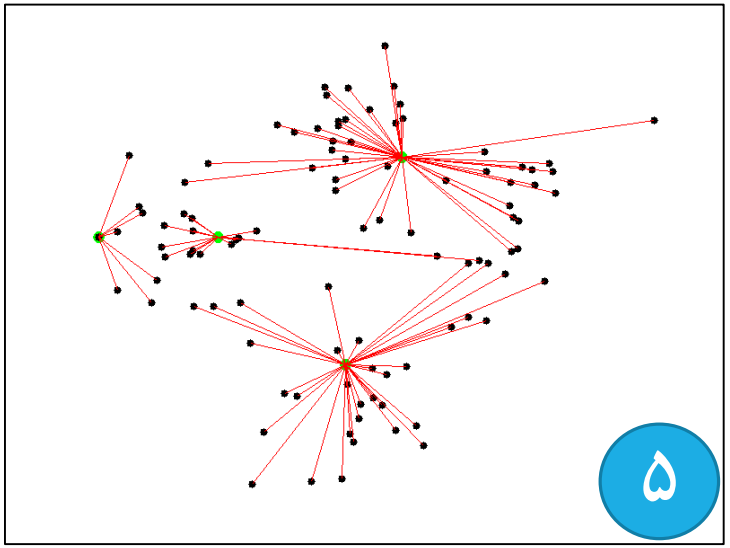


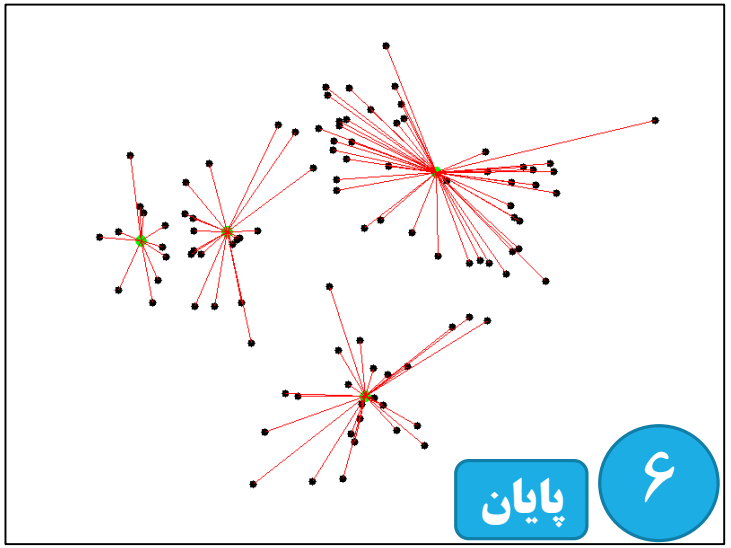



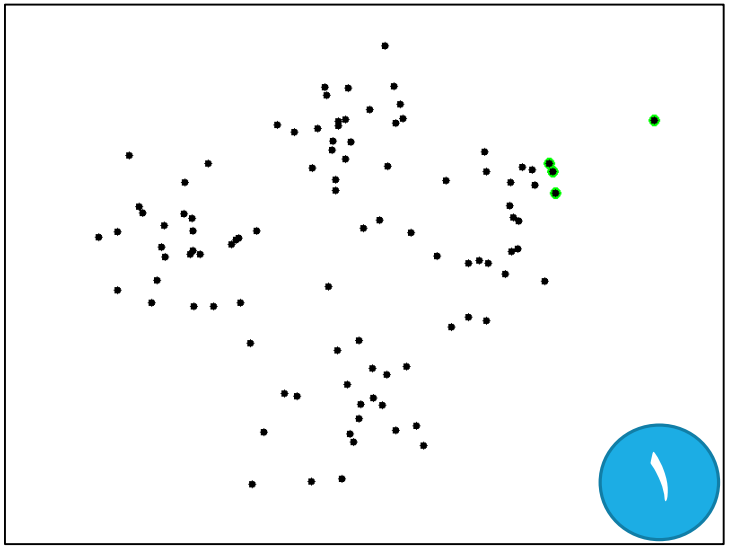
# K-Means, 4 left-most points, cont’d

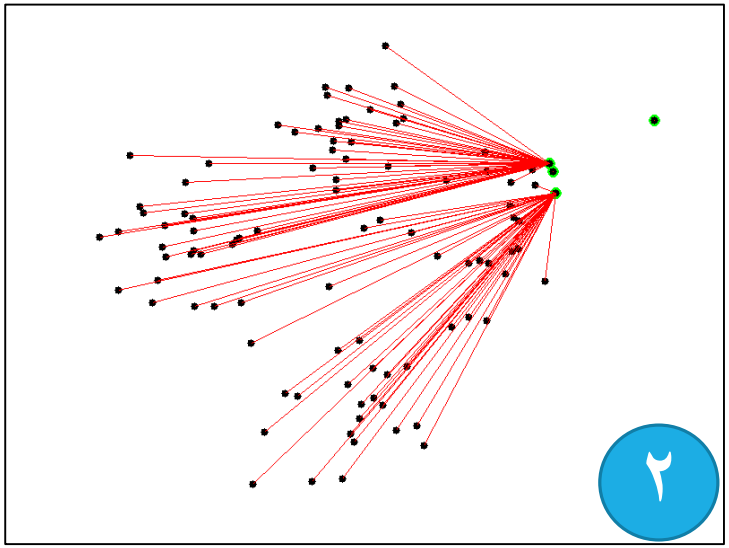


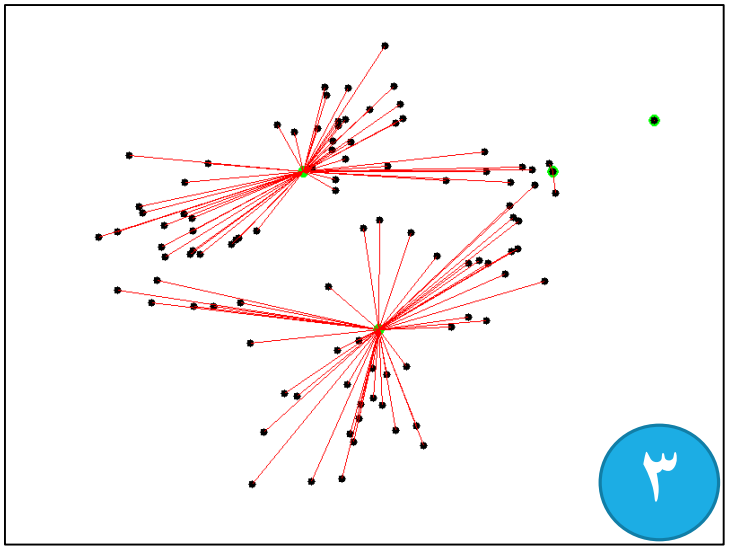



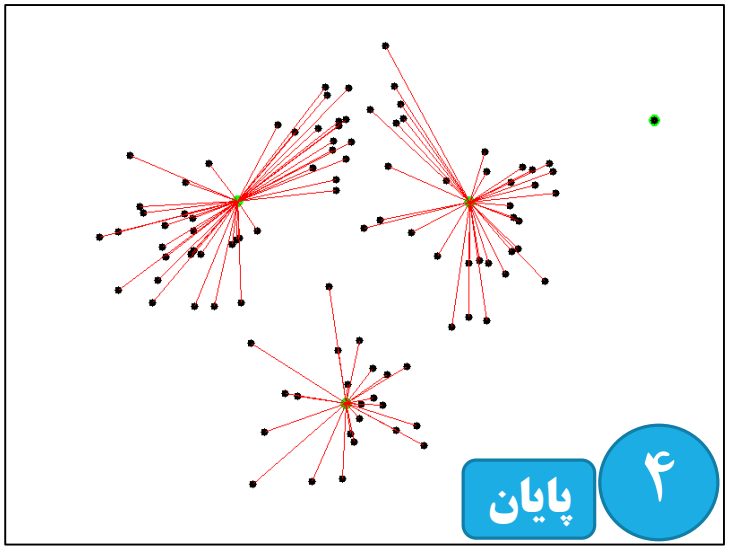

http://shabal.in/visuals/kmeans/2.html

# K-Means, 4 right-most points

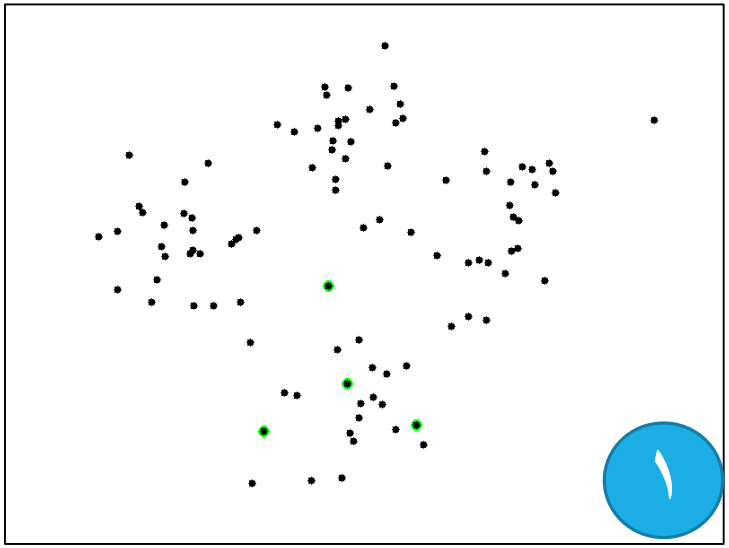


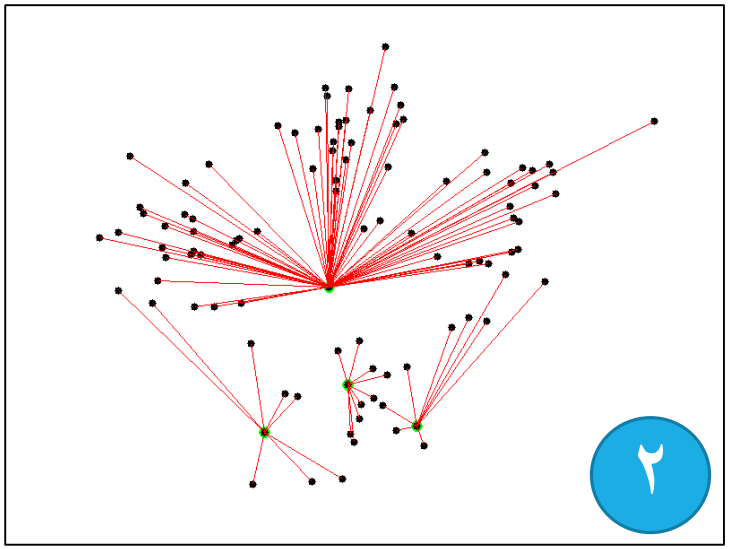



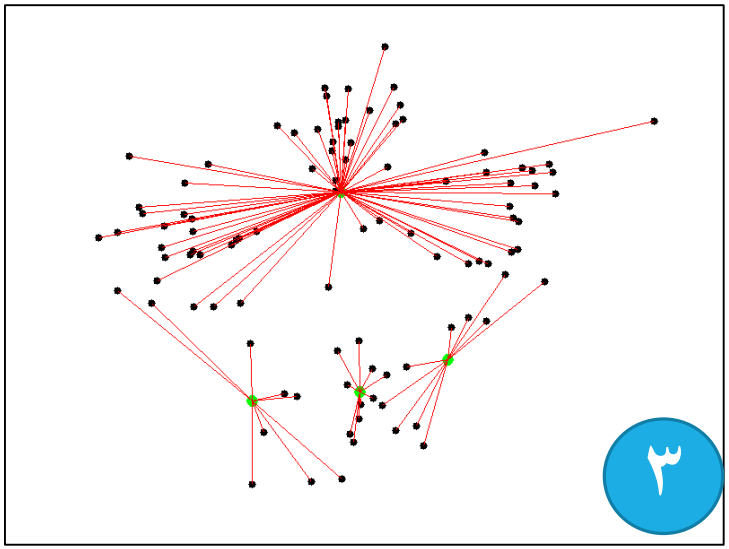



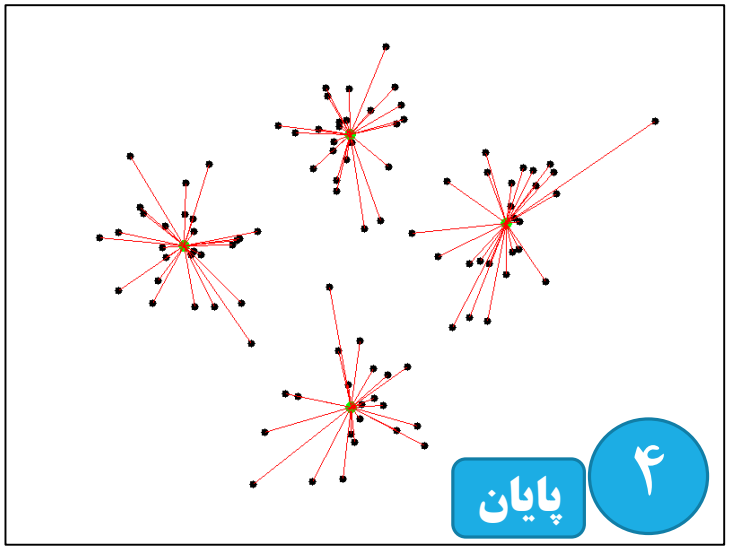



# K-Means, 4 random points



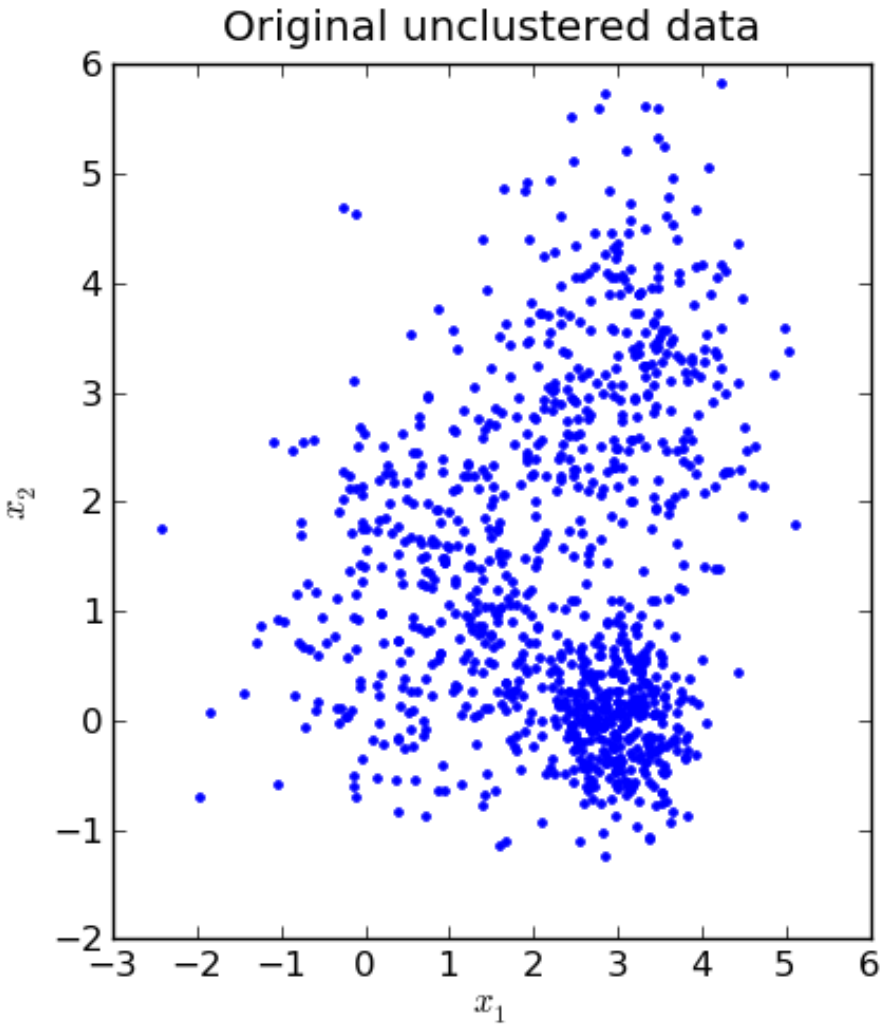
# k-Means: تعداد خوشه‌ها

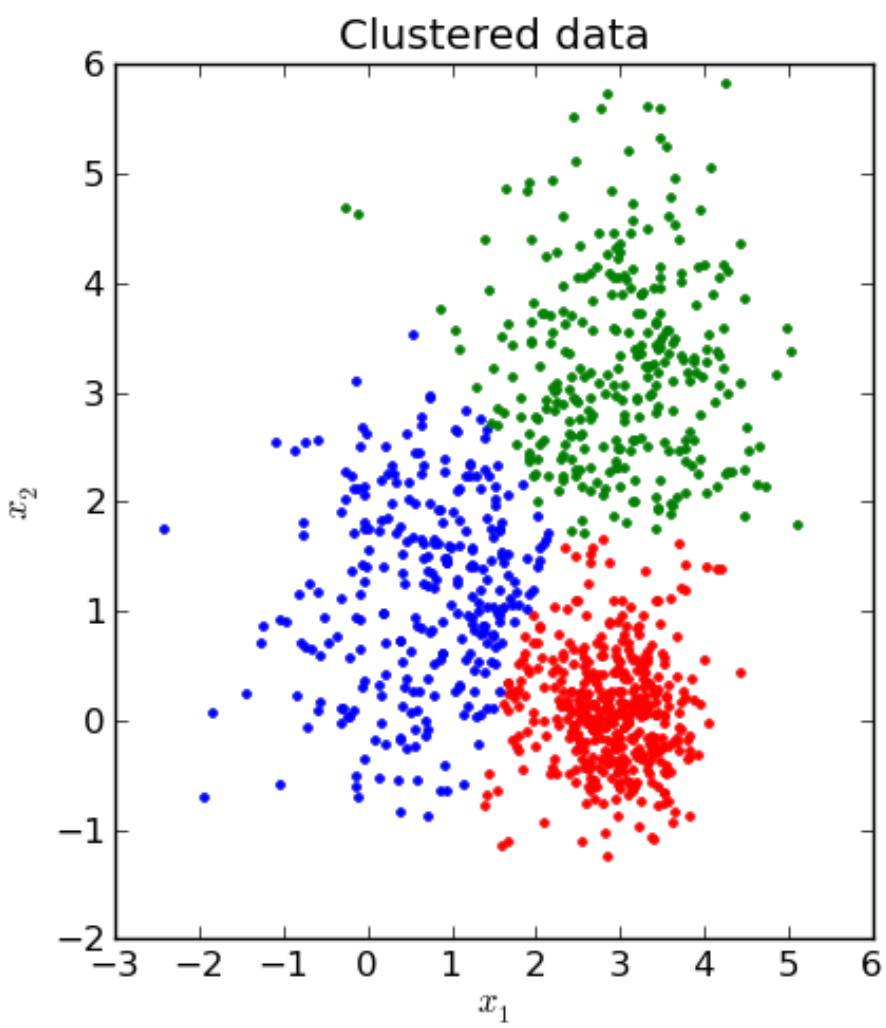
تعداد خوشه‌ها، ورودی الگوریتم است.

تعداد خوشه‌ها چه مقداری باید باشد؟!

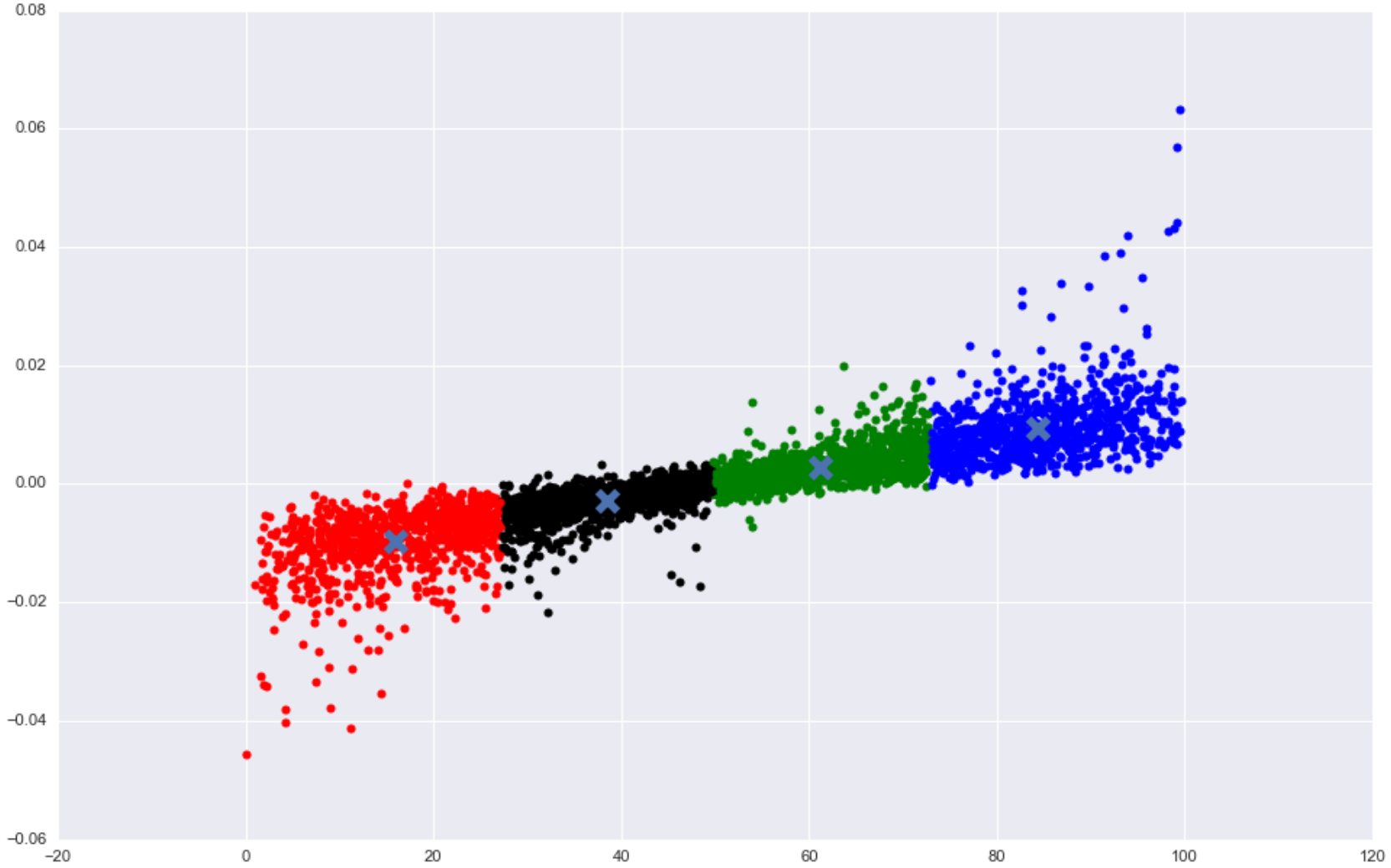
- می‌توان الگوریتم را برای محدوده‌ای از مقادیر (چند مرتبه مختلف) اجرا کرد و بهترین k را انتخاب نمود.

# k-Means: نمونه‌های اجرا، ۱



# k-Means: نمونه‌های اجرا، ۲

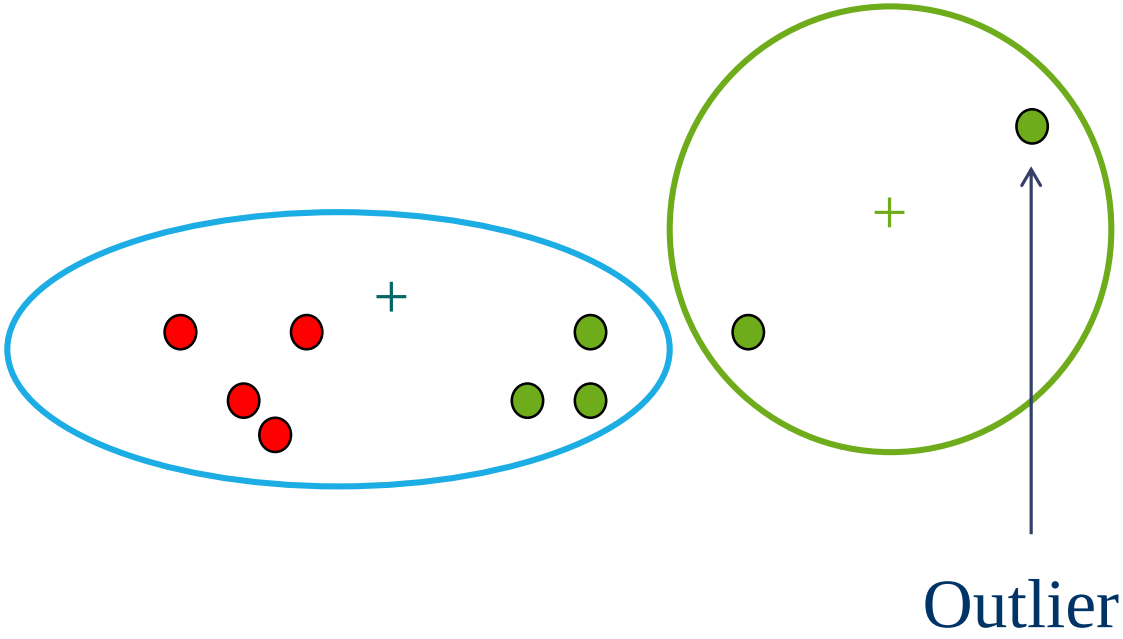


# k-Means: شکل خوشه‌ها

شکل خوشه‌ها معمولا دایره‌ای شکل و کروی شکل هستند

اندازه خوشه‌ها هم تقریبا یکسان هستند

این الگوریتم به وجود outlier و نویز نیز حساس است



# روش‌های افراز

k-Means: روشی مبتنی بر مرکز

**k-Medoids**: روشی مبتنی بر اشیاء داده‌ای

# k-Medoids

k-Means به outlierها حساس است چون آنها از اکثریت داده‌ها دور هستند و مقدار میانگین را از مقدار واقعی آن دور می‌کنند

مقدار میانگین نادرست، روی اختصاص بقیه داده‌ها به خوشه‌ها تاثیر می‌گذارد

این تاثیر به دلیل استفاده از تابع مربعات خطاها (*squared*-error) تشدید می‌شود

چطور می توان تاثیر حساسیت به حضور outlierها را کم کرد؟

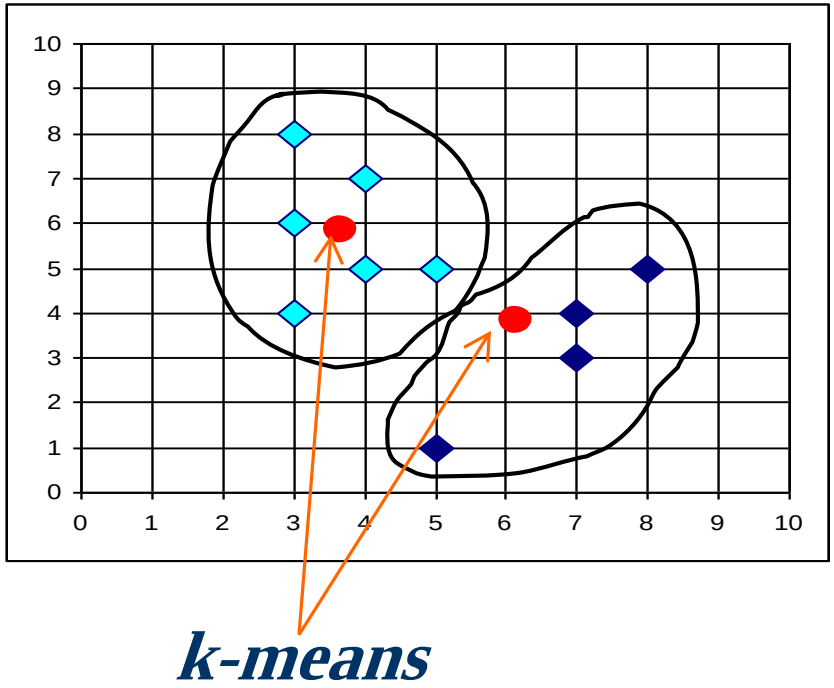
# k-Medoids

به جای استفاده از مقدار میانگین داده‌ها در یک خوشه بعنوان نقطه ارجاع، از یک نقطه واقعی برای نمایش خوشه استفاده شود. این نقطه یک نماینده داده‌ای برای هر خوشه خواهد بود.

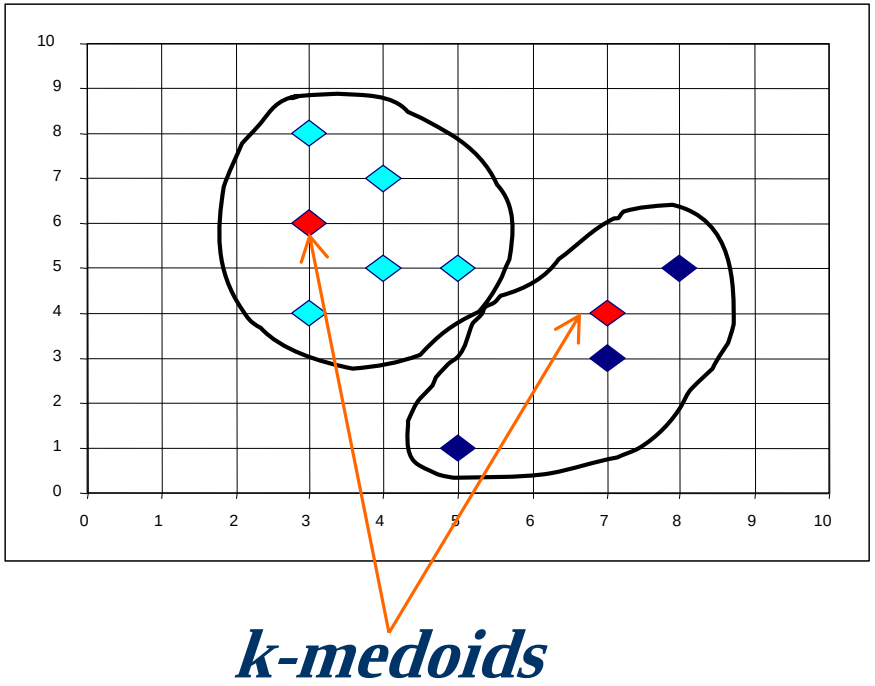
بقیه نقاط به خوشه‌ای اختصاص می‌یابند که به نماینده انتخاب شده آن خوشه، نزدیکتر باشد.

اصل شکل‌گیری روش k-Medoids، تقسیم بندی داده‌ها به k خوشه است بطوریکه معیار absolute-error به حداقل برسد

# مقایسه k-Medoids و k-Means




***k-means***




***k-medoids***

# k-Medoids

یک الگوریتم افراز k-Medoids:

## Partitioning Around Medoids (PAM)

### الگوریتم PAM

- نقاط نماینده (هسته‌ها) به صورت تصادفی انتخاب می‌شوند
- سپس بررسی می‌شود که اگر این نقطه (هسته) با نقطه دیگری که نماینده نیست عوض شود، آیا کیفیت بهبود می یابد؟
- همه جایگزین‌های ممکن آزمایش می‌شود.
- پروسه تکراری جایگزینی داده‌ی نماینده تا زمانی ادامه پیدا می‌کند که دیگر کیفیت خوشه‌بندی با هیچ جایگزینی قابل بهبود نباشد
- کیفیت با استفاده از تابع هزینه‌ای که میانگین تفاوت بین شیء داده‌ای و نماینده آن خوشه است، اندازه‌گیری می‌شود

# PAM, a *k*-Medoids partitioning algorithm

**Algorithm: *k*-medoids.** PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

**Input:**

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

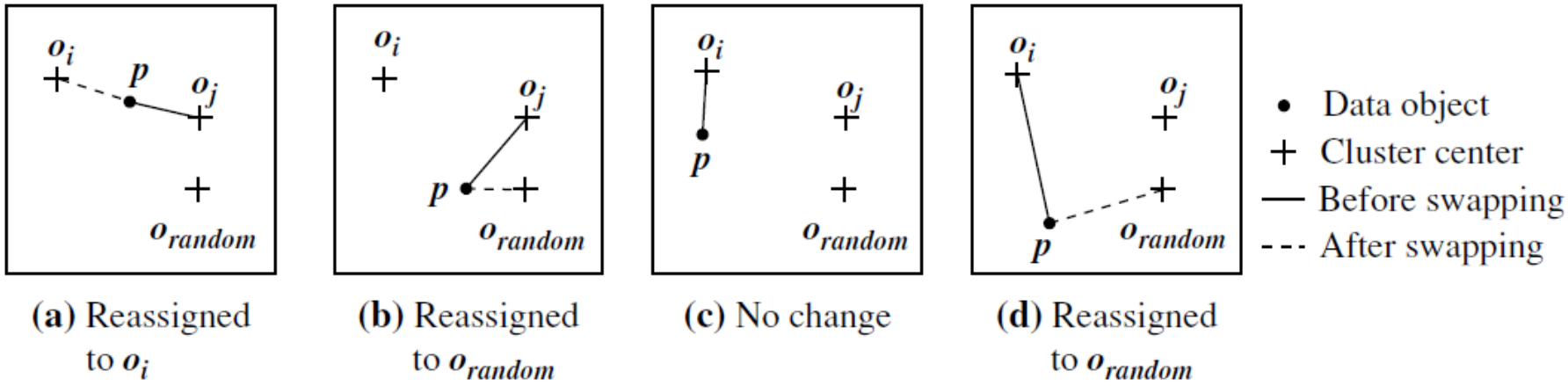
**Output:** A set of *k* clusters.

**Method:**

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, $o_{random}$;
(5) compute the total cost, *S*, of swapping representative object, $o_j$, with $o_{random}$;
(6) **if** $S < 0$ **then** swap $o_j$ with $o_{random}$ to form the new set of *k* representative objects;
(7) **until** no change;

# k-Medoids

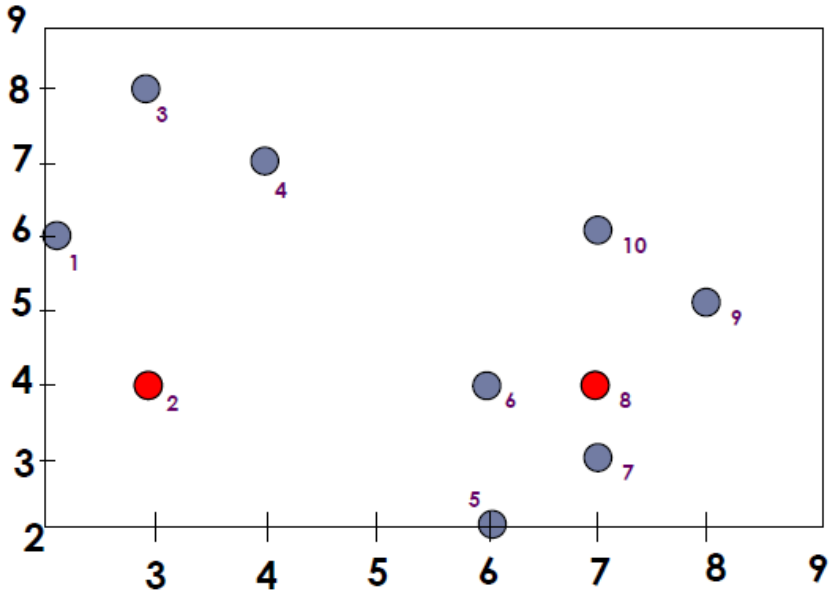
## ۴ حالت از تابع هزینه در k-Medoids



# مثال PAM

**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |




هدف: ایجاد دو خوشه می‌باشد

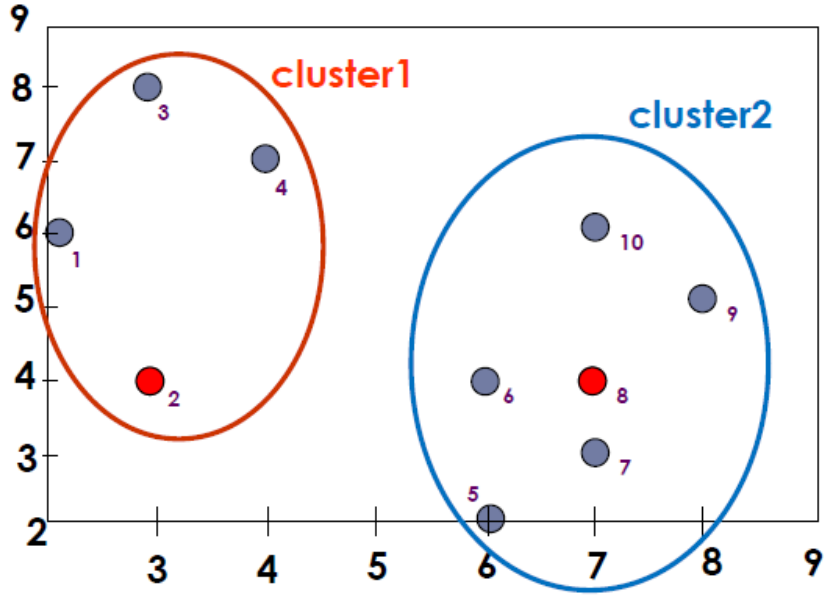
بصورت تصادفی دو داده را بعنوان Medoids انتخاب می کنیم

$O_2 = (3,4)$
$O_8 = (7,4)$

# مثال PAM

**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |




هر داده به نزدیکترین نماینده داده‌ای اختصاص می‌یابد. بنابراین خوشه‌های زیر را خواهیم داشت

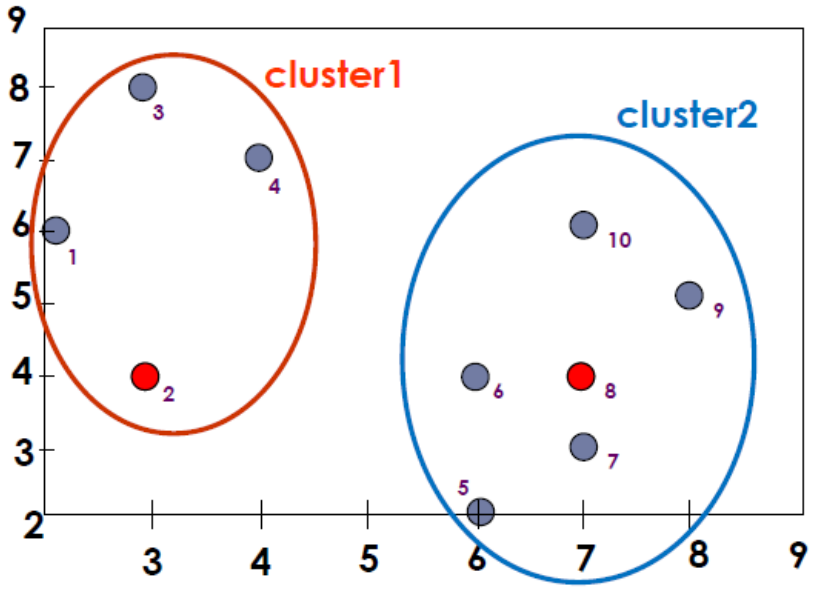
Cluster1 = $\{O_1, O_2, O_3, O_4\}$

Cluster2 = $\{O_5, O_6, O_7, O_8, O_9, O_{10}\}$

# مثال PAM

**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |




مقدار معیار خطای کامل absolute error را برای medoids $(o_2,o_8)$ محاسبه می کنیم

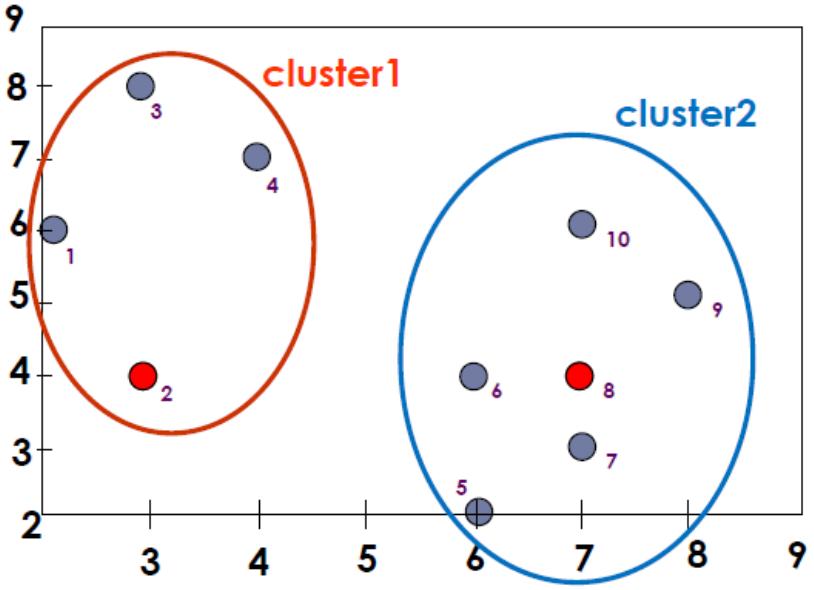
$$E=\sum_{i=1}^{k}\sum_{p\in C_i}|p-o_i|=|o_1-o_2|+|o_3-o_2|+|o_4-o_2|$$

$$+|o_5-o_8|+|o_6-o_8|+|o_7-o_8|+|o_9-o_8|+|o_{10}-o_8|$$

# مثال PAM

**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |



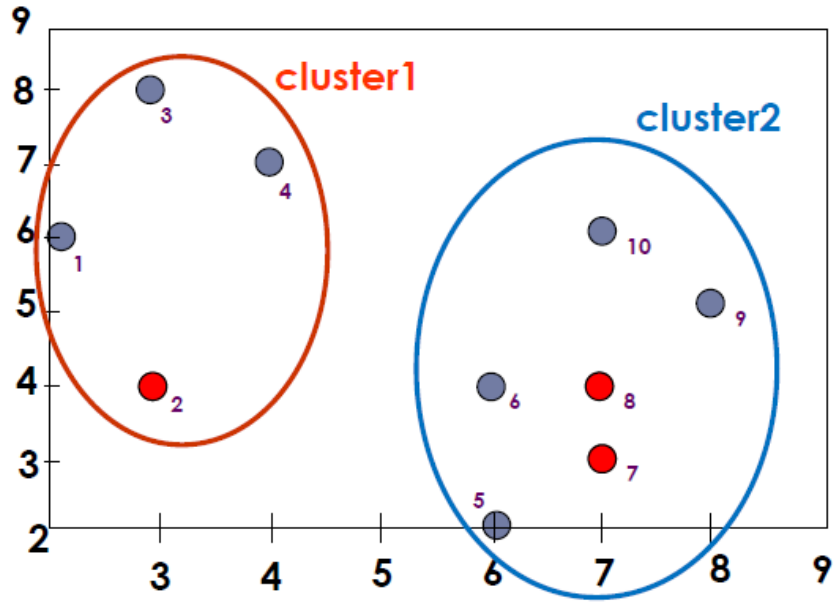

مقدار معیار خطای کامل: absolute error

$$E = (3+4+4)+(3+1+1+2+2) = 20$$

# مثال PAM

**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |




$o_7$ را بعنوان داده تصادفی انتخاب می کنیم

جای $o_7$ و$o_8$ را عوض می‌کنیم

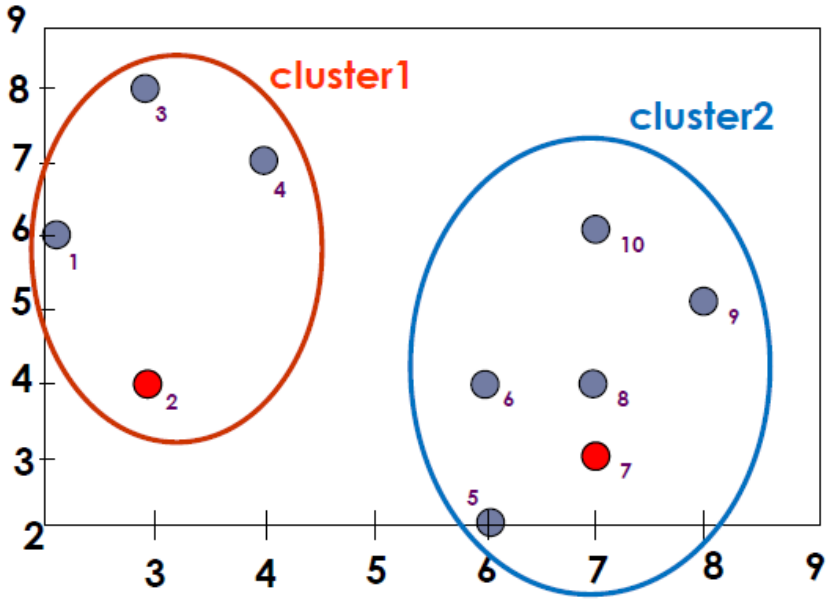
مقدار absolute error رابرای $medoids(o_2,o_7)$ محاسبه می‌کنیم

$$E = (3+4+4)+(2+2+1+3+3) = 22$$

# مثال PAM

**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |




→Compute the cost function

Absolute error [for $O_2$,$O_7$] – Absolute error [$O_2$,$O_8$]

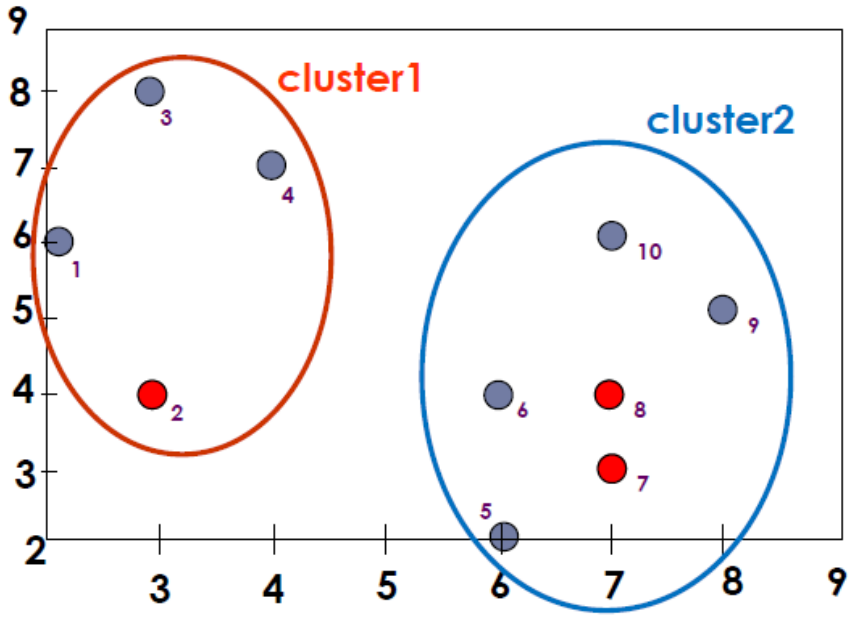
$$S = 22 - 20$$

S> 0 ⇒ it is a bad idea to replace $O_8$ by $O_7$

# مثال PAM

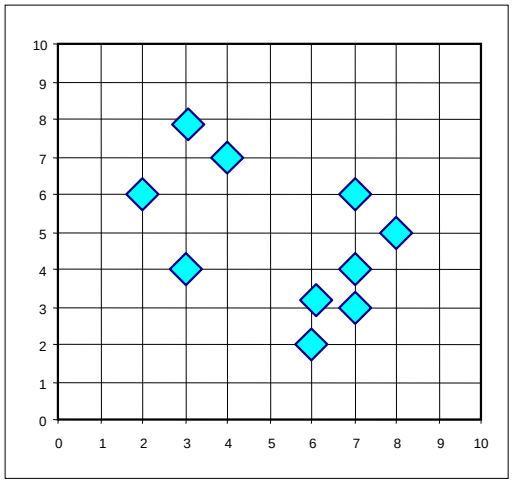
**Data Objects**

| | $A_1$ | $A_2$ |
|---|---|---|
| $O_1$ | 2 | 6 |
| $O_2$ | 3 | 4 |
| $O_3$ | 3 | 8 |
| $O_4$ | 4 | 7 |
| $O_5$ | 6 | 2 |
| $O_6$ | 6 | 4 |
| $O_7$ | 7 | 3 |
| $O_8$ | 7 | 4 |
| $O_9$ | 8 | 5 |
| $O_{10}$ | 7 | 6 |

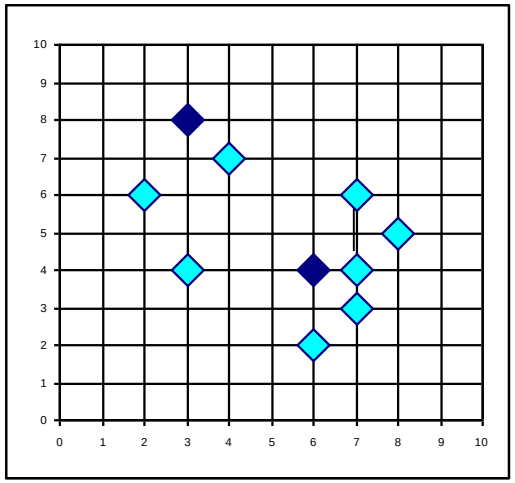



در این مثال، medoids در خوشه ۲، اختصاص اشیاء داده‌ای به خوشه را تغییر نمی‌دهد

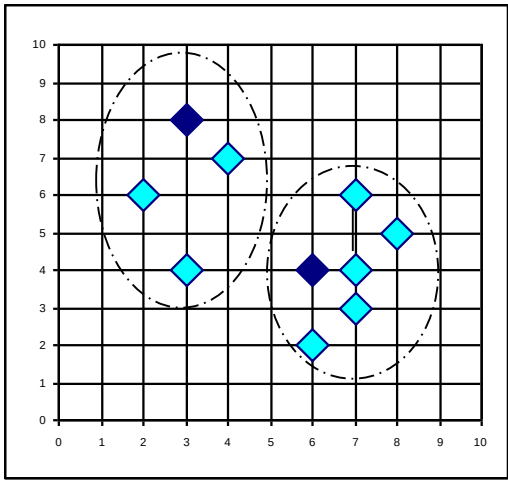
# k-Medoids algorithm (PAM)



Arbitrary choose k object as initial medoids

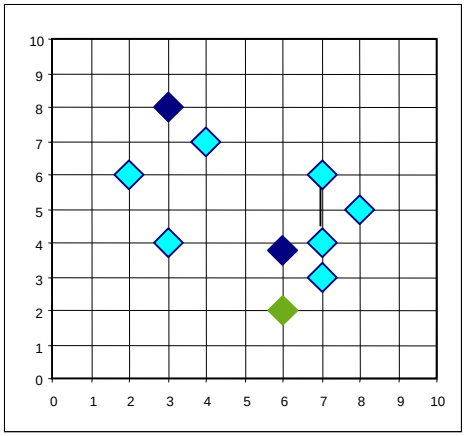


Assign each remaining object to nearest medoids

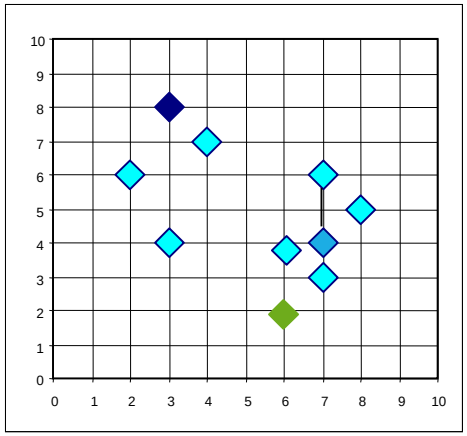


Total Cost = 20

Randomly select a nonmedoid object, $O_{ramdom}$



Compute total cost of swapping



Total Cost = 26

Swapping O and $O_{ramdom}$

If quality is improved.

K=2

**Do loop**

**Until no change**

# روش‌های خوشه‌بندی

روش‌های افراز Partitioning

**روش‌های سلسله‌مراتبی Hierarchical**

روش‌های مبتنی بر چگالی Density

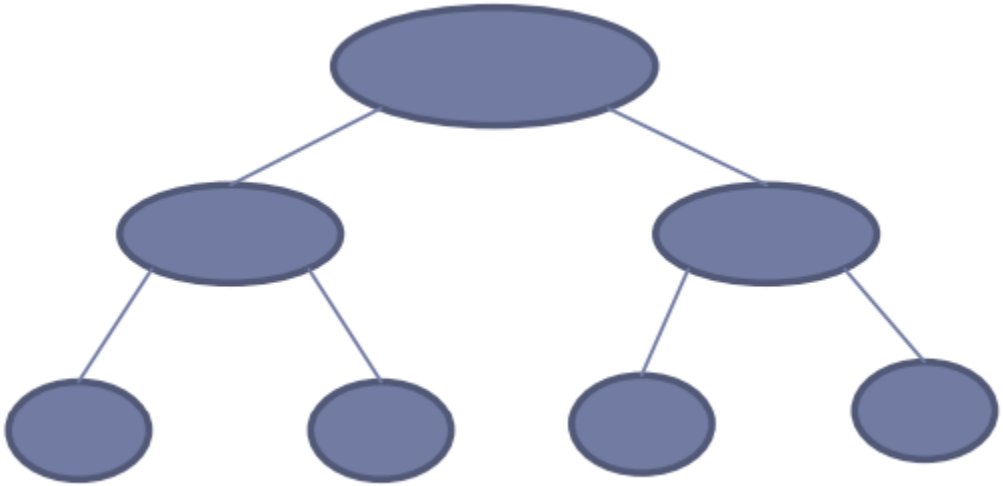
روش‌های مبتنی بر گرید Grid

# روش‌های سلسله‌مراتبی

گاهی اوقات می‌خواهیم خوشه‌بندی را بصورت سلسله‌مراتبی انجام دهیم

روش خوشه‌بندی سلسله‌مراتبی، گروه‌بندی اشیاء داده‌ای بصورت درختی از خوشه‌هاست

نمایش داده به فرم سلسله‌مراتبی برای خلاصه‌سازی و تصویرسازی داده مفید است



# روش‌های سلسله‌مراتبی

در روش خوشه‌بندی سلسله مراتبی

انتخاب نقطه شکستن یا ادغام بسیار مهم است

- زیرا زمانی که یک گروه از داده‌ها ادغام یا شکسته می‌شوند، فرایند در مرحله بعدی روی خوشه‌های ساخته شده مرحله قبلی انجام می‌شود

این امکان وجود ندارد که کاری که در مرحله قبل انجام شده است را برگردانیم یا اینکه داده‌های بین خوشه‌ها را جابجا کنیم

بنابراین اگر تصمیم‌گیری ادغام و شکستن به درستی انجام نشود، خوشه‌هایی با کیفیت پایین خواهیم داشت

# روش‌های سلسله‌مراتبی

روش سلسله مراتبی به دو دسته تقسیم‌بندی می‌شوند

- Agglomerative Hierarchical Clustering (AHC)

  خوشه‌بندی سلسله مراتبی تجمعی- روش پایین به بالا

- Divisive Hierarchical Clustering (DHC)

  خوشه‌بندی سلسله مراتبی تقسیمی- روش بالا به پایین

# روش خوشه‌بندی سلسله مراتبی تجمعی

در روش خوشه‌بندی سلسله مراتبی تجمعی (AHC)

- هر شیء خوشه‌ی خاص خودش را دارد
- بصورت تکراری خوشه‌ها با هم ادغام می‌شوند و خوشه‌های بزرگتر را می سازند
- تا زمانی که همه داده‌ها در یک خوشه قرار گیرند یا شرط خاتمه برآورده شود
- برای مرحله ادغام، الگوریتم دو خوشه‌ای را پیدا می‌کند که نزدیکترین به یکدیگر هستند (بر اساس معیار فاصله) و آنها را با هم ادغام می‌کند تا یک خوشه بسازد
- در هر تکرار، دو خوشه ترکیب می‌شوند که هر خوشه حداقل یک عضو دارد
- به n تکرار نیاز دارد

# روش خوشه‌بندی سلسله مراتبی تجمعی

## الگوریتم AHC

1. Compute the distance matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the distance matrix
6. **Until** only a single cluster remains

# روش خوشه‌بندی سلسله مراتبی تقسیمی

روش خوشه‌بندی سلسله مراتبی تقسیمی DHC

- با قرار دادن همه اشیاء در یک خوشه شروع می‌شود که ریشه سلسله مراتب است
- خوشه‌ی ریشه را به چندین زیرخوشه کوچکتر می‌شکند و بصورت بازگشتی این خوشه‌ها را به خوشه‌های کوچکتر تقسیم می‌کند
- پروسه تقسیم ادامه می‌یابد تا زمانی که هر خوشه در پایین ترین سطح، شامل یک شیء داده‌ای باشد یا هر شیء داده‌ای به اندازه کافی به بقیه شبیه باشد

در هر دو روش، کاربر، تعداد خوشه‌های مورد نظر را بعنوان شرط خروج مشخص می‌کند

# روش‌های سلسله‌مراتبی-مثال

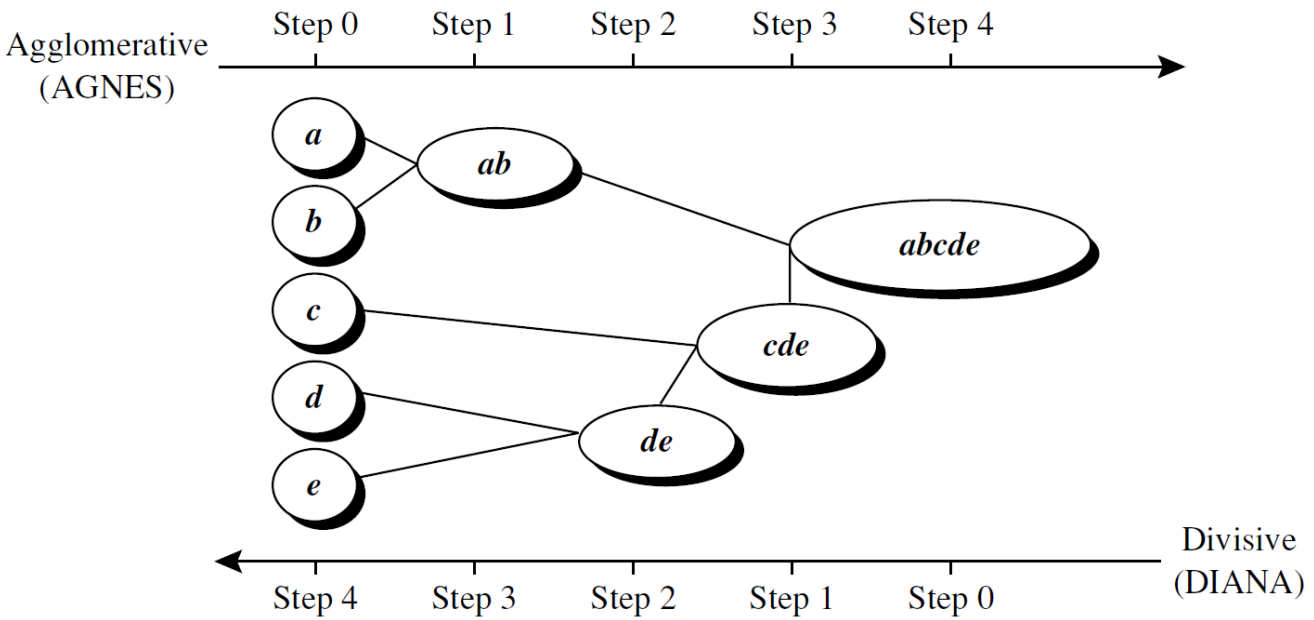
روش خوشه‌بندی سلسله مراتبی تجمعی

- AGNES (AGglomerative NESting)

روش خوشه‌بندی سلسله مراتبی تقسیمی

- DIANA (DIvisive ANAlysis)

$$\{a, b, c, d, e\}$$



# روش‌های سلسله‌مراتبی

## AGNES

- خوشه C1 و C2 ادغام می‌شوند: اگر یک داده در C1 و یک داده در C2 باشند، به شرطی با هم ادغام می‌شوند که کمترین فاصله اقلیدسی بین این دو داده نسبت به هر دو داده دیگری در خوشه‌ها متفاوت باشند

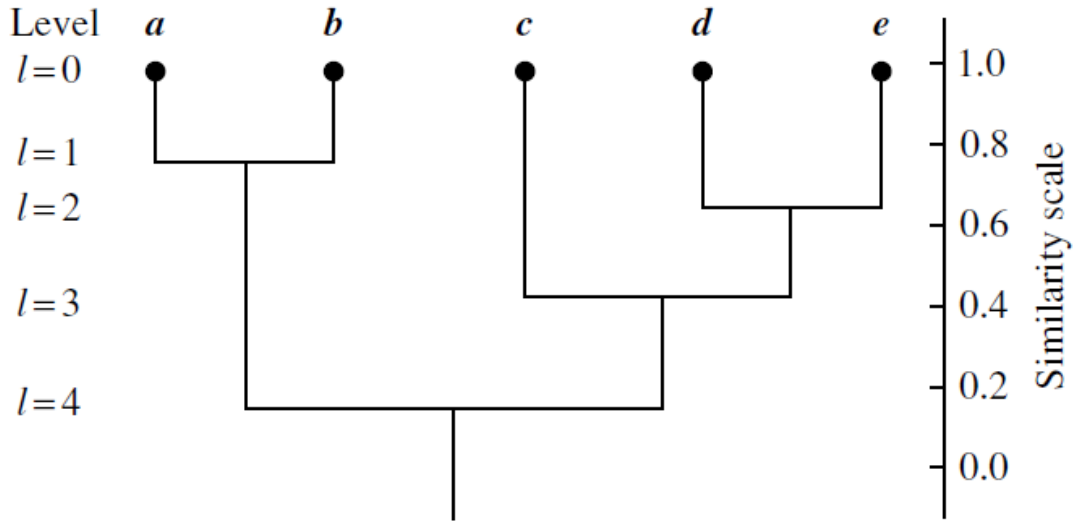
## DIANA

- یک خوشه شکسته می‌شود اگر حداکثر فاصله اقلیدسی بین نزدیکترین داده‌های همسایه در خوشه باشد

# روش سلسله مراتبی

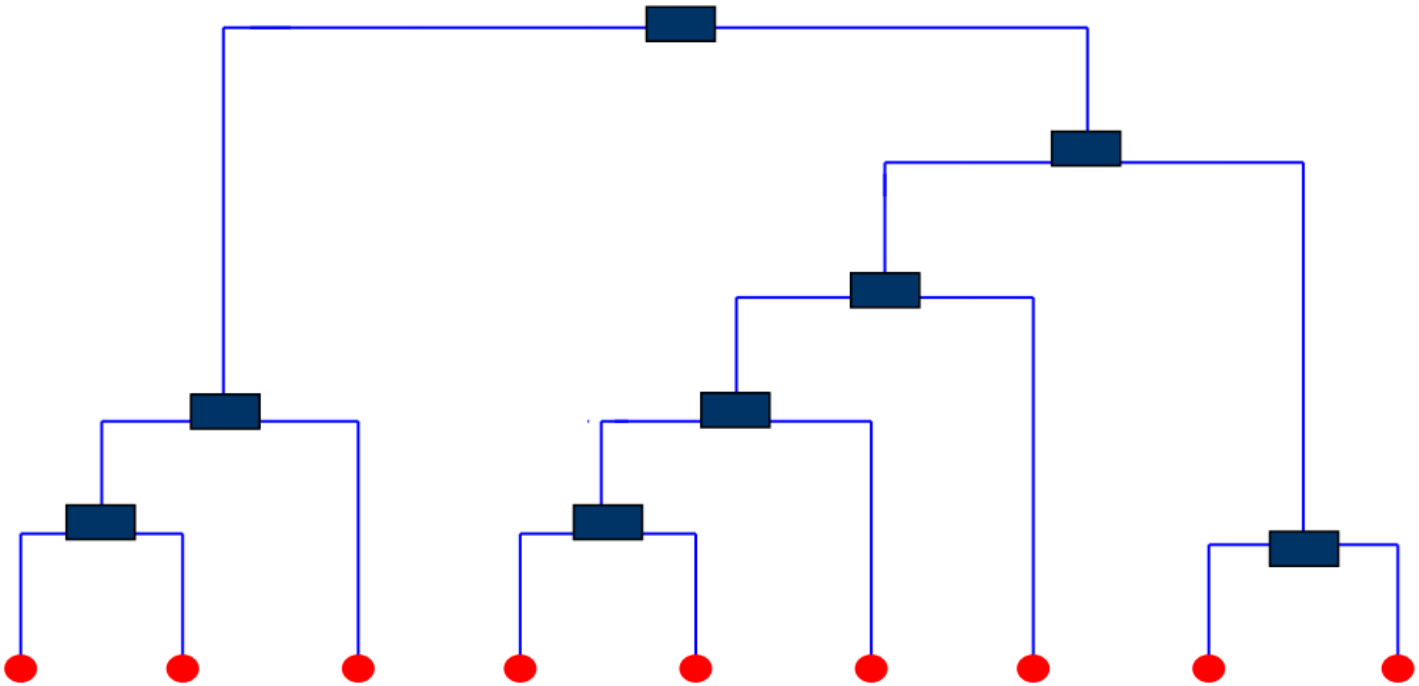
ساختار درختی روش سلسله مراتبی معمولا Dendrogram نامیده می‌شود

Dendrogram نشان می‌دهد که چطور داده‌ها قدم به قدم در روش تجمعی با هم ترکیب شده یا در روش تقسیمی تقسیم‌بندی می‌شوند



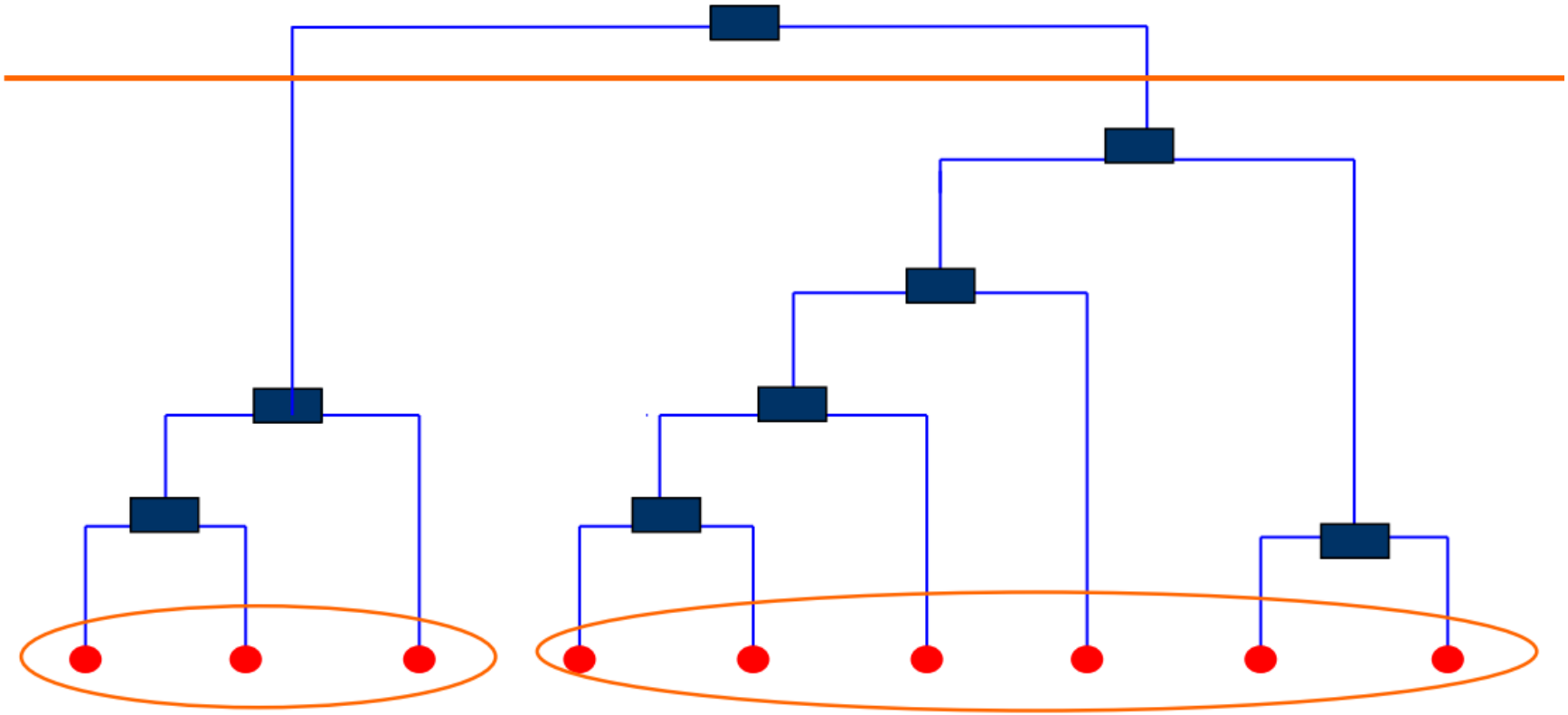
# روش سلسله مراتبی

در Dendrogram، هر گره (Node) در درخت، یک خوشه است، هر گره برگ یک خوشه تکی (تک عنصری) است.



# روش سلسله مراتبی

خوشه‌بندی اشیاء داده‌ای، بوسیله قطع کردن Dendrogram در سطح مورد نظر بدست می‌آید، سپس هر جزء به هم وصل شده یک خوشه را شکل می‌دهد

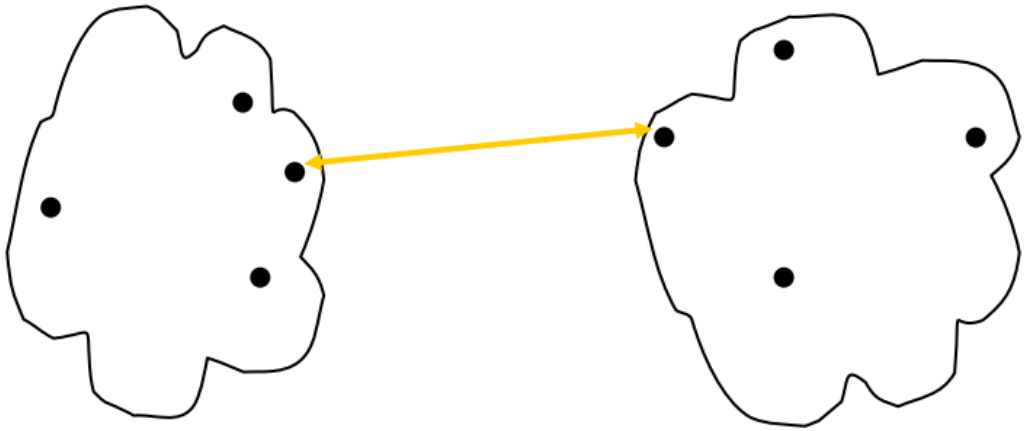


# معیار فاصله در روش سلسله مراتبی

فاصله حداقل

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

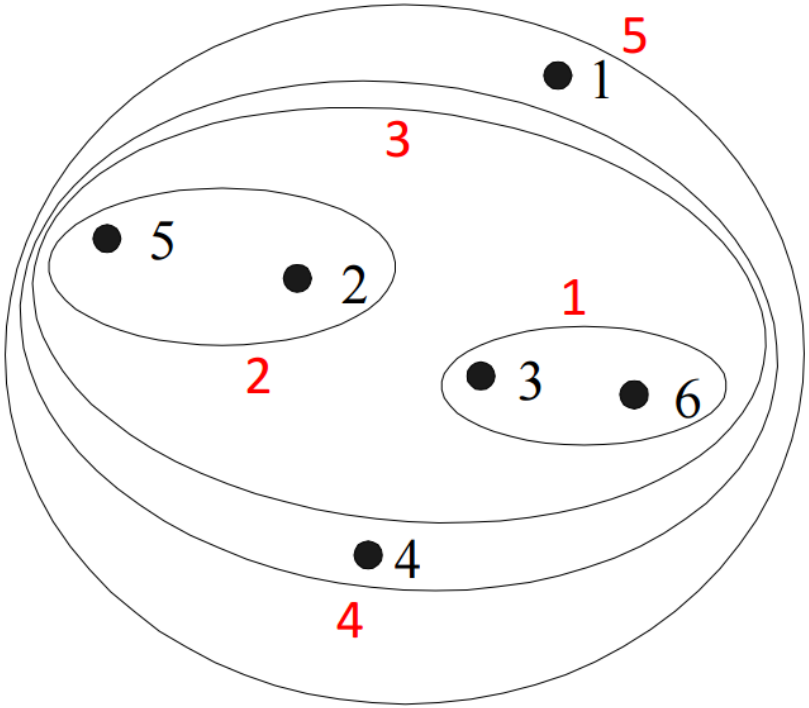
**|p-p'|** is the distance between two objects p and p'



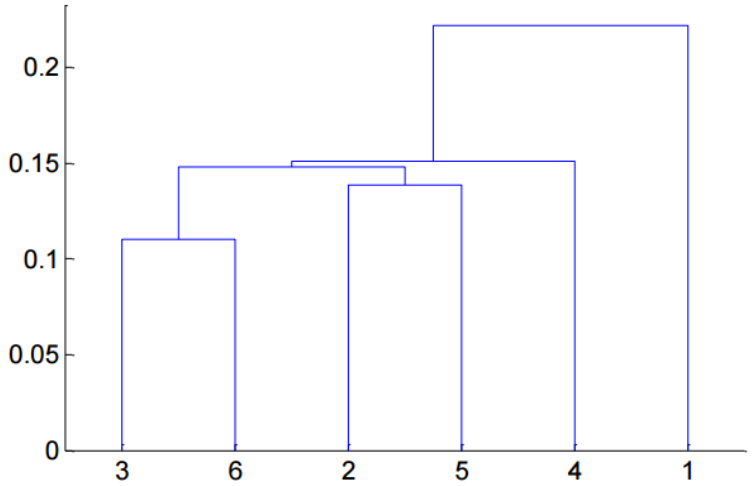
**single-linkage algorithm**

# معیار فاصله در روش سلسله مراتبی

فاصله حداقل




Nested Clusters



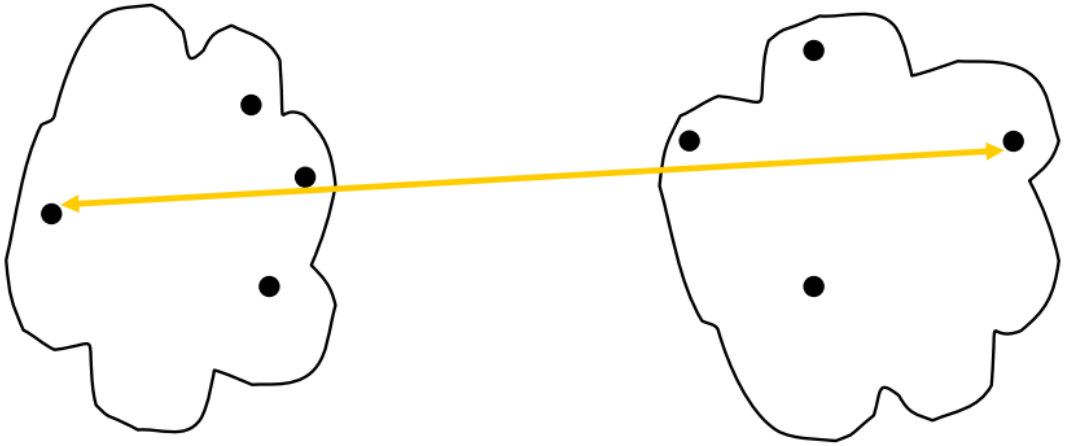

Dendrogram

# معیار فاصله در روش سلسله مراتبی

فاصله حداکثر

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

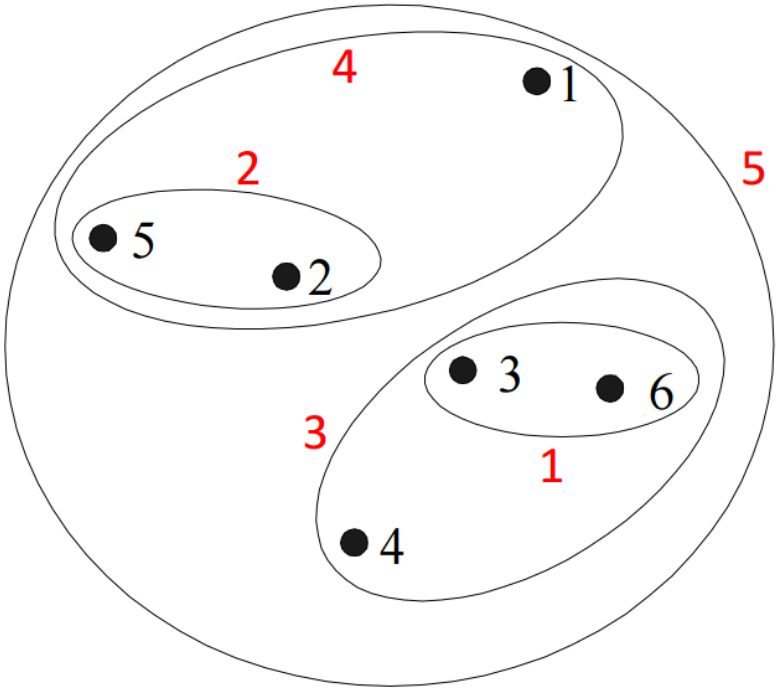
**|p-p'|** is the distance between two objects p and p'



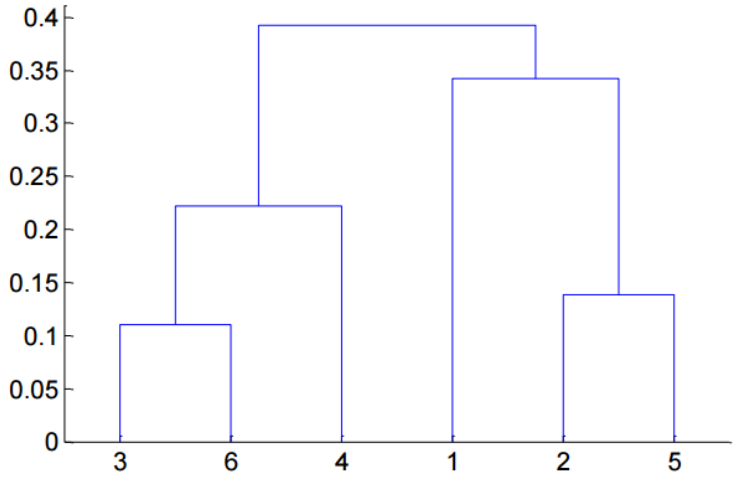
**complete-linkage algorithm**

# معیار فاصله در روش سلسله مراتبی

فاصله حداکثر




Nested Clusters



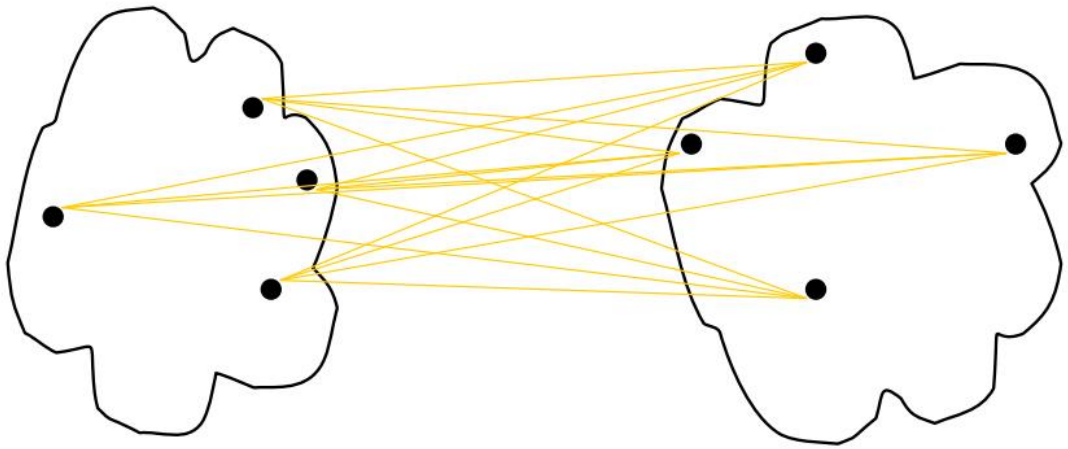

Dendrogram

# معیار فاصله در روش سلسله مراتبی

فاصله میانگین

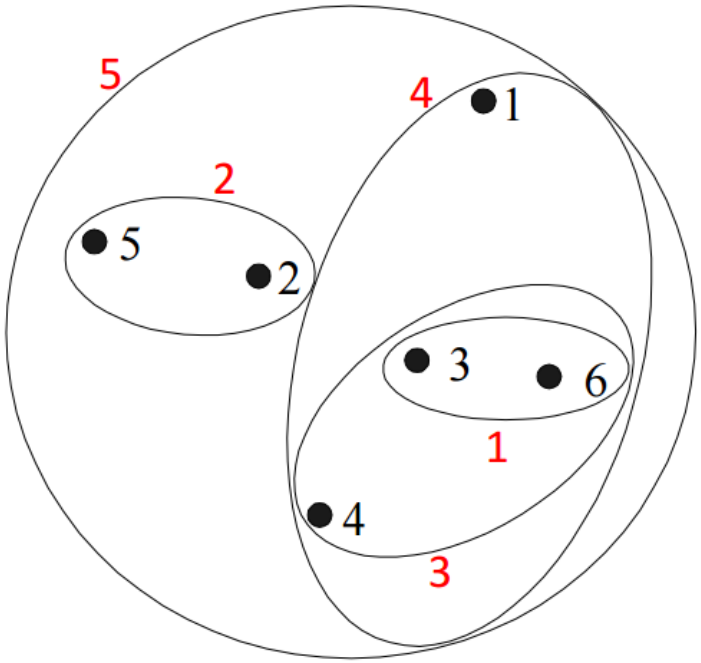
$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

- **|p-p'|** is the distance between two objects p and p'
- $\mathbf{n_i}$ and $\mathbf{n_j}$ are the number of objects in cluster $C_i$ and $C_j$ respectively

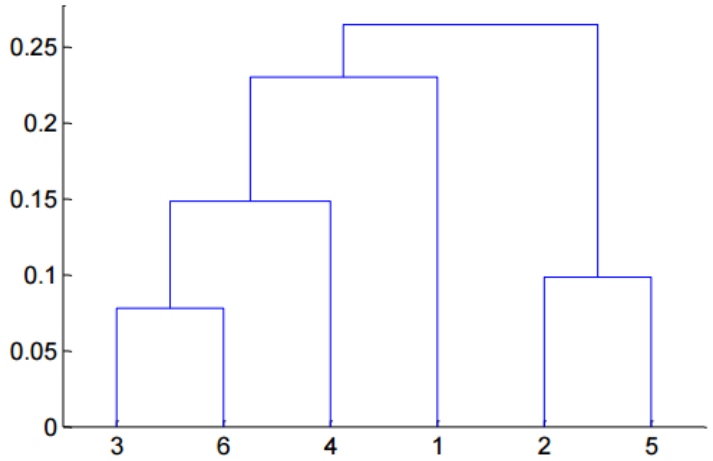


# معیار فاصله در روش سلسله مراتبی

فاصله میانگین




Nested Clusters



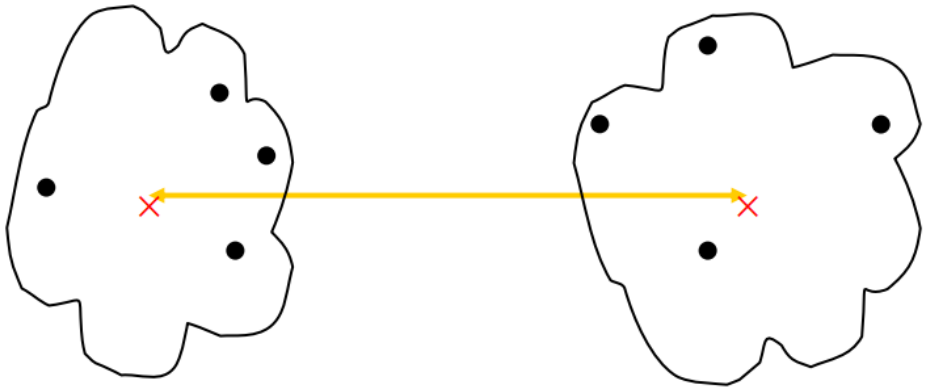

Dendrogram

# معیار فاصله در روش سلسله مراتبی

فاصله مرکز

$$d_{mean}(C_i, C_j) = | m_i - m_j |$$

$m_i$ and $m_j$ are the means for cluster $C_i$ and $C_j$ respectively



# معیارهای اتصال

معیارهای اتصال linkage measures

**Minimum distance:** $dist_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\}$

**Maximum distance:** $dist_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\}$

**Mean distance:** $dist_{mean}(C_i, C_j) = |m_i - m_j|$

**Average distance:** $dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$

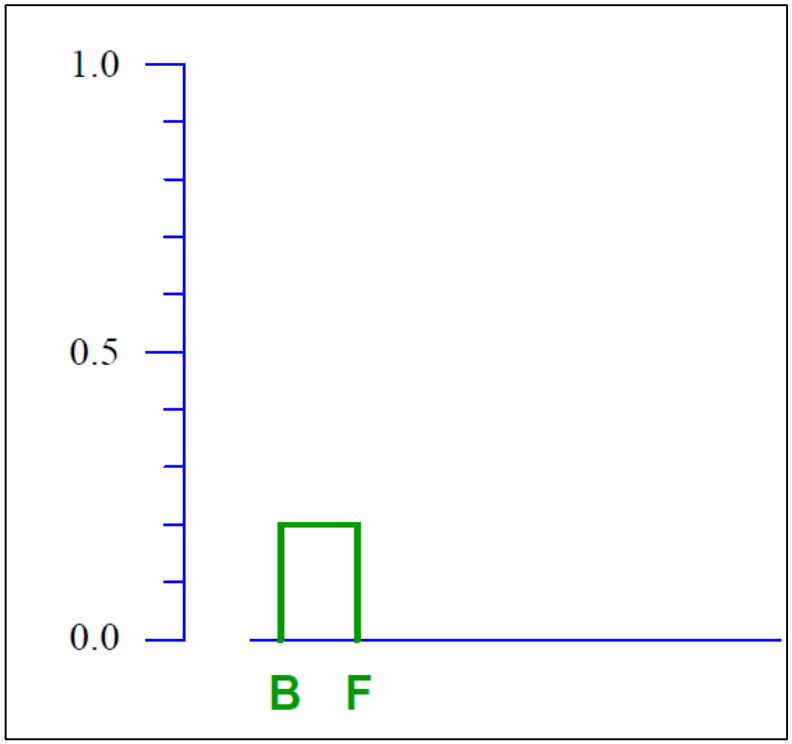
# مثال: روش سلسله مراتبی

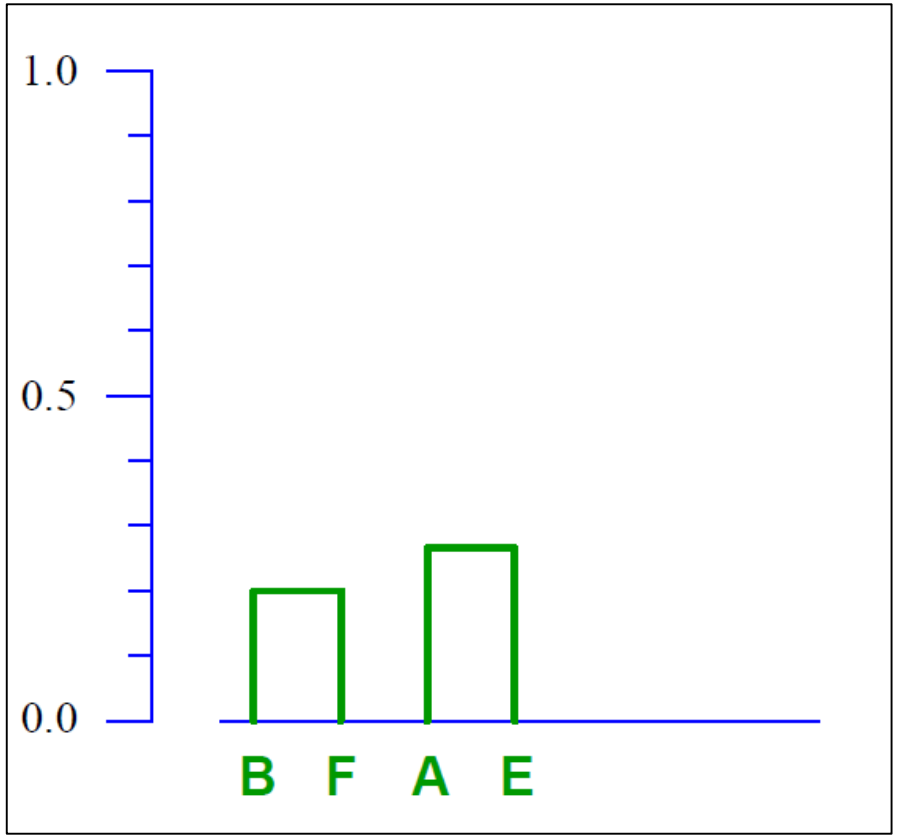
| samples | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 0 | 0.5000 | 0.4286 | 1.0000 | 0.2500 | 0.6250 | 0.3750 |
| B | 0.5000 | 0 | 0.7143 | 0.8333 | 0.6667 | 0.2000 | 0.7778 |
| C | 0.4286 | 0.7143 | 0 | 1.0000 | 0.4286 | 0.6667 | 0.3333 |
| D | 1.0000 | 0.8333 | 1.0000 | 0 | 1.0000 | 0.8000 | 0.8571 |
| E | 0.2500 | 0.6667 | 0.4286 | 1.0000 | 0 | 0.7778 | 0.3750 |
| F | 0.6250 | 0.2000 | 0.6667 | 0.8000 | 0.7778 | 0 | 0.7500 |
| G | 0.3750 | 0.7778 | 0.3333 | 0.8571 | 0.3750 | 0.7500 | 0 |

# مثال: روش سلسله مراتبی، ادامه

| samples | A | (B,F) | C | D | E | G |
|---|---|---|---|---|---|---|
| A | 0 | 0.6250 | 0.4286 | 1.0000 | 0.2500 | 0.3750 |
| (B,F) | 0.6250 | 0 | 0.7143 | 0.8333 | 0.7778 | 0.7778 |
| C | 0.4286 | 0.7143 | 0 | 1.0000 | 0.4286 | 0.3333 |
| D | 1.0000 | 0.8333 | 1.0000 | 0 | 1.0000 | 0.8571 |
| E | 0.2500 | 0.7778 | 0.4286 | 1.0000 | 0 | 0.3750 |
| G | 0.3750 | 0.7778 | 0.3333 | 0.8571 | 0.3750 | 0 |

# مثال: روش سلسله مراتبی، ادامه



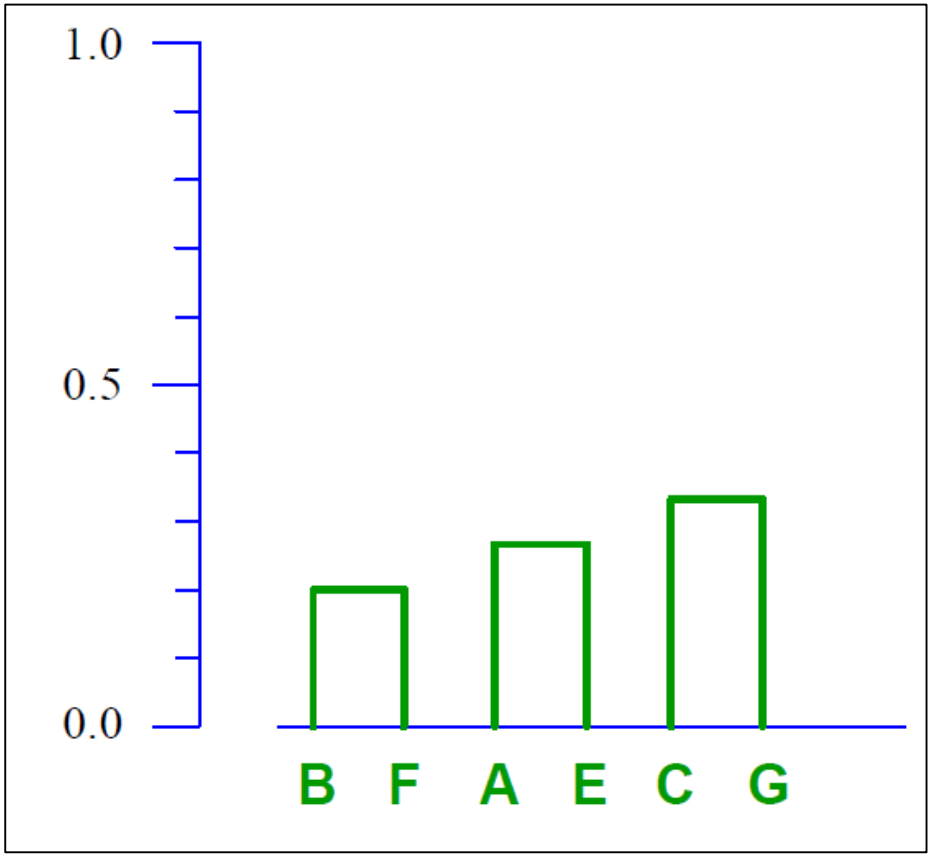
# مثال: روش سلسله مراتبی، ادامه

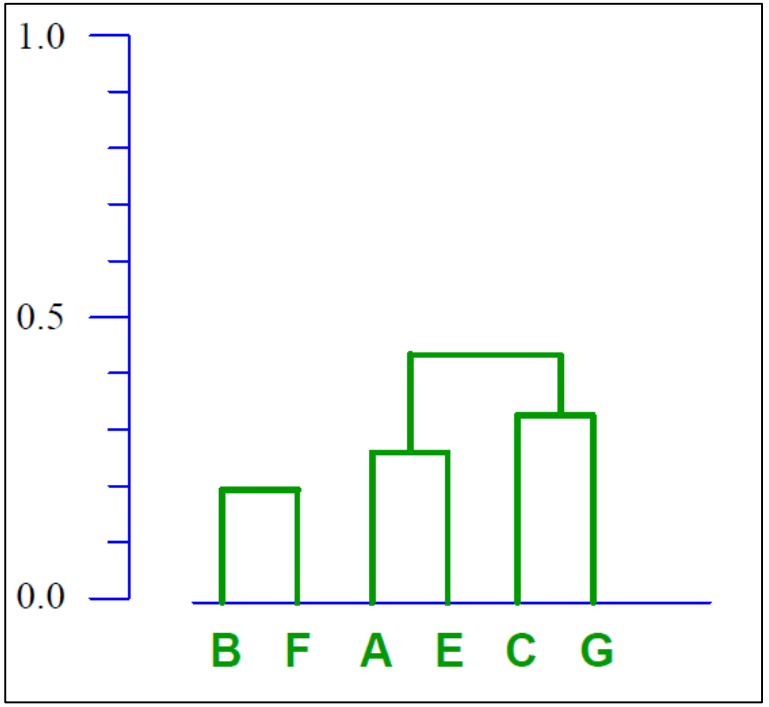
| samples | (A,E) | (B,F) | C | D | G |
|---|---|---|---|---|---|
| (A,E) | 0 | 0.7778 | 0.4286 | 1.0000 | 0.3750 |
| (B,F) | 0.7778 | 0 | 0.7143 | 0.8333 | 0.7778 |
| C | 0.4286 | 0.7143 | 0 | 1.0000 | 0.3333 |
| D | 1.0000 | 0.8333 | 1.0000 | 0 | 0.8571 |
| G | 0.3750 | 0.7778 | 0.3333 | 0.8571 | 0 |

# مثال: روش سلسله مراتبی، ادامه

| samples | (A,E) | (B,F) | (C,G) | D |
|---|---|---|---|---|
| (A,E) | 0 | 0.7778 | 0.4286 | 1.0000 |
| (B,F) | 0.7778 | 0 | 0.7778 | 0.8333 |
| (C,G) | 0.4286 | 0.7778 | 0 | 1.0000 |
| D | 1.0000 | 0.8333 | 1.0000 | 0 |

# مثال: روش سلسله مراتبی، ادامه



# مثال: روش سلسله مراتبی، ادامه

| samples | (A,E,C,G) | (B,F) | D |
|---|---|---|---|
| (A,E,C,G) | 0 | 0.7778 | 1.0000 |
| (B,F) | 0.7778 | 0 | 0.8333 |
| D | 1.0000 | 0.8333 | 0 |

| samples | (A,E,C,G,B,F) | D |
|---|---|---|
| (A,E,C,G,B,F) | 0 | 1.0000 |
| D | 1.0000 | 0 |

# روش‌های خوشه‌بندی

روش‌های افراز Partitioning

روش‌های سلسله‌مراتبی Hierarchical

**روش‌های مبتنی بر چگالی Density**

روش‌های مبتنی بر گرید Grid

# روش‌های مبتنی بر چگالی

خوشه‌بندی بر اساس چگالی است

ویژگی‌های اصلی

- کشف خوشه‌هایی با شکل دلخواه
- مدیریت کردن noise

انواع الگوریتم‌های خوشه‌بندی بر اساس چگالی عبارتند از:

- DBSCAN: Ester, et al. (KDD'96)
- OPTICS: Ankerst, et al (SIGMOD'99).
- DENCLUE: Hinneburg & D. Keim (KDD'98)
- CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

# روش‌های مبتنی بر چگالی

امکان کشف خوشه‌هایی با شکل دلخواه

# روش‌های مبتنی بر چگالی

دو پارامتر مهم در روش‌های مبتنی بر چگالی:

Eps: حداکثر شعاع همسایگی

MinPts: حداقل تعداد نقاط داده‌ای که در همسایگی با شعاع Eps وجود دارد (Eps-neighbourhood)

$N_{Eps}(p)$: {q belongs to D | dist(p,q) ≤ Eps}

# روش‌های مبتنی بر چگالی-مفاهیم اولیه

## نقطه هسته‌ای یا Core point

- نقطه‌ای که تعداد نقاطی که در همسایگی آن با شعاع Eps وجود دارد از حداقل تعداد نقاط (*MinPts*) بیشتر باشد

$$|N_{Eps}(q)| \geq MinPts$$

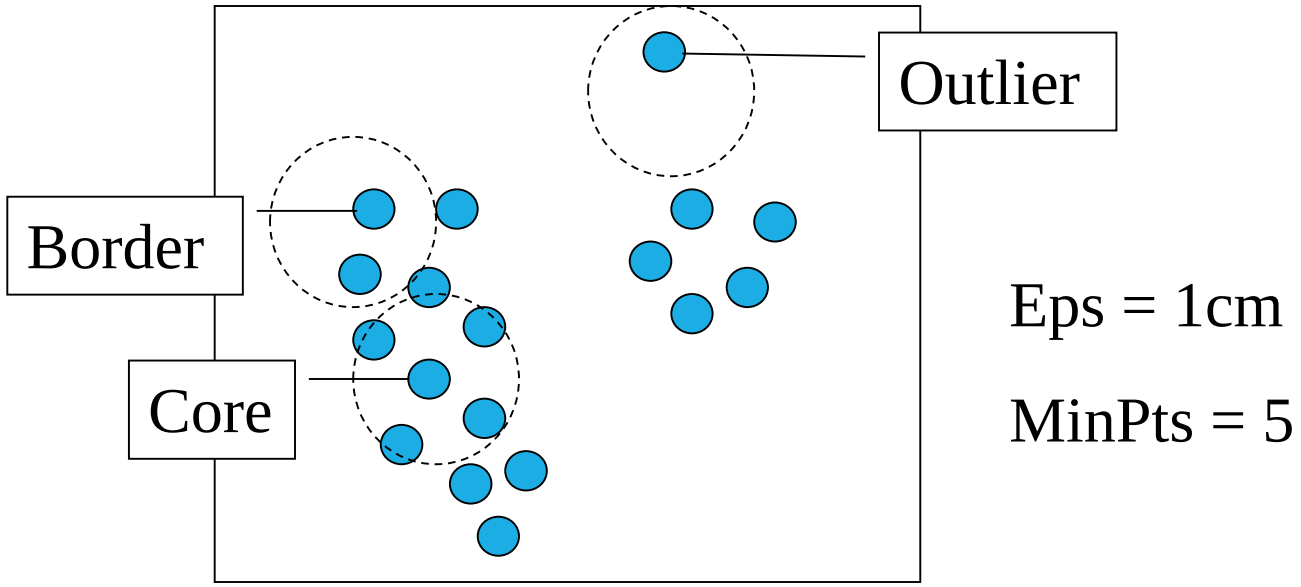
## نقطه حاشیه‌ای یا Border point

- نقطه‌ای که تعداد نقاطی که در همسایگی آن با شعاع Eps وجود دارد از حداقل تعداد نقاط (*MinPts*) کمتر باشد

## نقطه نویز Noise point

- نقطه‌ای که نتواند به صورت هسته‌ای یا حاشیه‌ای در نظر گرفته شود، نویز در نظر گرفته می‌شود

# روش‌های مبتنی بر چگالی-مفاهیم اولیه

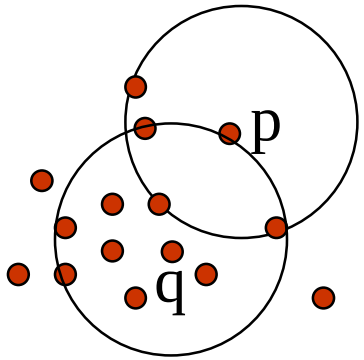


# روش‌های مبتنی بر چگالی-مفاهیم اولیه

Directly density-reachable:

نقطه p، Directly density-reachable از نقطه q است اگر دو شرط زیر را داشته باشد

- p متعلق به مجموعه همسایه q باشد
- q نقطه هسته‌ای باشد



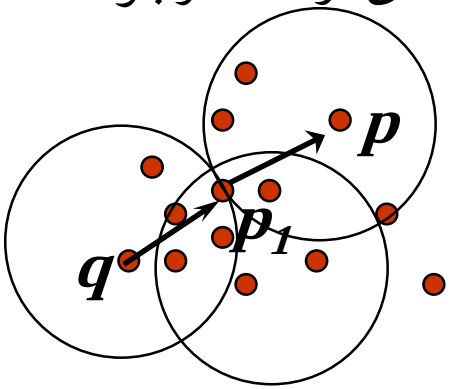

MinPts = 5

Eps = 1 cm

# روش‌های مبتنی بر چگالی-مفاهیم اولیه

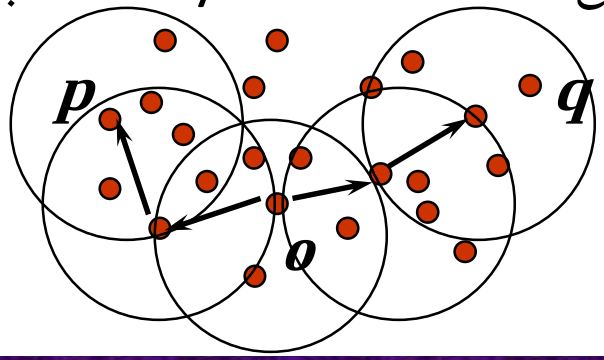
## Density-reachable:

- نقطه p، از نقطه q، density-reachable است، اگر مجموعه‌ای از نقاط وجود داشته باشد که Directly density-reachable باشند




## Density-connected:

- نقطه p، density-connected به نقطه q است، اگر نقطه‌ای مانند o وجود داشته باشد که p و q از طریق o، density-reachable باشند



# روش‌های مبتنی بر چگالی: روشDBSCAN

DBSCAN:

- Density Based Spatial Clustering of Applications with Noise

این روش بر اساس مفهوم «چگالی خوشه» است

خوشه بر اساس حداکثر مجموعه نقاطی که به یکدیگر density-connected هستند، تشکیل می‌شود

این روش می‌تواند خوشه‌هایی با شکل دلخواه را کشف کند

# روش‌های مبتنی بر چگالی: الگوریتم DBSCAN

Arbitrary select a point p

Retrieve all points density-reachable from p with respect to Eps and MinPts.

- If p is a core point, a cluster is formed.
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.

Continue the process until all of the points have been processed.

# روش‌های مبتنی بر چگالی: الگوریتم DBSCAN

**Algorithm: DBSCAN:** a density-based clustering algorithm.

**Input:**

- $D$: a data set containing $n$ objects,
- $\epsilon$: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

**Output:** A set of density-based clusters.

# روش‌های مبتنی بر چگالی: الگوریتم DBSCAN

**Method:**

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p′ in N
(9)              if p′ is unvisited
(10)                 mark p′ as visited;
(11)                 if the ε-neighborhood of p′ has at least MinPts points,
                     add those points to N;
(12)             if p′ is not yet a member of any cluster, add p′ to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

# مثال DBSCAN

***Exercise 5:* DBScan**

If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
The distance matrix is the same as the one in Exercise 1. Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to $\sqrt{10}$ ?

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| A1 | 0 | $\sqrt{25}$ | $\sqrt{36}$ | $\sqrt{13}$ | $\sqrt{50}$ | $\sqrt{52}$ | $\sqrt{65}$ | $\sqrt{5}$ |
| A2 | | 0 | $\sqrt{37}$ | $\sqrt{18}$ | $\sqrt{25}$ | $\sqrt{17}$ | $\sqrt{10}$ | $\sqrt{20}$ |
| A3 | | | 0 | $\sqrt{25}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{53}$ | $\sqrt{41}$ |
| A4 | | | | 0 | $\sqrt{13}$ | $\sqrt{17}$ | $\sqrt{52}$ | $\sqrt{2}$ |
| A5 | | | | | 0 | $\sqrt{2}$ | $\sqrt{45}$ | $\sqrt{25}$ |
| A6 | | | | | | 0 | $\sqrt{29}$ | $\sqrt{29}$ |
| A7 | | | | | | | 0 | $\sqrt{58}$ |
| A8 | | | | | | | | 0 |

# مثال DBSCAN، ادامه

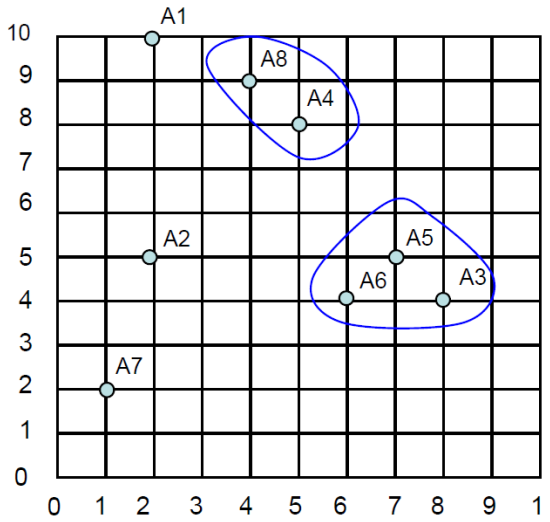
***Solution:***

What is the Epsilon neighborhood of each point?

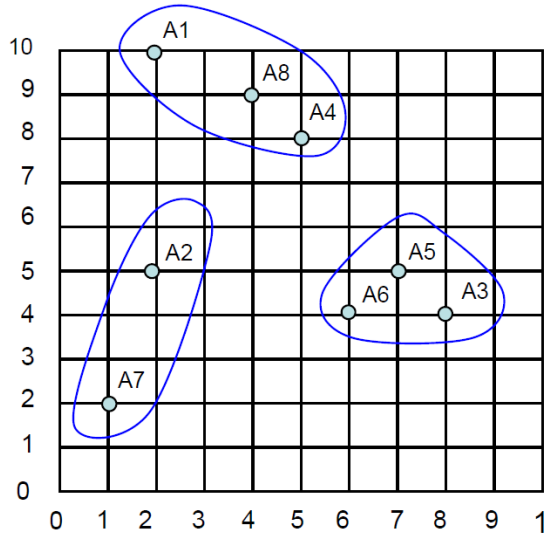
$N_2(A1)=\{\}$; $N_2(A2)=\{\}$; $N_2(A3)=\{A5, A6\}$; $N_2(A4)=\{A8\}$; $N_2(A5)=\{A3, A6\}$;
$N_2(A6)=\{A3, A5\}$; $N_2(A7)=\{\}$; $N_2(A8)=\{A4\}$

So A1, A2, and A7 are outliers, while we have two clusters C1={A4, A8} and C2={A3, A5, A6}

If Epsilon is $\sqrt{10}$ then the neighborhood of some points will increase:
A1 would join the cluster C1 and A2 would joint with A7 to form cluster C3={A2, A7}.



Epsilon = 2



Epsilon = $\sqrt{10}$

# روش‌های خوشه‌بندی

روش‌های افراز Partitioning

روش‌های سلسله‌مراتبی Hierarchical

روش‌های مبتنی بر چگالی Density

**روش‌های مبتنی بر گرید Grid**

# روش‌های مبتنی بر گرید

استفاده از ساختار داده سلولی multi-resolution

متدهای مبتنی بر گرید، فضای اشیاء را به تعداد محدودی سلول که در ساختار گرید شکل داده شده‌اند، فرموله می‌کند

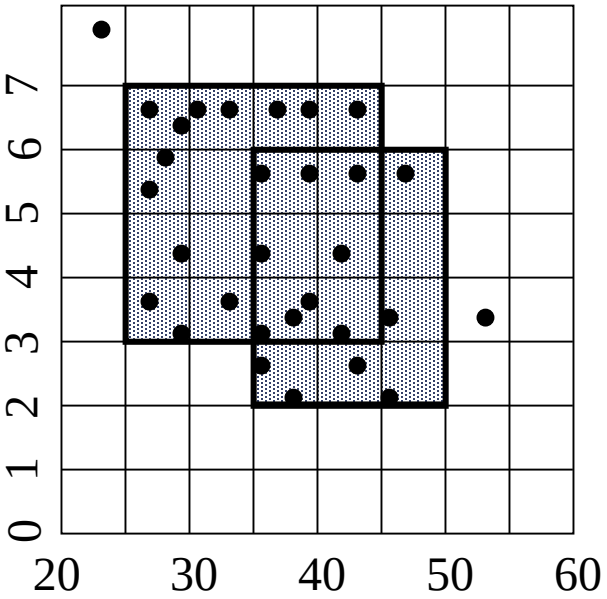
همه عملیات خوشه‌بندی در ساختار گرید اجرا می‌شود

مزیت اصلی این روش، زمان سریع پردازش آن است که مستقل از تعداد اشیاء داده‌ای است و تنها به تعداد سلول‌ها در هر بعد از فضای فرموله شده بستگی دارد

الگوریتمهای مبتنی بر گرید شامل:

- STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz
- WaveCluster by Sheikholeslami, Chatterjee, and Zhang
- CLIQUE: Agrawal, et al

# یک روش مبتنی بر گرید



# اندازه‌گیری کیفیت خوشه‌بندی

با توجه به اینکه متدهای خوشه‌بندی مختلفی را یاد گرفتیم، می خواهیم بدانیم چطور می‌توان کیفیت خوشه‌های تولید شده را اندازه‌گیری کرد

روش‌های مختلفی برای اندازه‌گیری کیفیت وجود دارند که به دو دسته زیر طبقه‌بندی می‌شوند

متدهای خارجی(extrinsic methods یا supervised methods):

- هنگامی استفاده می‌شود که ground truth وجود داشته باشد

متدهای داخلی (intrinsic methods یا unsupervised methods):

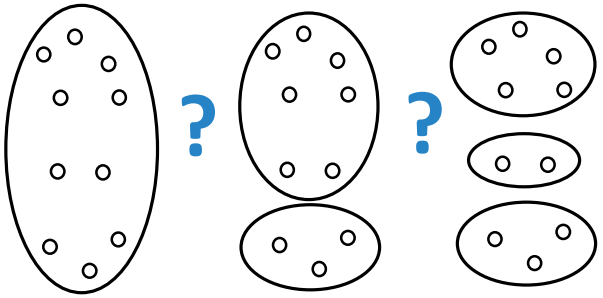
- هنگامی که ground truth وجود ندارد از این روش استفاده می‌شود که خوشه‌ها را از این لحاظ بررسی می‌کند که «چقدر خوب از هم جدا شده‌اند؟»

ground truth را می‌توان برچسب مشخص شده خوشه‌ها بر اساس نظر متخصص انسانی در نظر گرفت

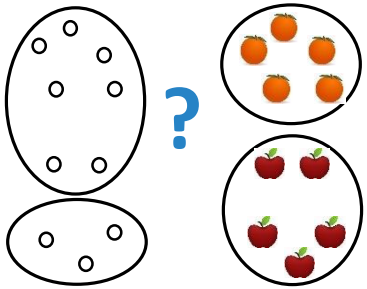
# اندازه‌گیری کیفیت خوشه‌بندی

Cluster Evaluation

- Internal
  - We don’t know anything about the desired labels




- External
  - We have some information about the labels



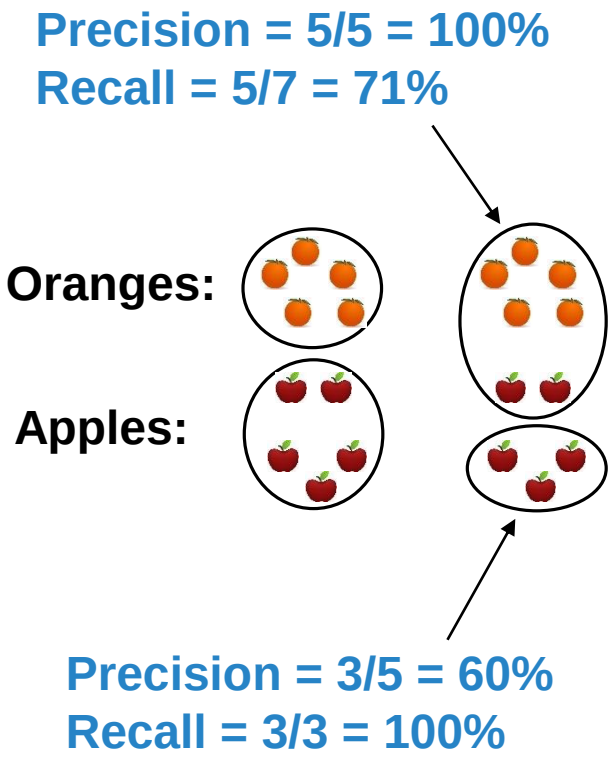
# اندازه‌گیری کیفیت خوشه‌بندی-متدهای خارجی

Precision

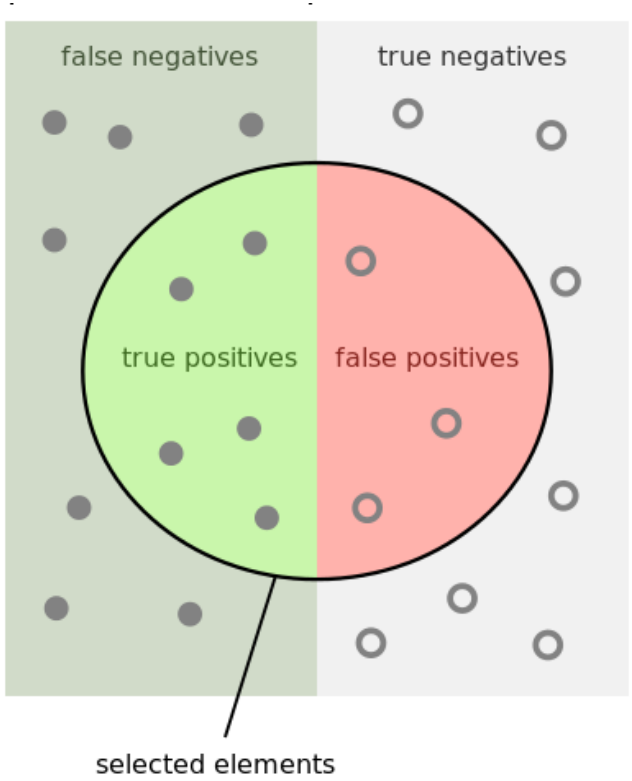
Recall

Precision = specificity (% of selected items that are correct)

Recall = sensitivity (% of correct items that are selected)



# اندازه‌گیری کیفیت خوشه‌بندی-متدهای خارجی




$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}} \qquad \text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

# اندازه‌گیری کیفیت خوشه‌بندی-متدهای خارجی

| | Correct | Not Correct |
|---|---|---|
| Selected | TP | FP |
| Not Selected | FN | TN |

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Precision = specificity (% of selected items that are correct)

$$Recall = \frac{TP}{TP + FN}$$

Recall = sensitivity (% of correct items that are selected)

# اندازه‌گیری کیفیت خوشه‌بندی-یک متد داخلی

## Silhouette Coefficient

Cohesion:

میزان نزدیک بودن عناصر داده‌ای در یک خوشه را بررسی می کند

Separation:

میزان جدا بودن یا متمایز بودن یک خوشه از خوشه‌های دیگر را اندازه گیری می‌کند

# Silhouette coefficient

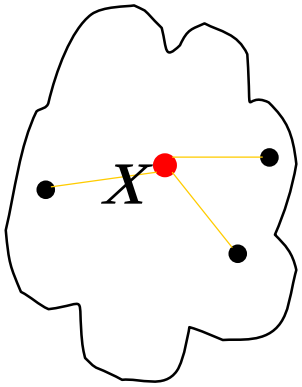
- *Cohesion* $a(x)$: average distance of $x$ to all other vectors in the same cluster.
- *Separation* $b(x)$: average distance of $x$ to the vectors in other clusters. Find the minimum among the clusters.
- *silhouette* $s(x)$:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

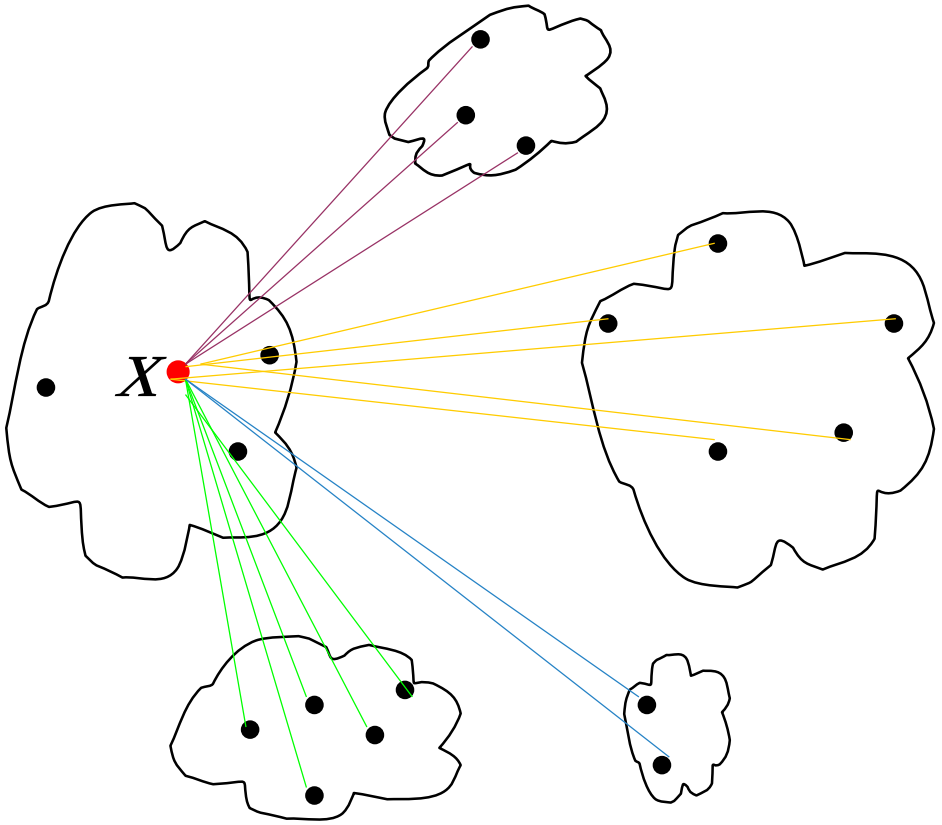
- $s(x)$ = [-1, +1]: -1=bad, 0=indifferent, 1=good
- Silhouette coefficient (SC):

$$SC = \frac{1}{N} \sum_{i=1}^{N} s(x)$$

# Silhouette coefficient




cohesion




separation

$a(x)$: average distance in the cluster

$b$(x): average distances to others clusters, find minimal