

Session 11

k-means Clustering

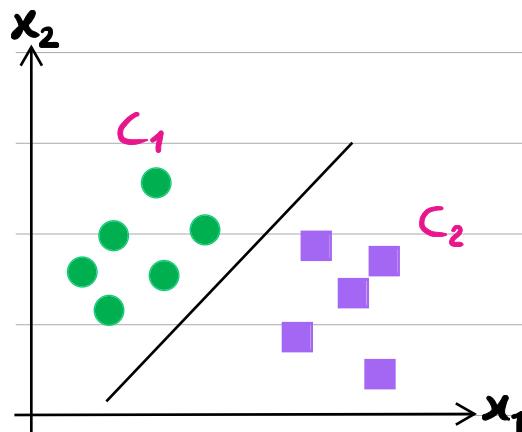


Supervised $\rightarrow (x_1, y_1) (x_2, y_2) (x_3, y_3) \dots (x_m, y_m)$

Unsupervised $\leftarrow k\text{-Means}$

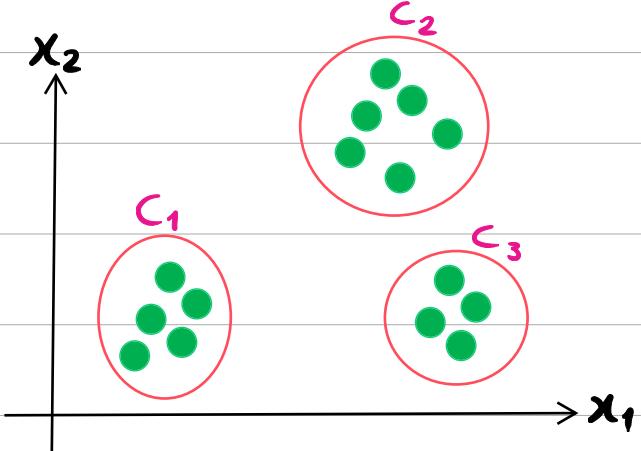
Unsupervised $\rightarrow (x_1) (x_2) (x_3) \dots (x_m)$

کمترین کاربرد k-means



classification

Supervised



Clustering

Unsupervised

خوشه بندی یا clustering: داده ها عضو خوشه cluster نهایت شباهت را باهم دارند ولی بین خوشه های مختلف بیشترین تفاوت وجود دارد

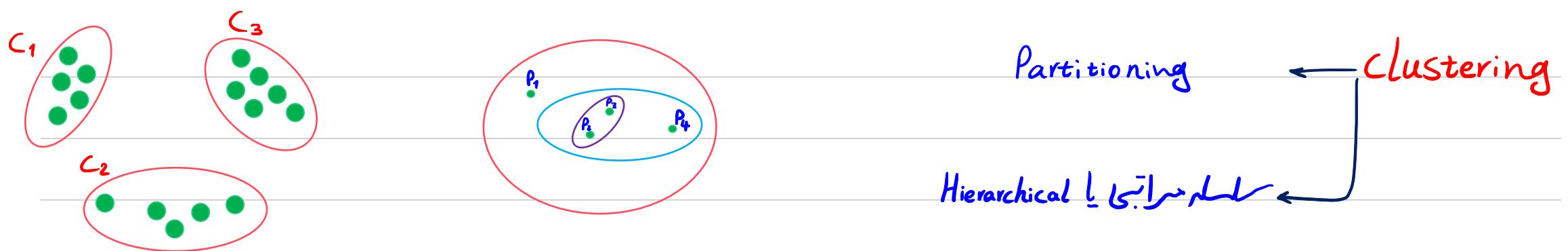
در چند کامپیوتری دانش این حالت ایده آل است و مادر تلاش حق ساختن آنیم.

high intra-clustering similarity \rightarrow Cohesive

شباهت یا پیوندگی درون cluster باید زیاد باشد. ناصله های بین داده ها کم

Low inter-cluster Similarity \rightarrow distinctive

تمایز بین cluster های متفاوت باید زیاد باشد. شباهت خارج cluster ها باید کم باشد.



Partitioned clustering

Hierarchical

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

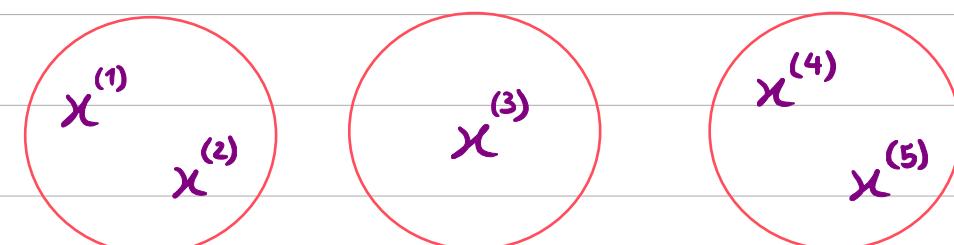
$$C = \{C_1, C_2, \dots, C_k\}$$

$$\checkmark \quad j, C_j \neq \emptyset$$

$$\bigcup_{j=1}^k C_j = X$$

ex: $X = \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}\} \rightarrow m=5$ ← خوشه‌بندی افزایشی Partitional Clustering

$$C = \{C_1, C_2, C_3\} \rightarrow k=3$$



$$\checkmark \quad i, j, C_i \cap C_j = \emptyset$$



- شرط:
1. داده‌ها (x) در clusters های مختلف شوند به طوری که هیچ cluster تکی نشود.
 2. اجتناب عناصر موجود در cluster ها برابر با X شود.
 3. تمام cluster ها دو و یک استراکچر تکی نشود.

ex:

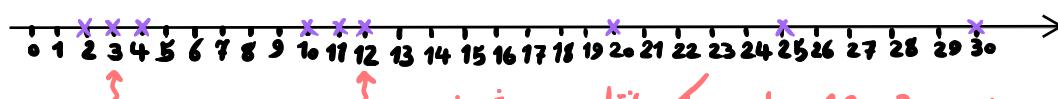
$$X = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$$

$$M = 9$$

$$k = 2$$

: k-means

①



۱. انتخاب (centroid) mean لـ k ندوم.

C_1

C_2

$$\boxed{2, 3, 4}$$

$$\boxed{10, 11, 12, 20, 25, 30}$$

۲. با توجه به مرکزها cluster ها را دوباره指派 (reassigned).

۳. بـ دست آوردن centroid های جدید با توجه به داده ها mean ←

$$\text{بـ دست آوردن مرکز جدید} \rightarrow C_1 = \frac{2+3+4}{3} = 3$$

$$C_2 = \frac{10+11+12+20+25+30}{6} = 18$$

②



۴. تکرار مراحل ۲ و ۳ تا زمانی که تغیری در cluster ها ثابت شود باشند.

C_1

C_2

$$\boxed{2, 3, 4, 10}$$

$$\boxed{11, 12, 20, 25, 30}$$

$$\text{بـ دست آوردن مرکز جدید} \rightarrow C_1 = \frac{2+3+4+10}{4} = 4.75$$

$$C_2 = \frac{11+12+20+25+30}{5} = 19.6$$

③

C_1

$$\boxed{2, 3, 4, 10, 11, 12}$$

$$\boxed{20, 25, 30}$$

$$C_1 = \frac{2+3+4+10+11+12}{6} = 7$$

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

۵. ماتریس ندایم ← پایان cluster را

ex:

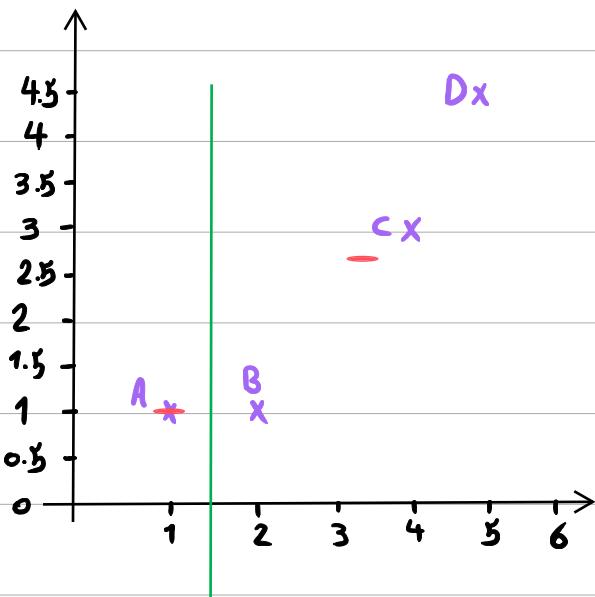
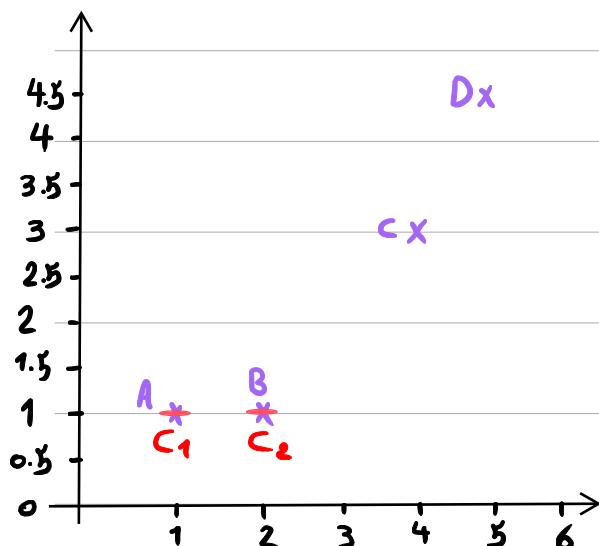
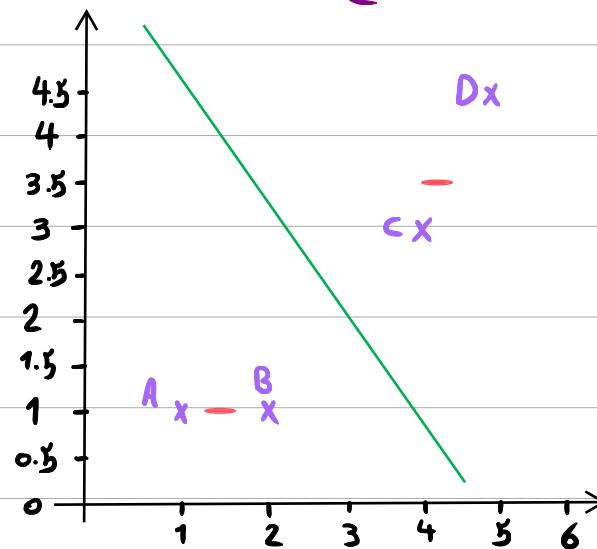
	x	y
$x^{(1)}$	A	1
$x^{(2)}$	B	2
$x^{(3)}$	C	4
$x^{(4)}$	D	5

$$k=2$$

$$m=4$$

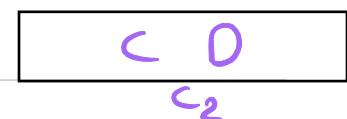
انتساب اولیه مراحلز ←

$$\begin{cases} C_1 = A \\ C_2 = B \end{cases}$$

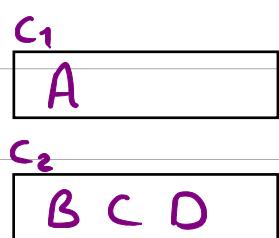


$$\text{dis}(B, A) = \sqrt{(2-1)^2 + (1-1)^2} = 1^{\min}$$

$$\text{dis}(x^2, C_2) = \sqrt{(2-3.6)^2 + (1-3.6)^2} = \sqrt{9.32}$$



$$\left. \begin{aligned} \text{dis}(C, A) &= \sqrt{(4-1)^2 + (3-1)^2} = \sqrt{11} \\ \text{dis}(C, B) &= \sqrt{(4-2)^2 + (3-1)^2} = \sqrt{8} \end{aligned} \right\} \xrightarrow{\min} C_2$$



$$C_1 = A$$

$$C_2 = (3.6, 2.6)$$

$$C_1 = (1.5, 1)$$

$$\left. \begin{aligned} \text{dis}(D, A) &= \sqrt{(5-1)^2 + (4-1)^2} = \sqrt{25} \\ \text{dis}(D, B) &= \sqrt{(5-2)^2 + (4-1)^2} = \sqrt{18} \end{aligned} \right\} \xrightarrow{\min} C_2$$

$$x = \frac{2+4+5}{3} = 3.6$$

$$y = \frac{1+3+4}{3} = 2.6$$

$$C_2 = (4.5, 3.5)$$

/ حل

k -means \rightarrow ورودی: $X \rightarrow x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}$

خروجی: $C \rightarrow C_1, C_2, C_3, \dots, C_k \in \mathbb{R}$

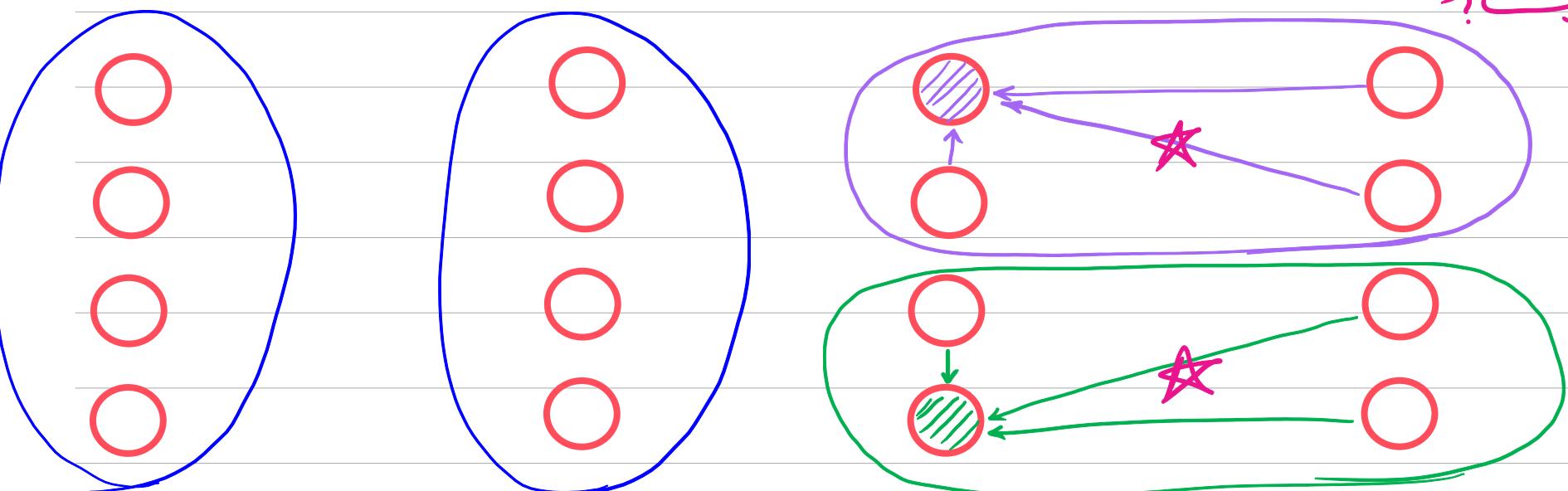
تابع هدف $\rightarrow C_1, C_2, C_3, \dots, C_k$ to Minimize:

$$\sum_{i=1}^m \min_{j=1, \dots, k} d^2(x^{(i)}, C_j)$$

$$\sum_{i=1}^m \min_{j=1, \dots, k} d^2(x^{(i)}, C_j) \xrightarrow{\text{فاصله اقلیدس}} \sum_{i=1}^m \min \left\| x^{(i)} - C_j \right\|^2$$

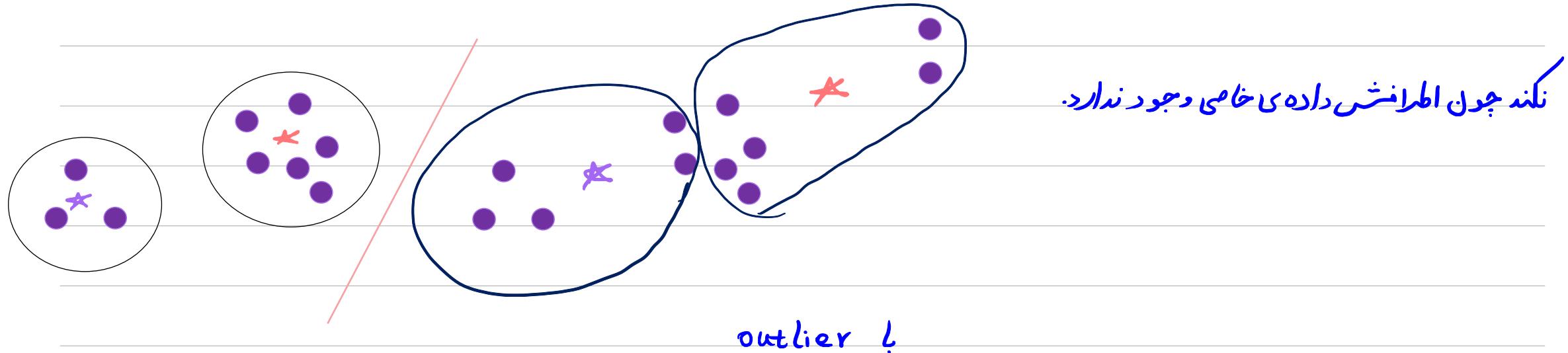
$$\Rightarrow J(C) = \sum_{j=1}^k \sum_{x^{(i)} \in C_j} \left\| x^{(i)} - C_j \right\|^2 \xrightarrow{\text{نکته داده را که می‌دانیم cluster را نمایش می‌کند}}$$

نکته درباره k -means



؟ برای رفع مشکل Local minimum چه کنم؟

می‌توان مراکز اولیه را داده‌هایی که دور هستند را در نظر بگیرید که outlier باشند آن داده باعث می‌شود در روند الگوریتم آن خوش‌رشد



حریت نسبت به انتخاب Centroid ها

