

فرادرس

فرادرس
فراترازیک کلاس درس
www.faradars.org

آموزش یادگیری ماشین (Machine Learning) (تئوری - عملی) - بخش دوم

درس ششم: خوشه‌بندی

مدرس:

فرشید شیراوند

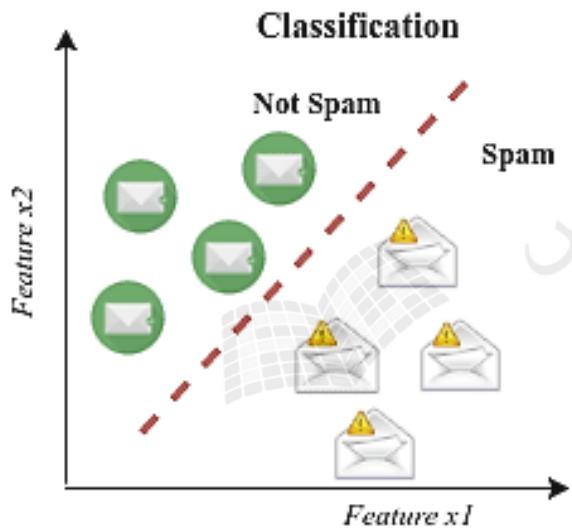
دانشجوی دکترای بیو انفورماتیک

دانشگاه تهران



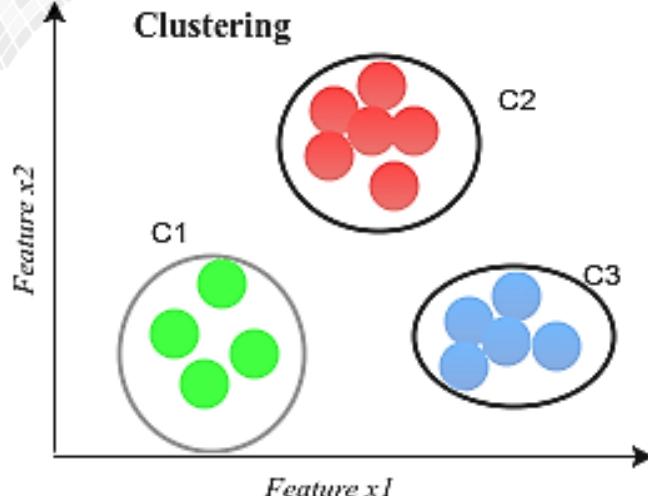
آموزش بدون ناظر

$(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$

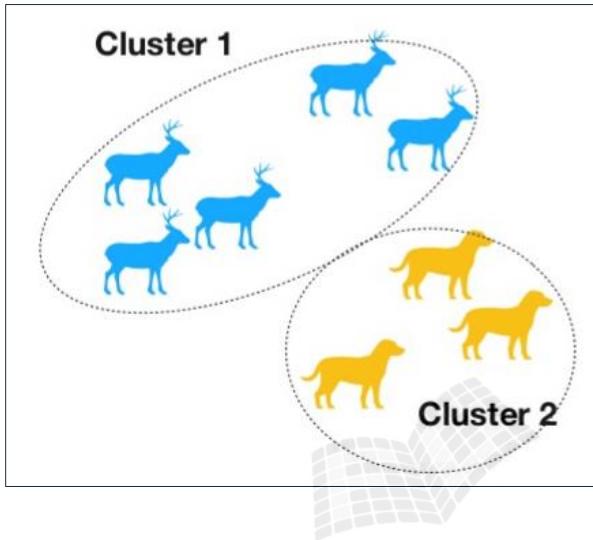


شیوه آموزش: بدون نظارت (unsupervised)

$(x_1), (x_2), \dots, (x_m)$



خوشه‌بندی



انتساب اشیاء به خوشه‌ها به‌طوری که اشیاء یک خوشه:

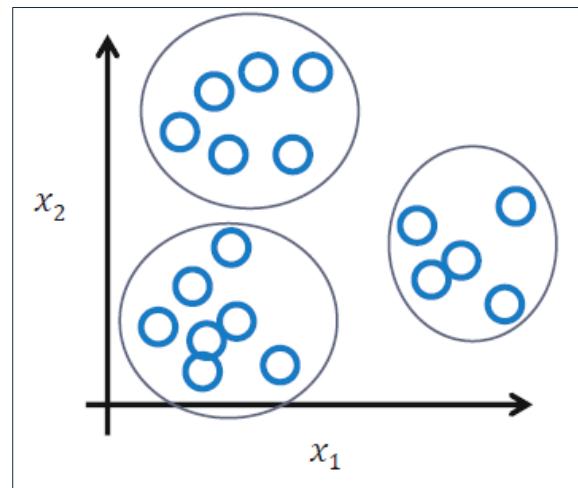


- بیشترین شباهت را با هم داشته باشند.
- بیشترین تفاوت بین خوشه‌های مختلف موجود باشد.

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



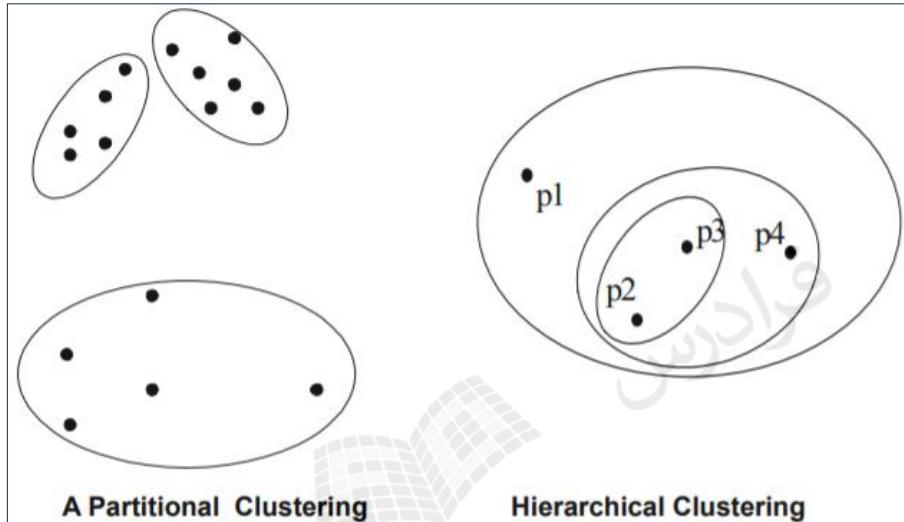
کاربردهای خوشه‌بندی



- بازیابی اطلاعات (کلستر 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 خوشبندی



۱- پارتیشن‌بندی (Partitioning)

۲- سلسله مراتبی (Hierarchical)

خوشه‌بندی افزایی

Partitional Clustering

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

$$N = 5$$

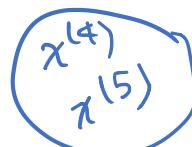
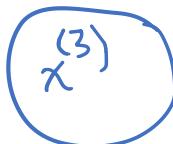
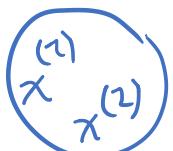
$$\mathcal{X} = \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}\}$$

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

$$K = 3$$

$$\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$$

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



مثال

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

K-means
 $K=2$ مراکن $\begin{cases} C_1 = 3 \\ C_2 = 12 \end{cases}$



2, 3, 4

10, 11, 12, 20, 25, 30

$\begin{cases} C_1 = 3 \\ C_2 = 18 \end{cases}$

2, 3, 4, 10

11, 12, 20, 25, 30

$\begin{cases} C_1 = 4.75 \\ C_2 = 19.6 \end{cases}$

2, 3, 4, 10, 11, 12

20, 25, 30

$\begin{cases} C_1 = 7 \\ C_2 = 25 \end{cases}$

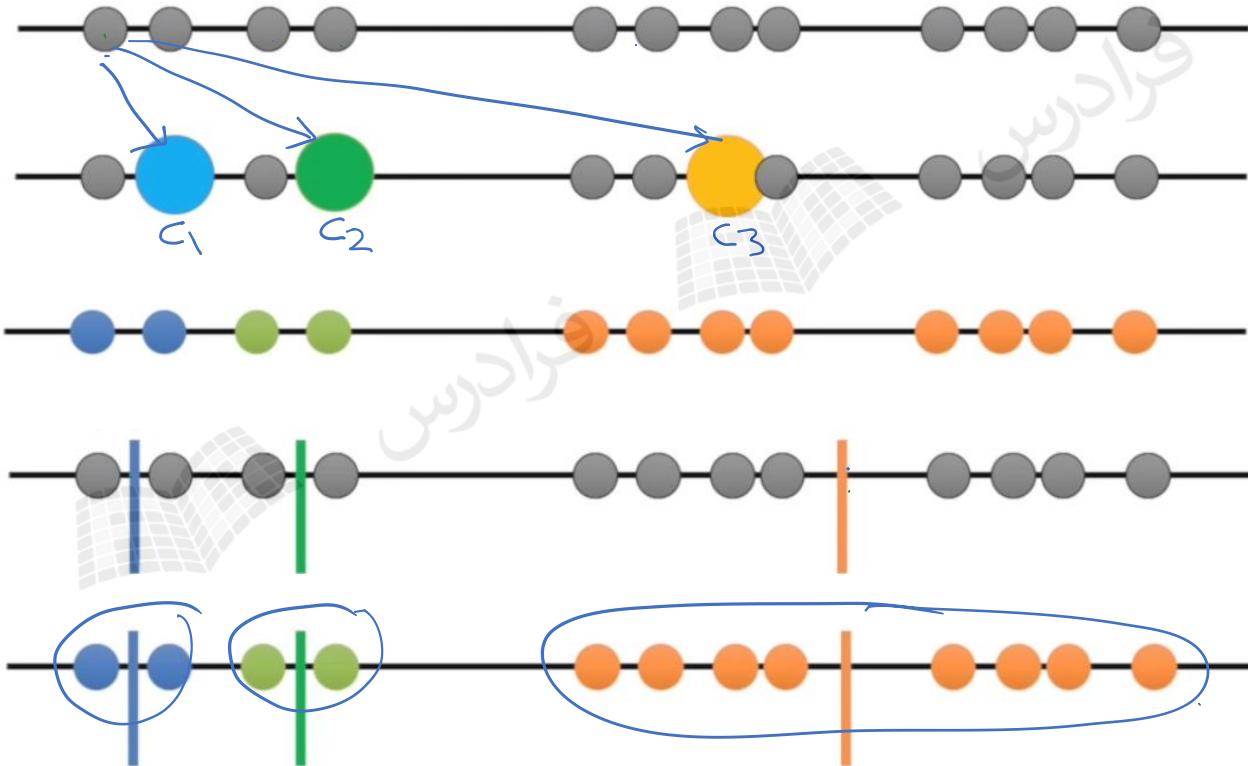
2, 3, 4, 10, 11, 12

20, 25, 30

نیز

مثال

$\kappa = 3$



مثال

$k=2$

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

A
 c_1

B, C, D
 c_2

$$C \rightarrow A : \sqrt{3^2 + 2^2}$$

$$C \rightarrow B : \sqrt{2^2 + 2^2}$$

A, B

C, D

$$c_1 = (1, 1)$$

$$c_2 = \left(\frac{7}{3}, \frac{8}{3}\right)$$

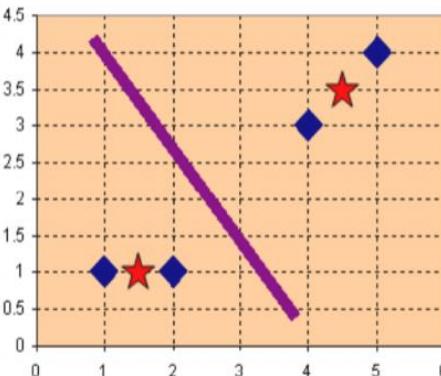
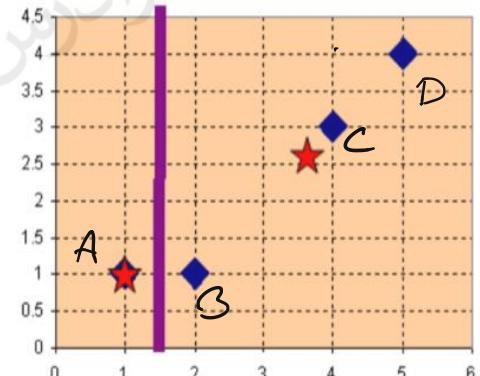
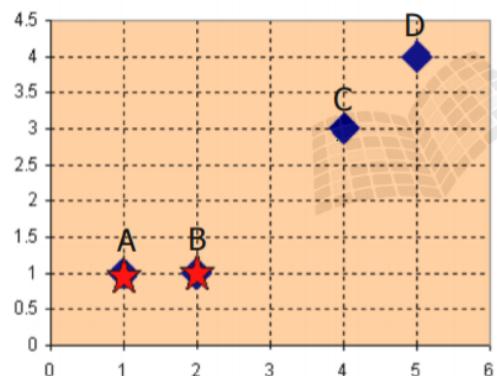
$$c_1 = \left(\frac{3}{2}, 1\right)$$

$$c_2 = \left(\frac{9}{2}, \frac{7}{2}\right)$$

A, B

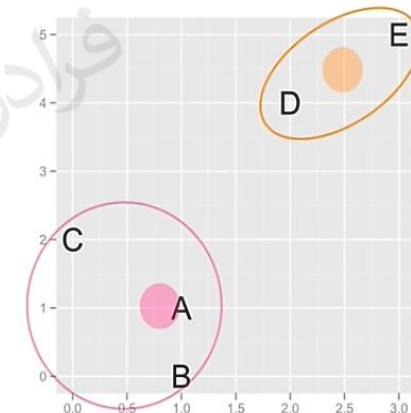
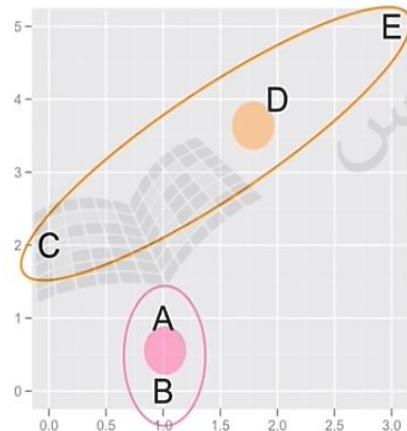
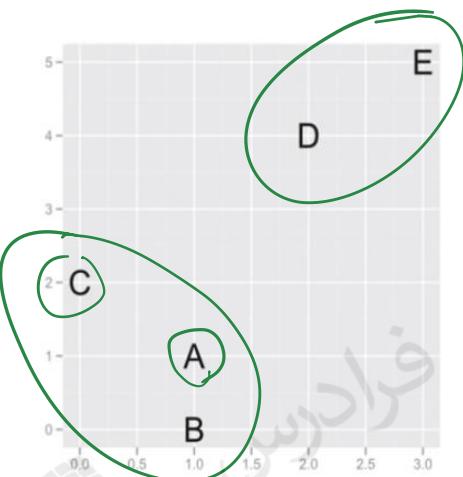
C, D

$$c_{CC} = \frac{1}{2}$$



تمرین

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



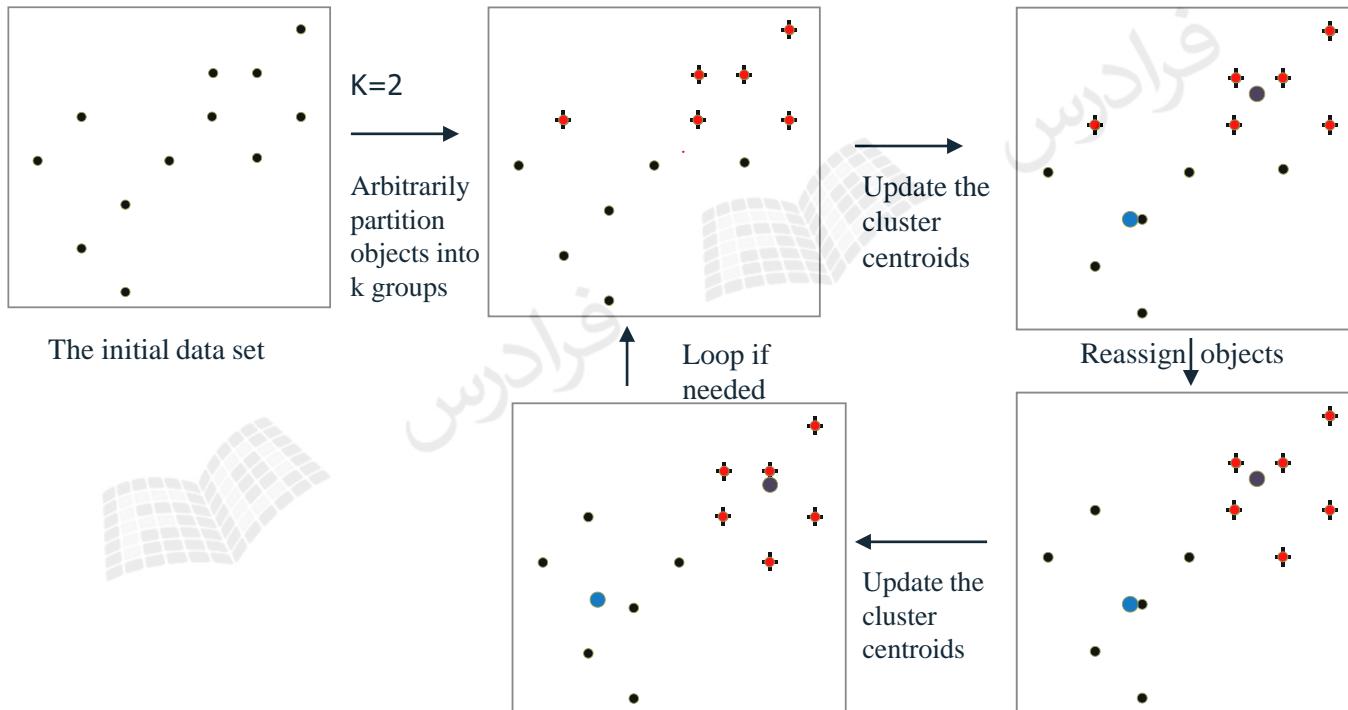
$k=2$

A, B

C, D, E

$1, \frac{1}{2}$

K-Means



K-Means

(The Lloyd's method)

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

Repeat until converges (or other stopping criterion):

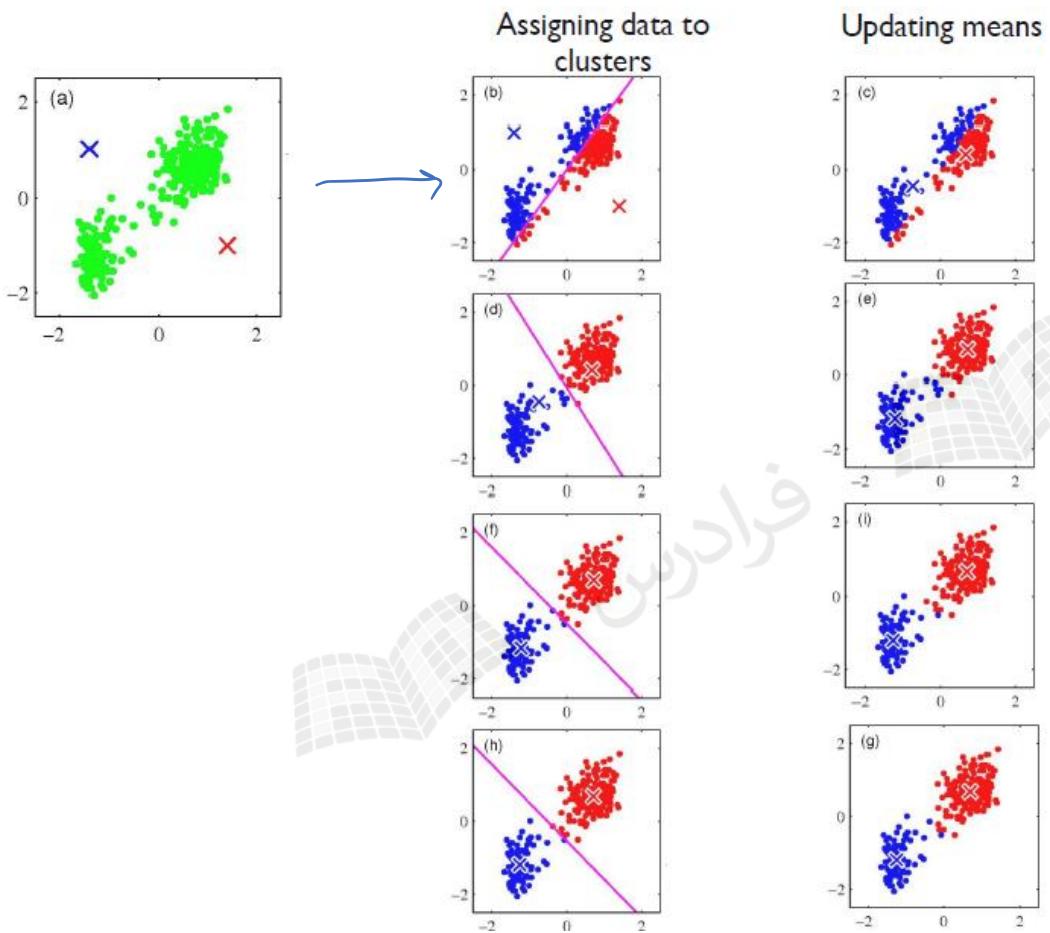
for i=1 to N do:

 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

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

K-Means



K-means Clustering

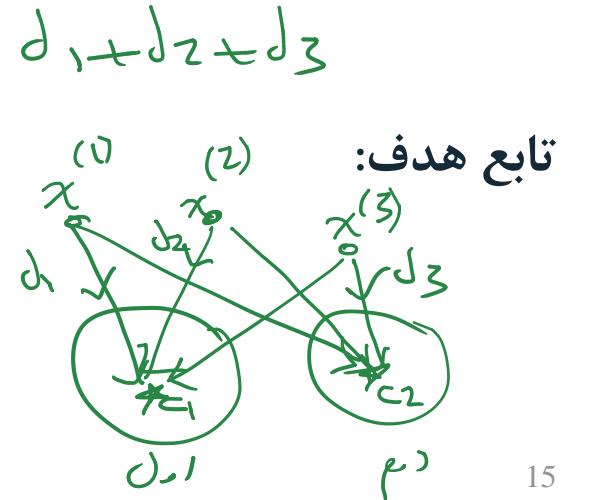
a set $x^{(1)}, \dots, x^{(N)}$ of data points and an integer K
(in d-dim feature space)

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

choose c_1, c_2, \dots, c_K to minimize:

$$N=3 \\ K=2$$

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



تابع هدف

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

$$\sum_{j=2}^{k=2}$$

NP-hard

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

اقلیدسی:

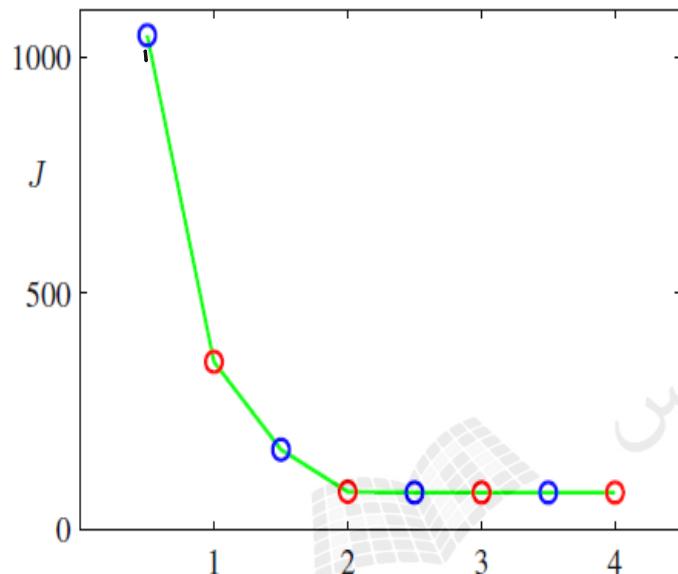
$$x^{(1)}, x^{(2)}, x^{(3)}$$

$$J(C) = \sum_{j=1}^K \sum_{x^{(i)} \in C_j} \|x^{(i)} - c_j\|^2$$

$$x^{(1)}, x^{(2)}, c_1$$

$$x^{(3)}, c_2$$

K-Means همگرایی



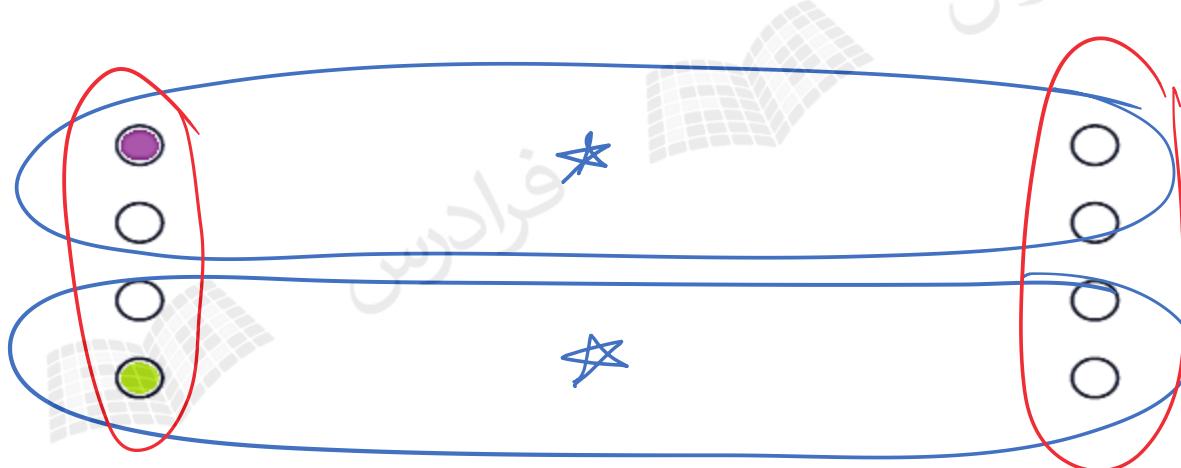
✓ الگوریتم k-means همواره همگرا است.

در هر دو فاز مقدار تابع هزینه کم می شود.

بهینه محلی

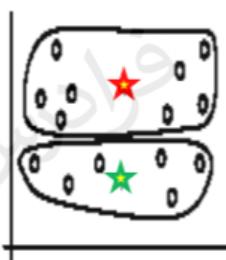
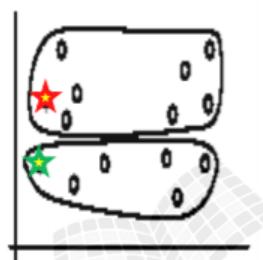
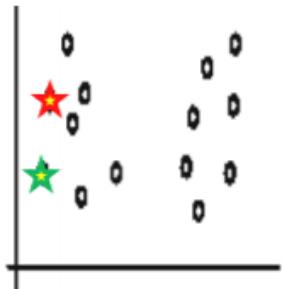
$K=2$

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



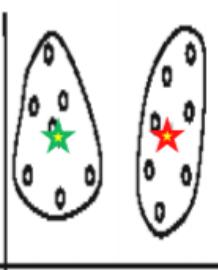
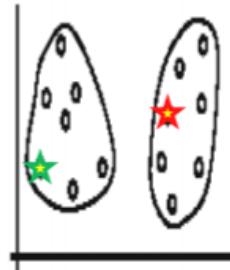
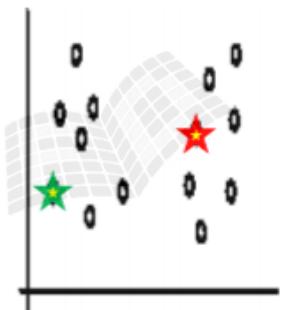
k-means++

Sensitivity to initial seeds



Iteration 1

Iteration 2

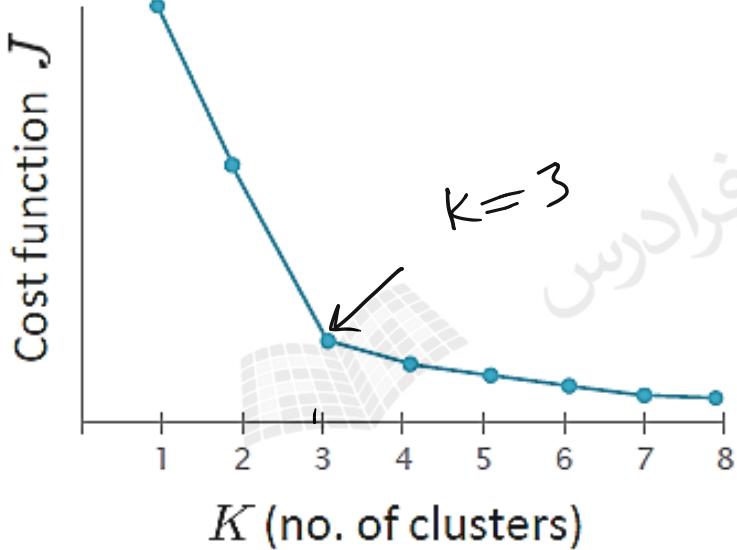


Iteration 1

Iteration 2

انتخاب تعداد خوشه‌ها

elbow



آریهی

نقاط قوت K-Means

- پیاده‌سازی ساده
- پیچیدگی زمانی الگوریتم:

$O(nkt)$

تعداد رکوردها

تعداد کلاسات

تعداد
تکرار

$t \ll n$

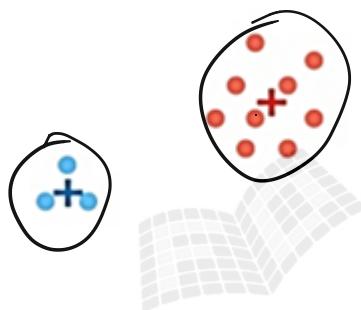
$k \ll n$

نقاط ضعف K-Means

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



حساس بودن به داده‌های پرت

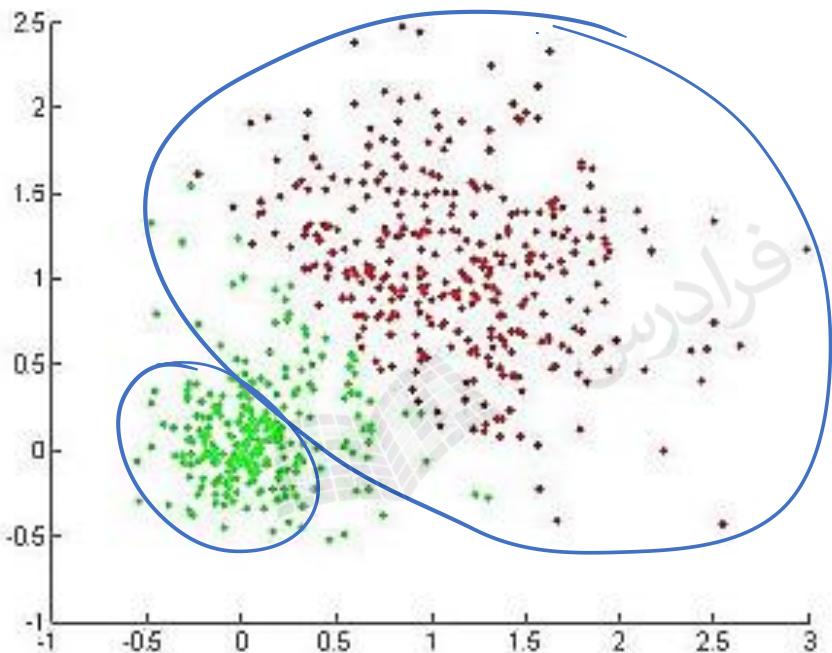


بدون outlier



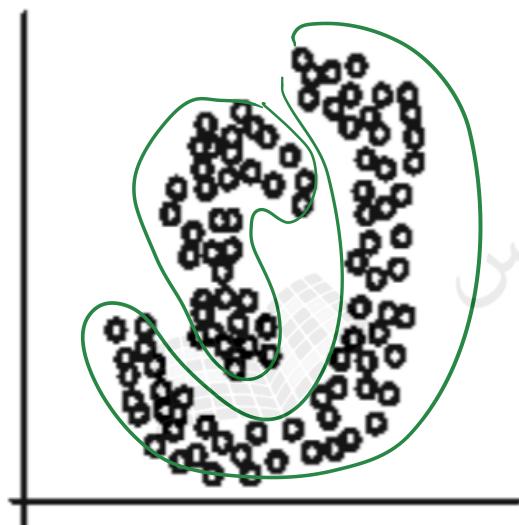
با outlier

مثال

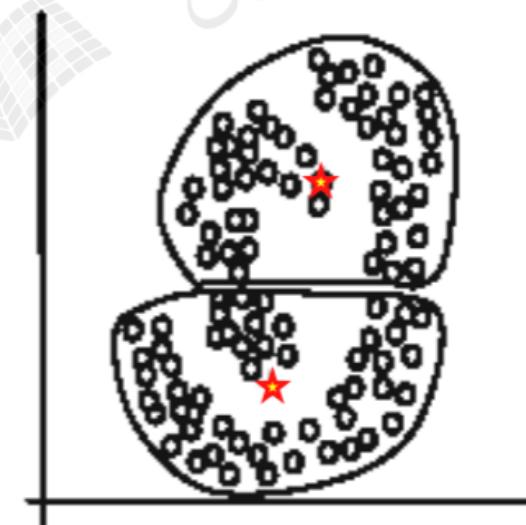


مثال

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



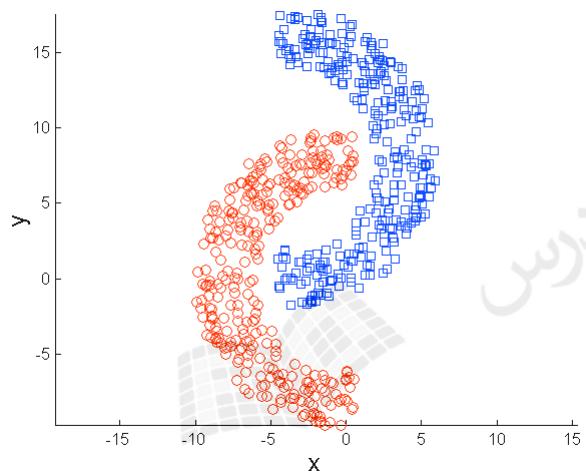
(A): Two natural cluster



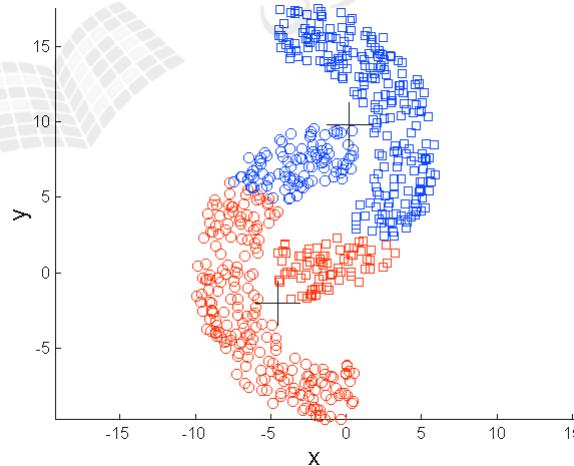
(B): k-means cluster

مثال

Non-globular Shapes



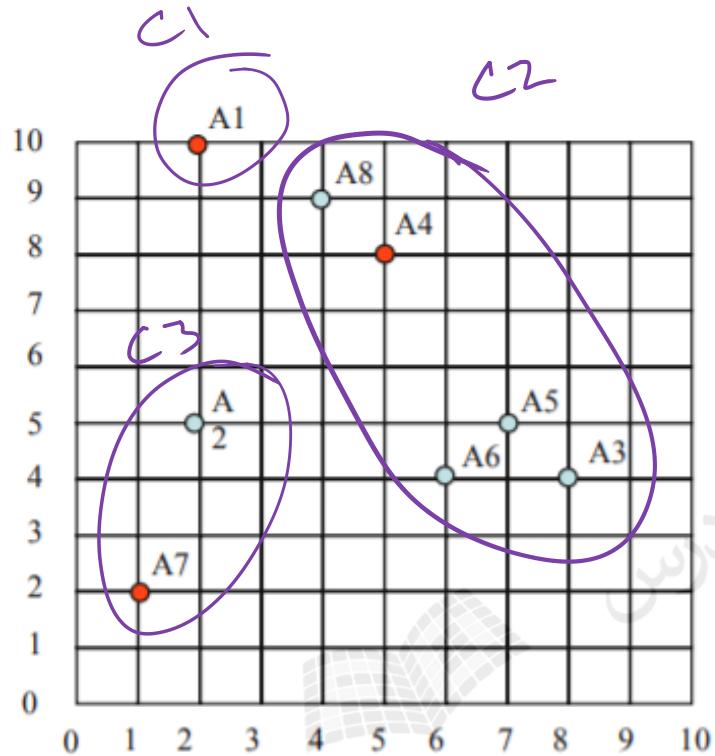
Original Points



K-means (2 Clusters)

مثال

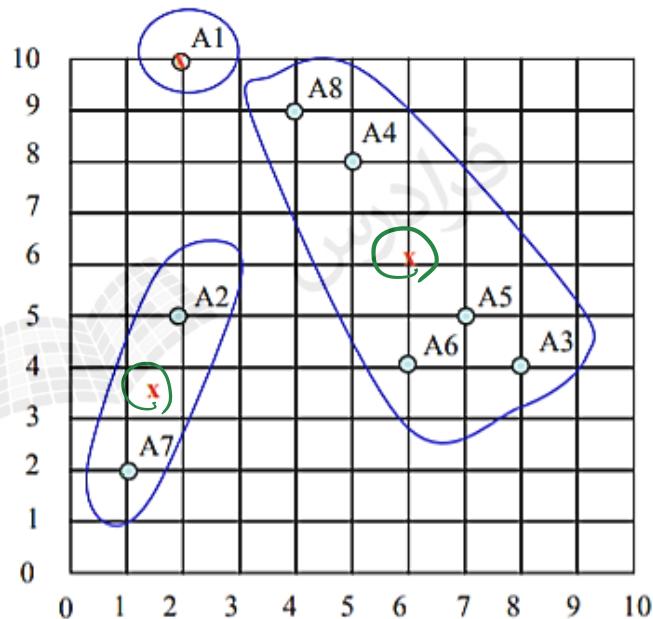
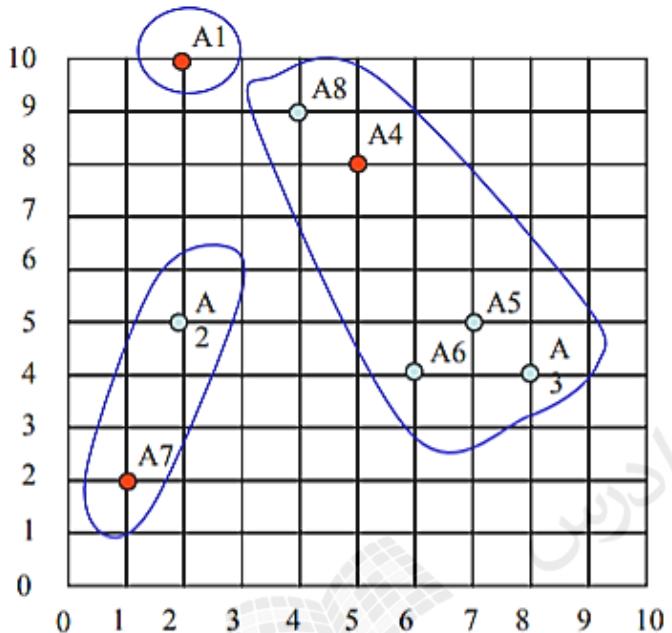
$$k = 3$$



$$d(a,b) = |x_2 - x_1| + |y_2 - y_1|$$

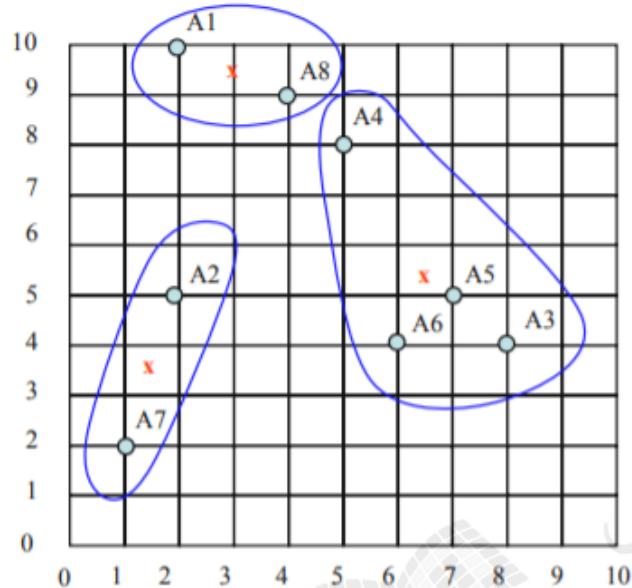
| | | $C_1 \downarrow$ | $C_2 \downarrow$ | $C_3 \downarrow$ | cluster |
|----|--------|------------------|------------------|------------------|---------|
| A1 | (2,10) | 0 | 5 | 9 | 1 |
| A2 | (2,5) | 5 | 6 | 4 | 3 |
| A3 | (8,4) | 12 | 7 | 9 | 2 |
| A4 | (5,8) | 5 | 0 | 10 | 2 |
| A5 | (7,5) | 10 | 5 | 9 | 2 |
| A6 | (6,4) | 10 | 5 | 7 | 2 |
| A7 | (1,2) | 9 | 10 | 0 | 3 |
| A8 | (4,9) | 3 | 2 | 10 | 2 |

$$|2-5| + |7-8| = 3+2$$



→ $(8+5+7+6+4)/5 = 6$. • $(4+8+5+4+9)/5) = 6$

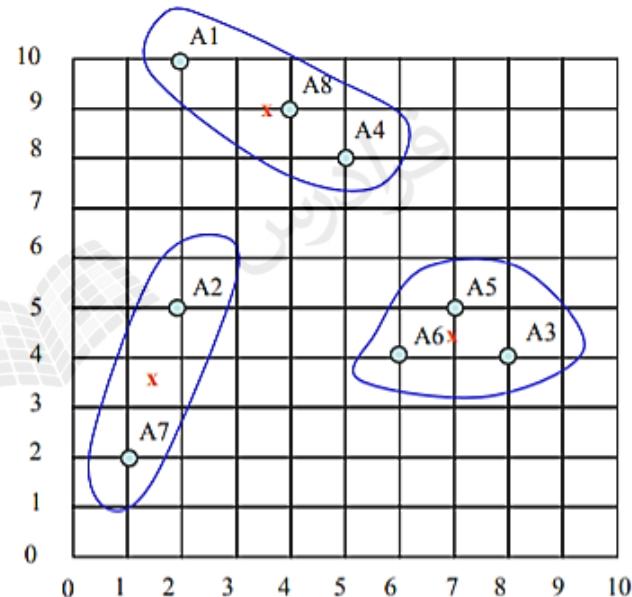
$(2+1)/2 = 1.5$ • $(5+2)/2) = 3.5$



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

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

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



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

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

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

خوشه‌بندی سلسله مراتبی

(Hierarchical clustering)

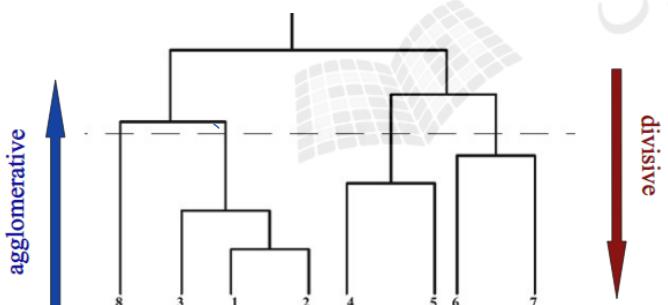
خوشه‌بندی سلسله مراتبی

۱- روش‌های تجمعی (Agglomerative) (روش پایین به بالا)

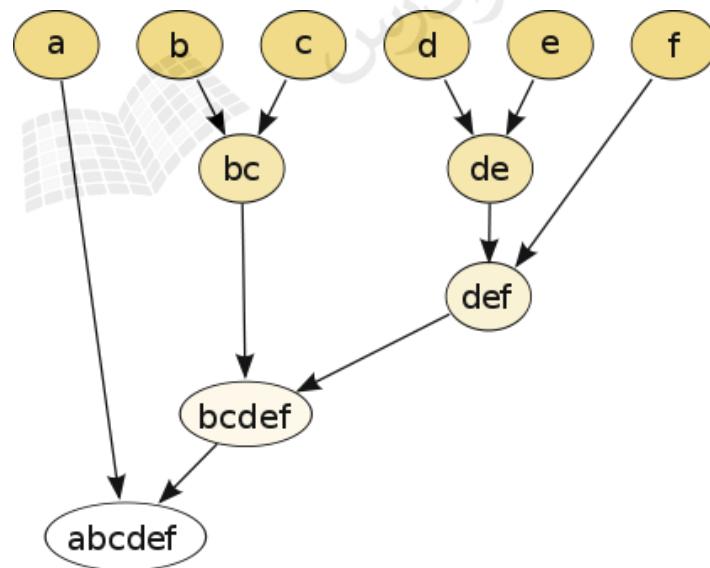
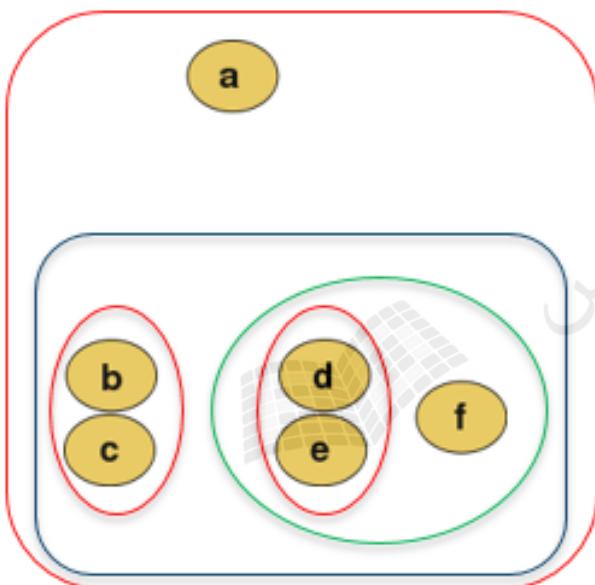
با هر داده در یک کلاستر جدا شروع می‌کند. به طور تکراری، در هر مرحله کلاسترها نزدیک به هم را ترکیب می‌کند تا حد اکثر یک کلاستر باقی بماند. (یا شرایط توقف دیگر)

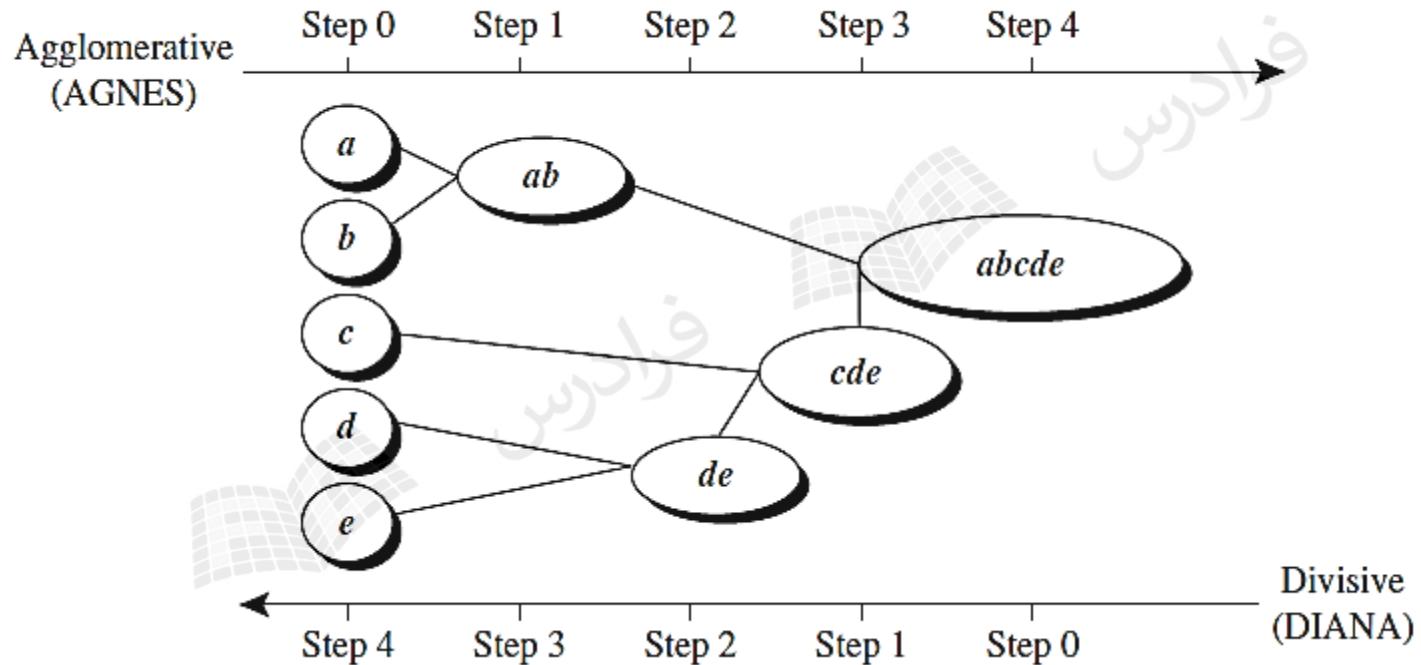
۲- روش‌های تقسیمی (Divisible) (روش بالا به پایین)

با کل داده‌ها به عنوان یک کلاستر شروع می‌کند. به طور تکراری، داده‌ها را در یکی از کلاسترها تقسیم می‌کند تا هنگامی که فقط یک داده در هر کلاستر باشد. (یا شرایط توقف دیگر)



خوشه‌بندی سلسله مراتبی و نمودار دندروگرام آن





فاصله دو کلاستر

- Single link

Minimum distance between different pairs of data

$$dist_{SL}(\mathcal{C}_i, \mathcal{C}_j) = \min_{x \in \mathcal{C}_i, x' \in \mathcal{C}_j} dist(x, x')$$

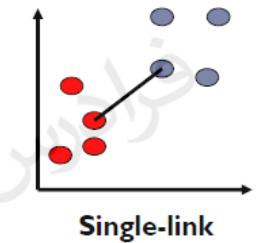
- Complete link

Maximum distance between different pairs of data

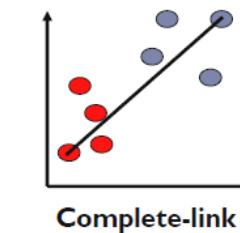
$$dist_{CL}(\mathcal{C}_i, \mathcal{C}_j) = \max_{x \in \mathcal{C}_i, x' \in \mathcal{C}_j} dist(x, x')$$

- Averaged link

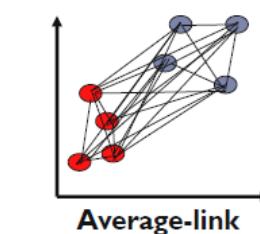
Average distance between pairs of element



Single-link



Complete-link

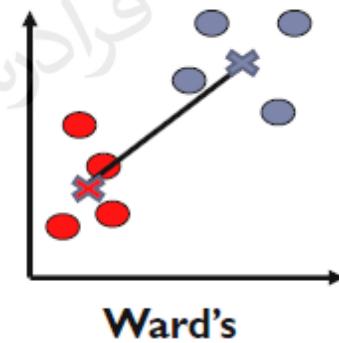


Average-link

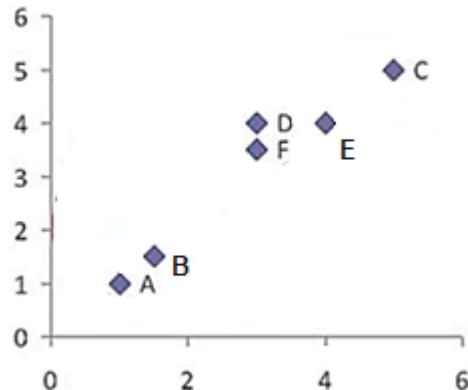
- **Centroid**

Distance between centroids(centers of gravity)

$$dist_{Ward}(\mathcal{C}_i, \mathcal{C}_j) = \frac{|\mathcal{C}_i||\mathcal{C}_j|}{|\mathcal{C}_i| + |\mathcal{C}_j|} dist(\mathbf{c}_i, \mathbf{c}_j)$$



مثال (Single link)



$$F \rightarrow D : \sqrt{(3-3)^2 + (3.5-4)^2} = 0.5$$

| | X | Y |
|---|-----|-----|
| A | 1 | 1 |
| B | 1.5 | 1.5 |
| C | 5 | 5 |
| D | 3 | 4 |
| E | 4 | 4 |
| F | 3 | 3.5 |

| | A | B | C | D | E | F |
|---|------|------|------|------|------|------|
| A | 0.00 | 0.71 | 5.66 | 3.61 | 4.24 | 3.20 |
| B | 0.71 | 0.00 | 4.95 | 2.92 | 3.54 | 2.50 |
| C | 5.66 | 4.95 | 0.00 | 2.24 | 1.41 | 2.50 |
| D | 3.61 | 2.92 | 2.24 | 0.00 | 1.00 | 0.50 |
| E | 4.24 | 3.54 | 1.41 | 1.00 | 0.00 | 1.12 |
| F | 3.20 | 2.50 | 2.50 | 0.50 | 1.12 | 0.00 |



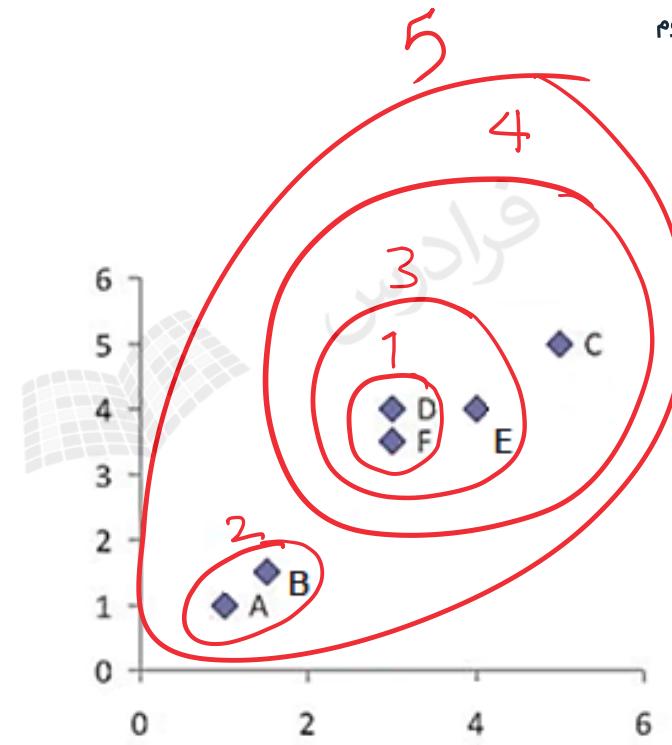
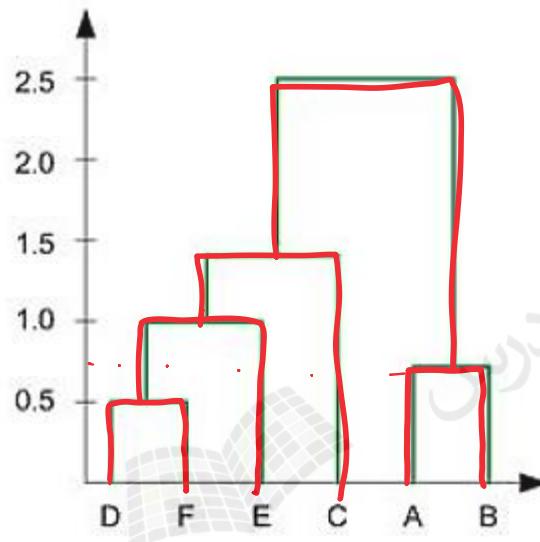
| | A | B | C | (D,F) | E |
|-------|------|------|------|-------|---|
| A | 0 | | | | |
| B | 0.71 | 0 | | | |
| C | 5.66 | 4.95 | 0 | | |
| (D,F) | 3.20 | 2.50 | 2.24 | 0 | |
| E | 4.24 | 3.54 | 1.41 | 1 | 0 |

| | (A,B) | C | (D,F) | E |
|-------|-------|------|-------|---|
| (A,B) | 0 | | | |
| C | 4.95 | 0 | | |
| (D,F) | 2.50 | 2.24 | 0 | |
| E | 3.54 | 1.41 | 1 | 0 |



| | (A,B) | C | (D,F),E |
|---------|-------|------|---------|
| (A,B) | 0 | | |
| C | 4.95 | 0 | |
| (D,F),E | 2.50 | 1.41 | 0 |

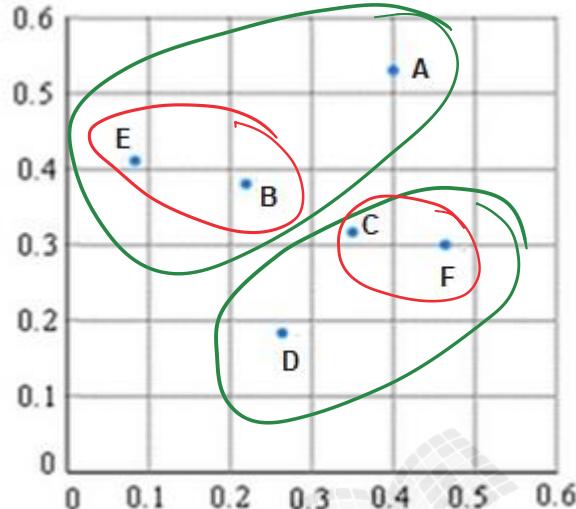
=> ...



مثال

| | X | Y |
|---|------|------|
| A | 0.40 | 0.53 |
| B | 0.22 | 0.38 |
| C | 0.35 | 0.32 |
| D | 0.26 | 0.19 |
| E | 0.08 | 0.41 |
| F | 0.45 | 0.30 |

| | A | B | C | D | E | F |
|---|------|------|------|------|------|---|
| A | 0 | | | | | |
| B | 0.24 | 0 | | | | |
| C | 0.22 | 0.15 | 0 | | | |
| D | 0.37 | 0.20 | 0.15 | 0 | | |
| E | 0.34 | 0.14 | 0.28 | 0.29 | 0 | |
| F | 0.23 | 0.25 | 0.11 | 0.22 | 0.39 | 0 |

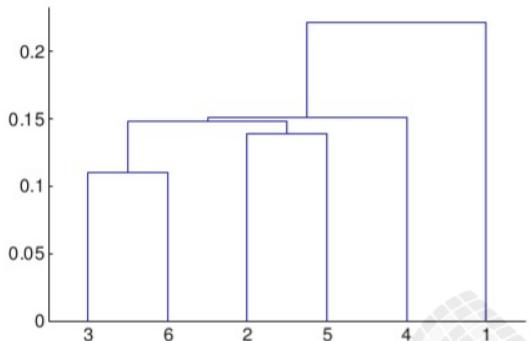


| | A | B | (C,F) | D | E |
|-------|------|------|-------|------|---|
| A | 0 | | | | |
| B | 0.24 | 0 | | | |
| (C,F) | 0.23 | 0.25 | 0 | | |
| D | 0.37 | 0.20 | 0.22 | 0 | |
| E | 0.34 | 0.14 | 0.39 | 0.29 | 0 |

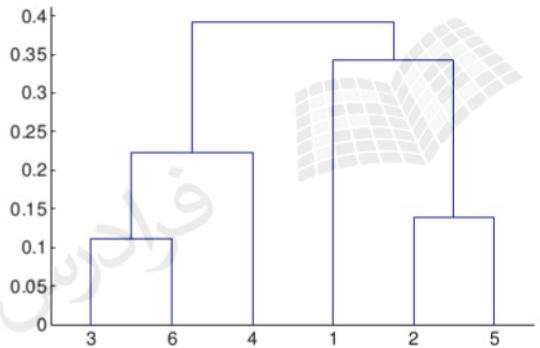
$$dist(\{C, F\} \rightarrow \{A\}) = \max \left\{ \underbrace{\{C\} \rightarrow \{A\}}, \underbrace{\{F\} \rightarrow \{A\}} \right\} = 0.23$$

$$dist(\{C, F, D\} \rightarrow \{A, B, E\}) = \max \left\{ \underbrace{\dots}_{\text{محاسبه}}, \dots, \dots \right\} = 0.39$$

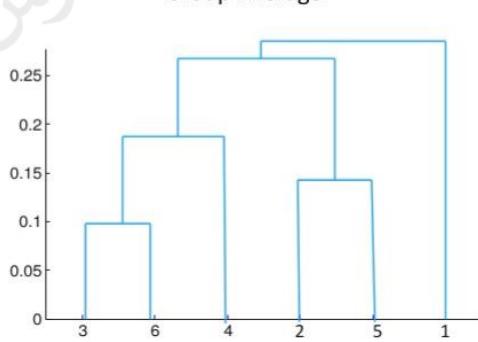
MIN



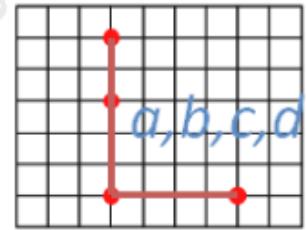
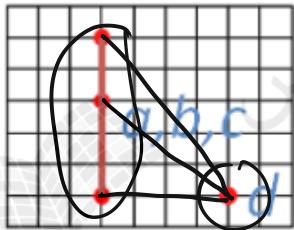
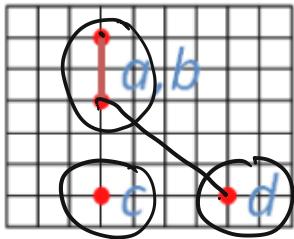
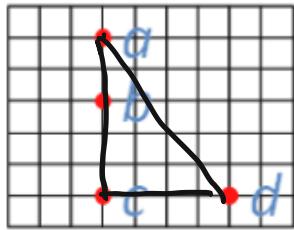
MAX



Group Average



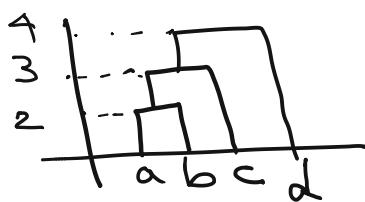
Single-Link



| | a | b | c | d |
|---|---|---|---|---|
| a | 0 | | | |
| b | 2 | 0 | | |
| c | 5 | 3 | 0 | |
| d | 6 | 5 | 4 | 0 |

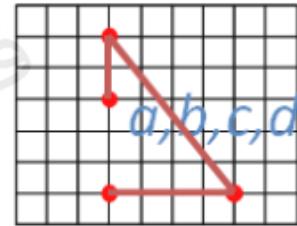
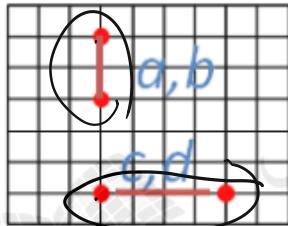
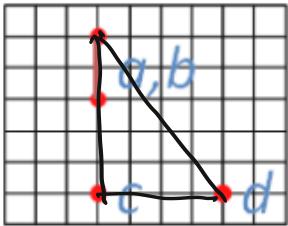
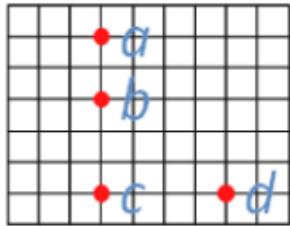
| | (a,b) | c | d |
|-------|-------|---|---|
| (a,b) | 0 | | |
| c | 3 | 0 | |
| d | 5 | 4 | 0 |

| | ((a,b),c) | d |
|-----------|-----------|---|
| ((a,b),c) | 0 | |
| d | 4 | 0 |



$$\sqrt{5^2 + 4^2} = \sqrt{41}$$

Complete-Link



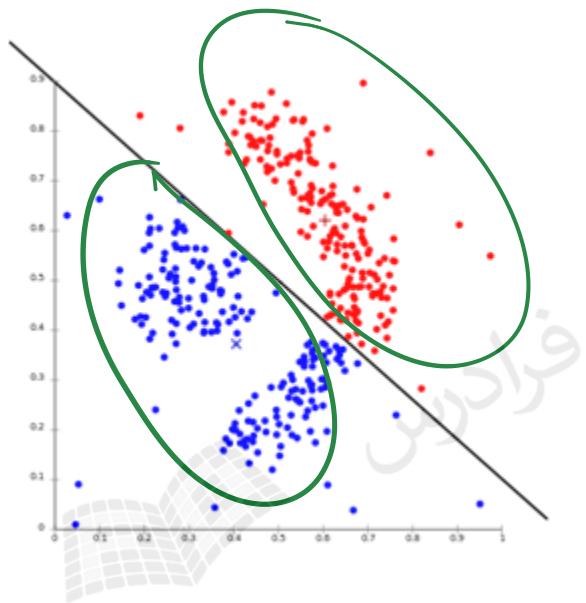
| | a | b | c | d |
|---|-----|---|---|---|
| a | 0 | | | |
| b | 2 | 0 | | |
| c | 5 | 3 | 0 | |
| d | 6.. | 5 | 4 | 0 |

| | (a,b) | c | d |
|-------|-------|---|---|
| (a,b) | 0 | | |
| c | 5 | 0 | |
| d | 6... | 4 | 0 |

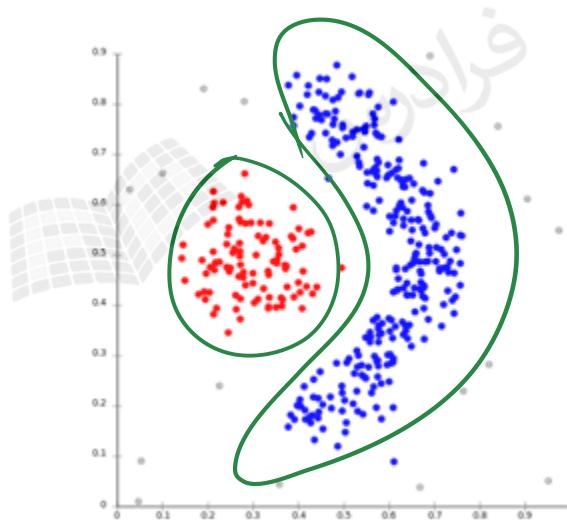
$$\sqrt{25 + 16}$$

$$\max \left\{ \begin{array}{l} 5 \\ 3 \\ - \\ - \end{array} \right\}$$

- - -



K-means



Single-Linkage

K-Means مقایسه خوشبندی سلسله مراتبی با

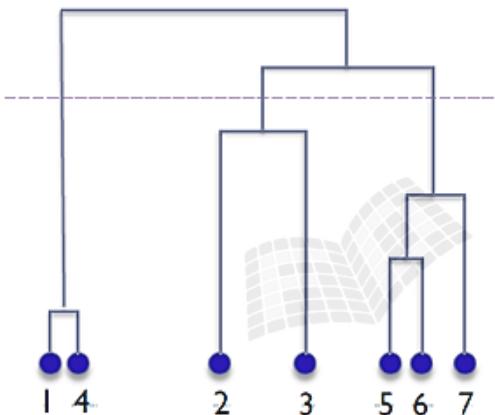
- روش سلسله مراتبی شهود بهتری برای انسان دارد. ✓
- روش k-means در بهنیه محلی گیر می‌کند. در سلسله مراتبی ممکن است حتی به مینیمم محلی هم نرسد. ✓
- هزینه زمانی در k-means کمتر از سلسله مراتبی است. ✓

Single link clustering $O(n^2)$

Complete link clustering $O(n^2 \log(n))$

مشکل بودن تعیین تعداد کلاستر

- انتخاب تعداد کلاسترها در k-means مشکل است، ولی در سلسله مراتبی از روی دندروگرام تا حدی قابل تشخیص است.



خوشه بندی مبتنی بر چگالی

(Density-based Clustering)

خوشه بندی مبتنی بر چگالی

- از مفهوم چگالی داده ها استفاده می کنند.
- محدود به اشکال محدب نمی باشند.
- لازم نیست مقدار k ، از ابتدا مشخص باشد.
- نسبت به نویز و داده های پرت مقاوم هستند.
- برای داده های زیاد مناسب هستند.
- به پارامترهای چگالی حساس هستند.

روش DBSCAN

از روش های معروف خوشه بندی مبتنی بر چگالی.

(Density-Based Spatial Clustering of Applications with Noise)

: ماکریم شعاع همسایگی. **EPS**

: حداقل تعداد نقاط در یک EPS همسایگی آن نقطه. **MinPts**



نقاط هسته ای، حاشیه ای و نویزی

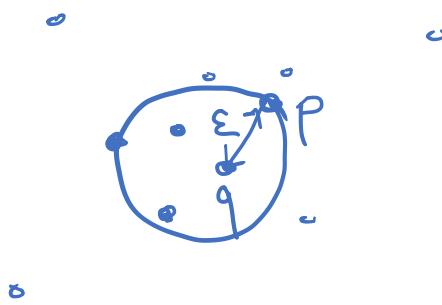
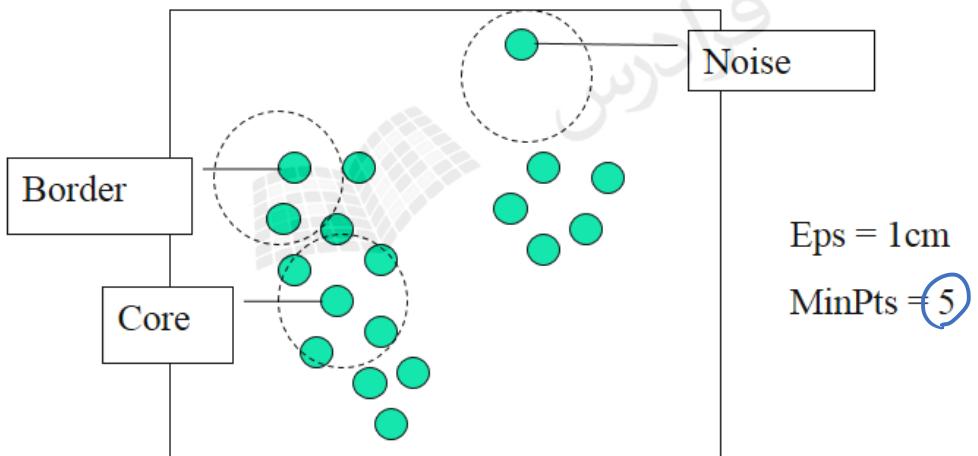
$N_{Eps}(q) : \{p \text{ belongs to } D \mid \text{dist}(p,q) \leq Eps\}$

همسایه های q:

$|N_{Eps}(q)| \geq MinPts$: نقطه هسته ای ✓

(Core point) نقطه حاشیه ای ✓

(Border point) نقطه نویزی ✓



Directly density_reachable

A point p is **directly density-reachable** from a point q w.r.t. Eps , MinPts if :

- 1) $p \in N_{\text{Eps}}(q)$
- 2) $|N_{\text{Eps}}(q)| \geq \text{MinPts}$

Density-Reachable:

p_1, \dots, p_n

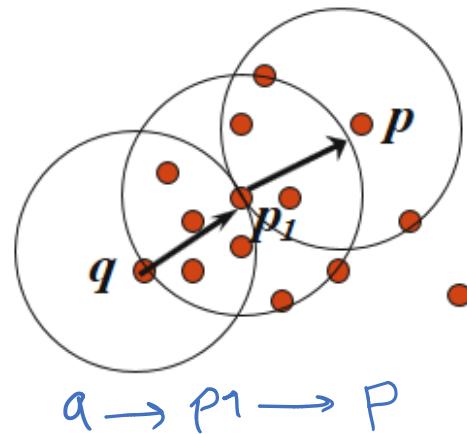
$p_i \rightarrow p_{i+1}$

$p_1 = q$, $p_n = p$

$q \in \text{نیز} p_1$

\downarrow
 \downarrow

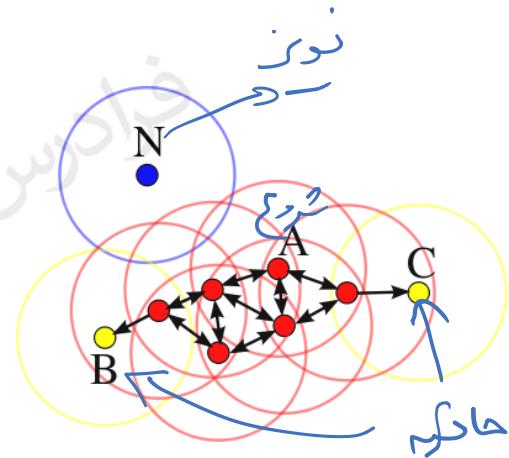
$q \in \text{نیز} p$



الگوریتم DBSCAN

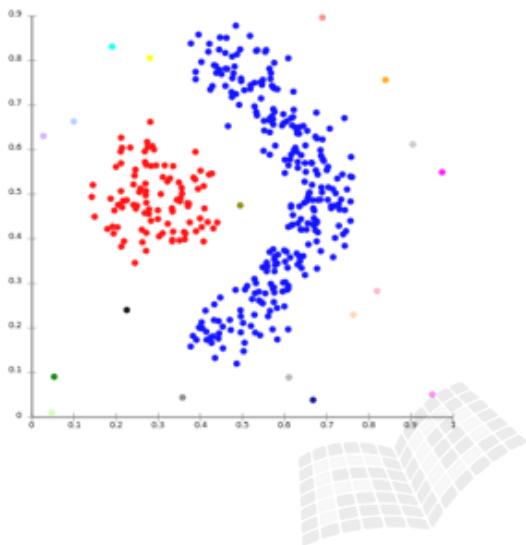
- Arbitrary select a point p
- Retrieve all points density-reachable from p w.r.t. Eps and MinPts
- If p is a core point, a cluster is formed
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed.

$O(n \log n)$
 $O(n^2)$

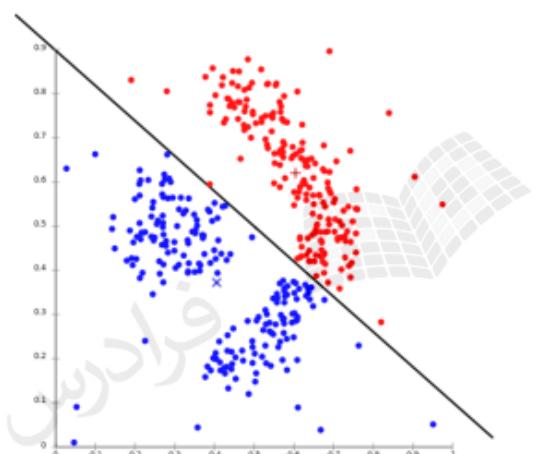


Red: core points ,
 Yellow: border points ,
 Blue: noise point

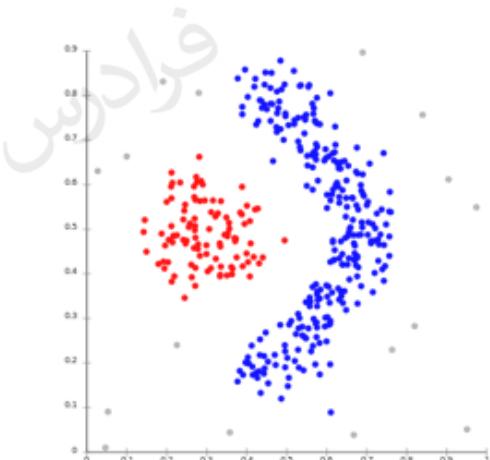
مقایسه



Single-Linkage clustering



K-means clustering



DBSCAN

مزایای DBSCAN

- نسبتا سریع.
- عدم نیاز به اعلام تعداد کلاسترها به الگوریتم.
- قابل استفاده برای هر نوع شکل.
- مقاوم در برابر نویز و داده های پرت.
- غیر حساس نسبت به ترتیب نقاط در دیتابیس.
- فقط نیاز به انتخاب درست دو پارامتر.

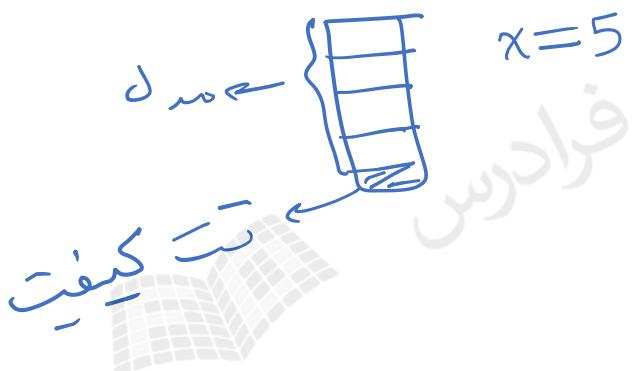
DBSCAN معاایب

- حساسیت شدید به پارامترها.
- با داده هایی که چگالی نواحی آن تفاوت زیادی دارد، خوب کار نمی کند.

ارزیابی الگوریتم های خوشه بندی

تعیین مناسب k

k



$$K \approx \sqrt{\frac{n}{2}}$$

✓ قاعده تجربی:

✓ استفاده از Cross Validation

- تقسیم داده ها به X قسمت
- ایجاد مدل خوش بندی با $X-1$ قسمت
- استفاده از قسمت باقی مانده برای کیفیت خوش بندی
- تکرار مراحل به تعداد X بار برای مقادیر مختلف k و انتخاب مقداری که کمترین خطا را در مرحله قبل دارد.

اندازه‌گیری کیفیت خوشبندی

۱- روش‌های با ناظر یا **Extrinsic Methods**

مانند شاخص Rand

۲- روش‌های بدون ناظر یا **Intrinsic Methods**

معیار SSE و Graph-based ✓

Rand شاخص

$$RandIndex = \frac{n_{11} + n_{00}}{n_{11} + n_{00} + n_{10} + n_{01}}$$

Diagram illustrating the components of the Rand Index formula:

- n_{11} is labeled TP (True Positive).
- n_{00} is labeled TN (True Negative).
- n_{10} is labeled FP (False Positive).
- n_{01} is labeled FN (False Negative).

معیارهای ارزیابی بدون ناظر

$$Graph-based = \sum_{i=1}^K \frac{Separation(i)}{Cohesion(i)}$$

میان جدایی صریح‌تر از بقیه
میان چینه‌گر خود را

میان نزدیکی

$$SSE = \sum_{i=1}^K \sum_{x \in k_i} (x - c_i)^2$$

مقدار خطا را

این اسلایدها بر مبنای نکات مطرح شده در فرادرس
«آموزش یادگیری ماشین (Machine Learning) (تئوری - عملی) - بخش دوم»
تهییه شده است.

برای کسب اطلاعات بیشتر در مورد این آموزش به لینک زیر مراجعه نمایید.

faradars.org/fvdm94062