

یادگیری بدون نظارت

سید ناصر رضوی n.razavi@tabrizu.ac.ir

۱۳۹۵

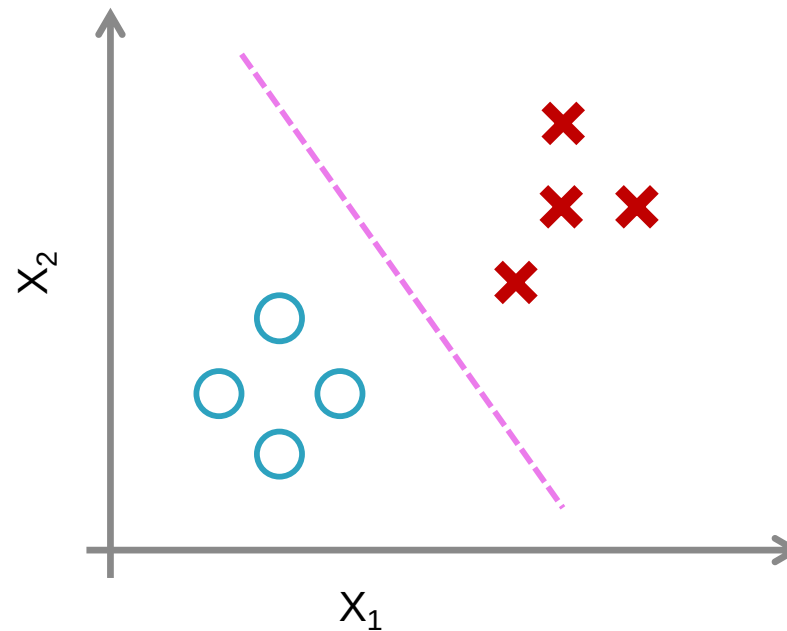
فهرست مطالب

- یادگیری بدون نظارت
- کاربردها
- خوشه‌بندی
- الگوریتم K-means
- بهبود خوشه‌بندی
- الگوریتم دو بخشی‌ساز
- خوشه‌بندی سلسله‌مراتبی

یادگیری نظارت شده

۳

□ یادگیری نظارت شده. به ازای هر نمونه، پاسخ درست داده شده است.



مجموعه‌ی آموزشی

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \dots, (x^{(m)}, y^{(m)})\}$$

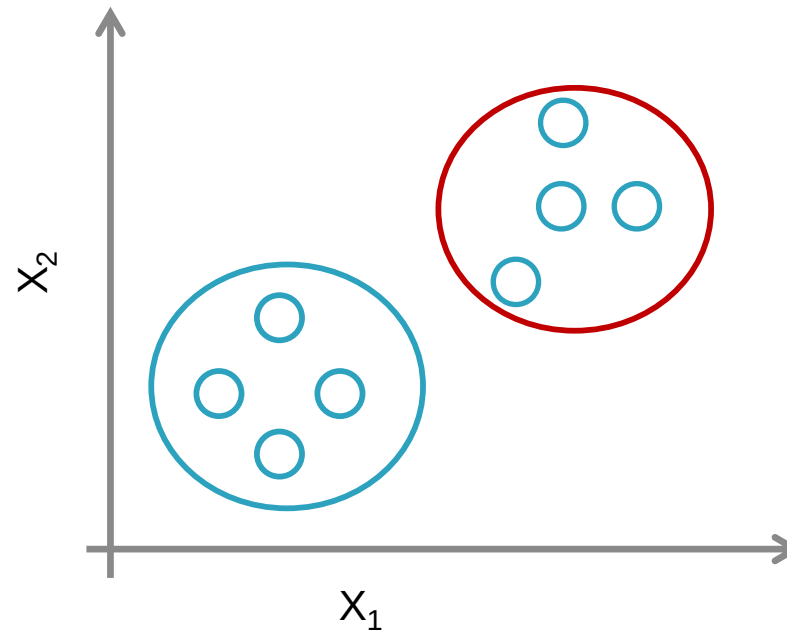
یادگیری بدون نظارت

۴

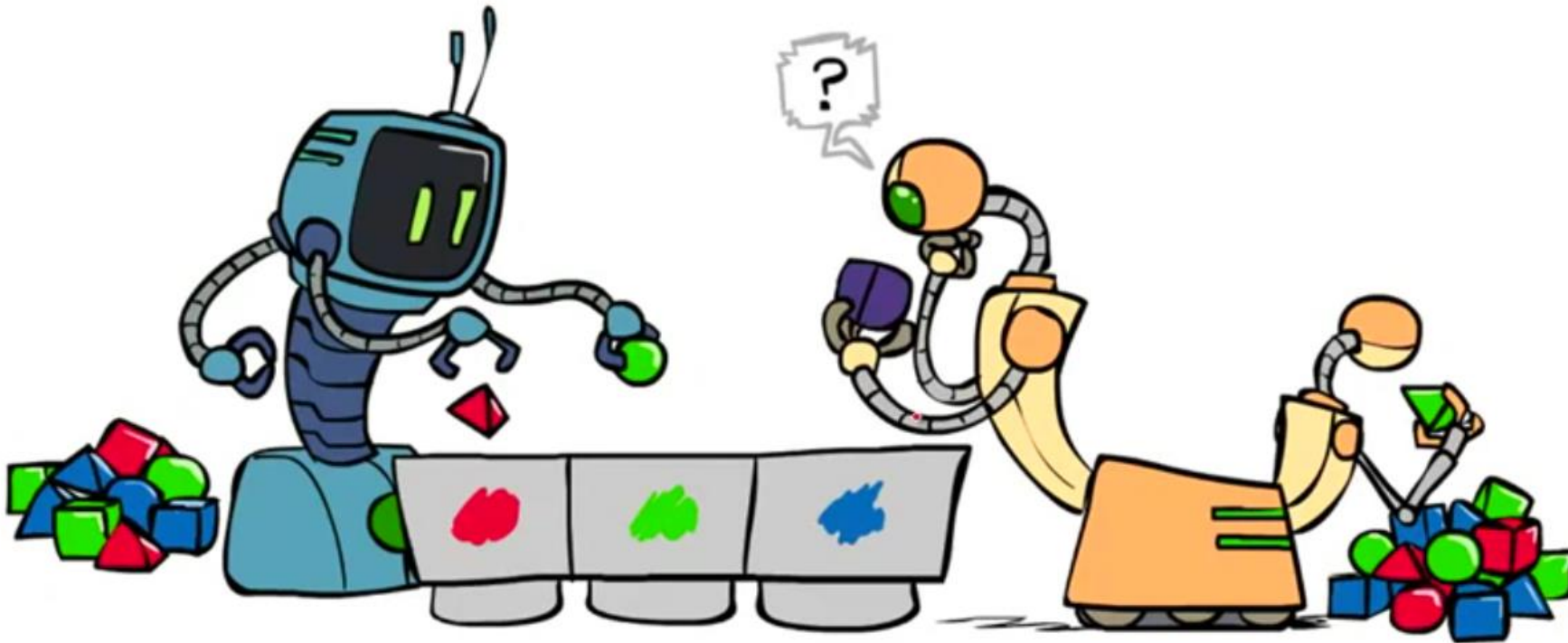
□ یادگیری بدون نظارت. عدم آگاهی از پاسخ‌های درست.

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

مجموعه‌ی آموزشی

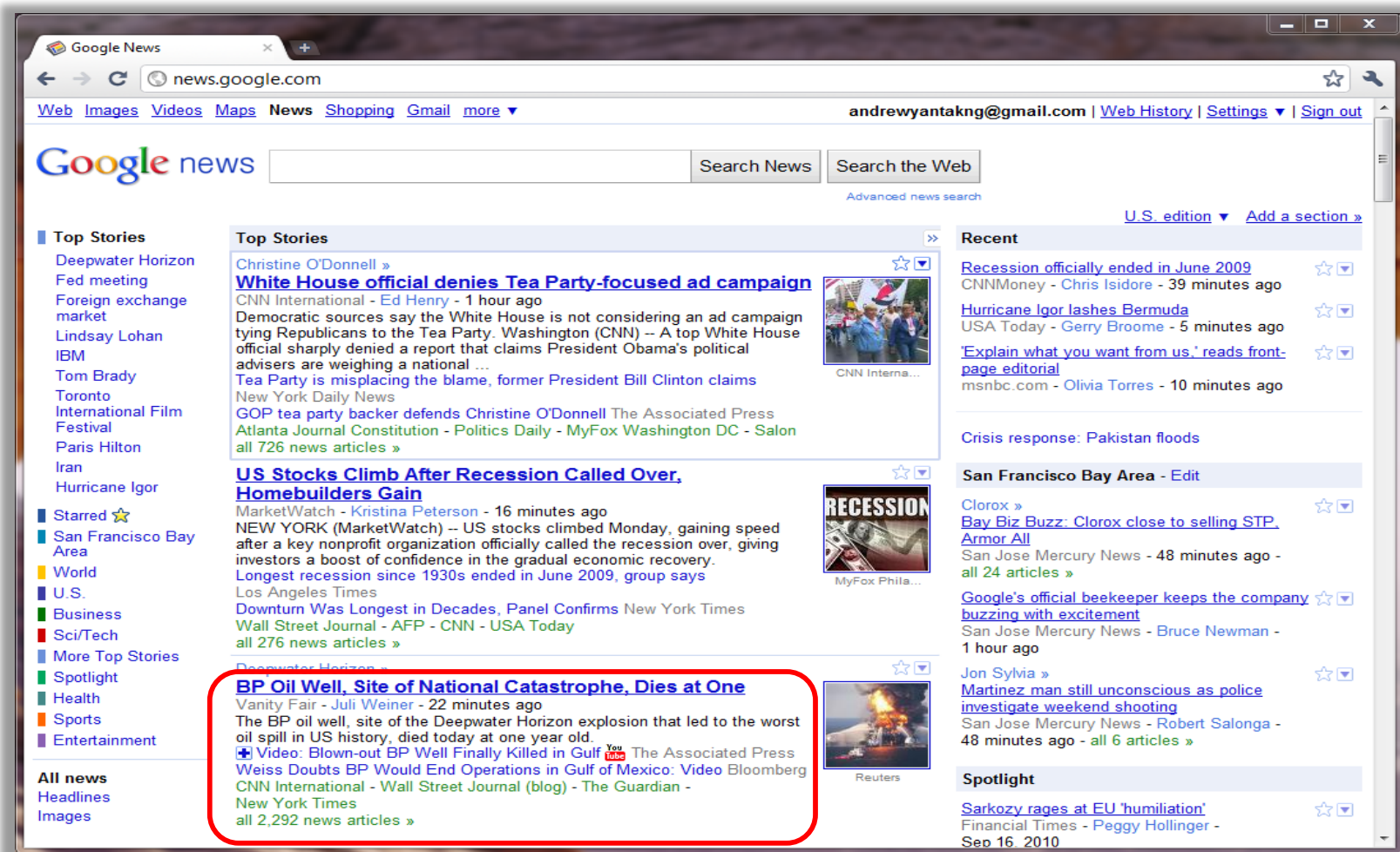


□ هدف. تشخیص ساختار در داده‌های ورودی [گروه‌بندی داده‌های مشابه]



کاربرد خوشه‌بندی: گروه‌بندی اخبار مرتبط

۶



کاربرد خوشه‌بندی: گروه‌بندی اخبار مرتبط

۷

The diagram illustrates the flow of information from Google News to The Source and then to CNN and the Guardian, illustrating the use of clustering in news aggregation.

Google News (Top Left): A screenshot of the Google News homepage. A red box highlights the article "BP Oil Well, Site of National Catastrophe, Dies at One" by CNN International. A red arrow points from this article to The Source.

The Source (Top Right): A screenshot of The Source website. A red arrow points from the Google News article to this page. The article title is "BP Kills Macondo, But Its Legacy Lives On" by James Heron. A red arrow points from this article to CNN.

CNN (Bottom Left): A screenshot of the CNN website. A red arrow points from The Source to this page. The article title is "Allen: Well is dead, but much Gulf Coast work remains" by the CNN Wire Staff. A red arrow points from this article to the Guardian.

guardian.co.uk (Bottom Right): A screenshot of the Guardian website. A red arrow points from CNN to this page. The article title is "BP oil spill cost hits nearly \$10bn" by Julia Kollewe.

Red Arrows: Four red arrows indicate the flow of information from Google News to The Source, and then to CNN and the Guardian, illustrating the use of clustering in news aggregation.

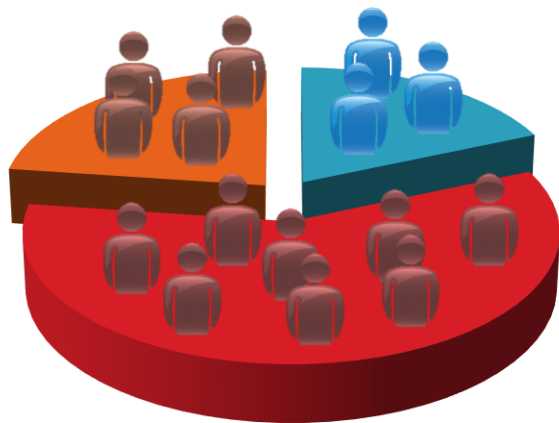
سید ناصر رضوی - یادگیری ماشین - یادگیری بدون نظارت

چند کاربرد دیگر از یادگیری بدون نظارت

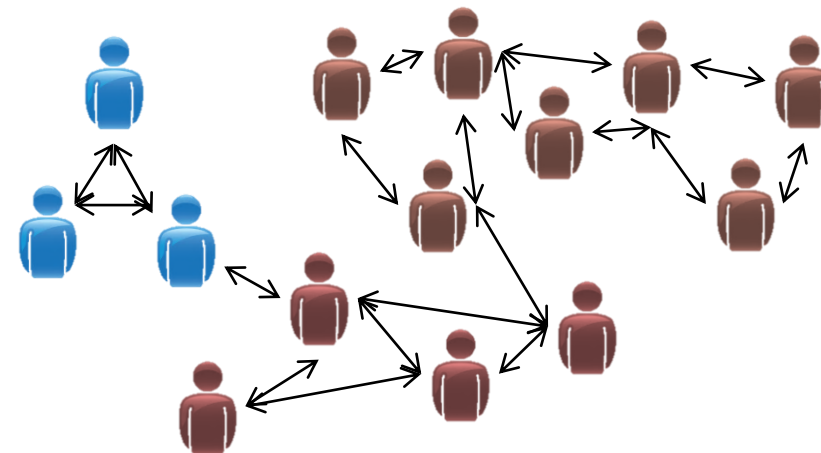
۸



سازماندهی کلاسترهای محاسباتی (مرکز داده‌ها)



بخش‌بندی بازار



تحلیل شبکه‌های اجتماعی

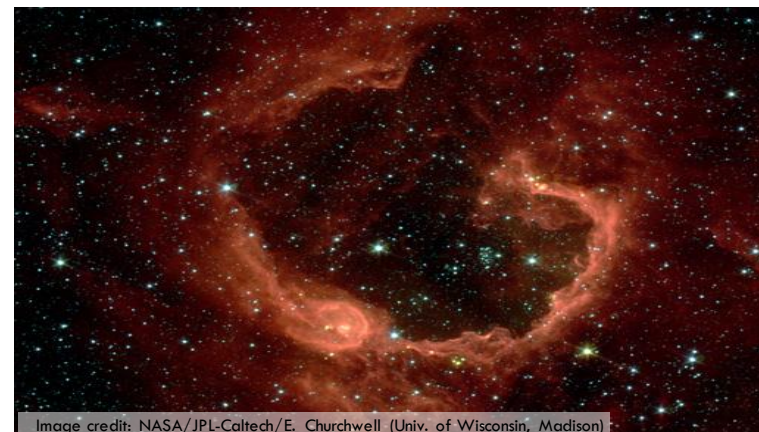
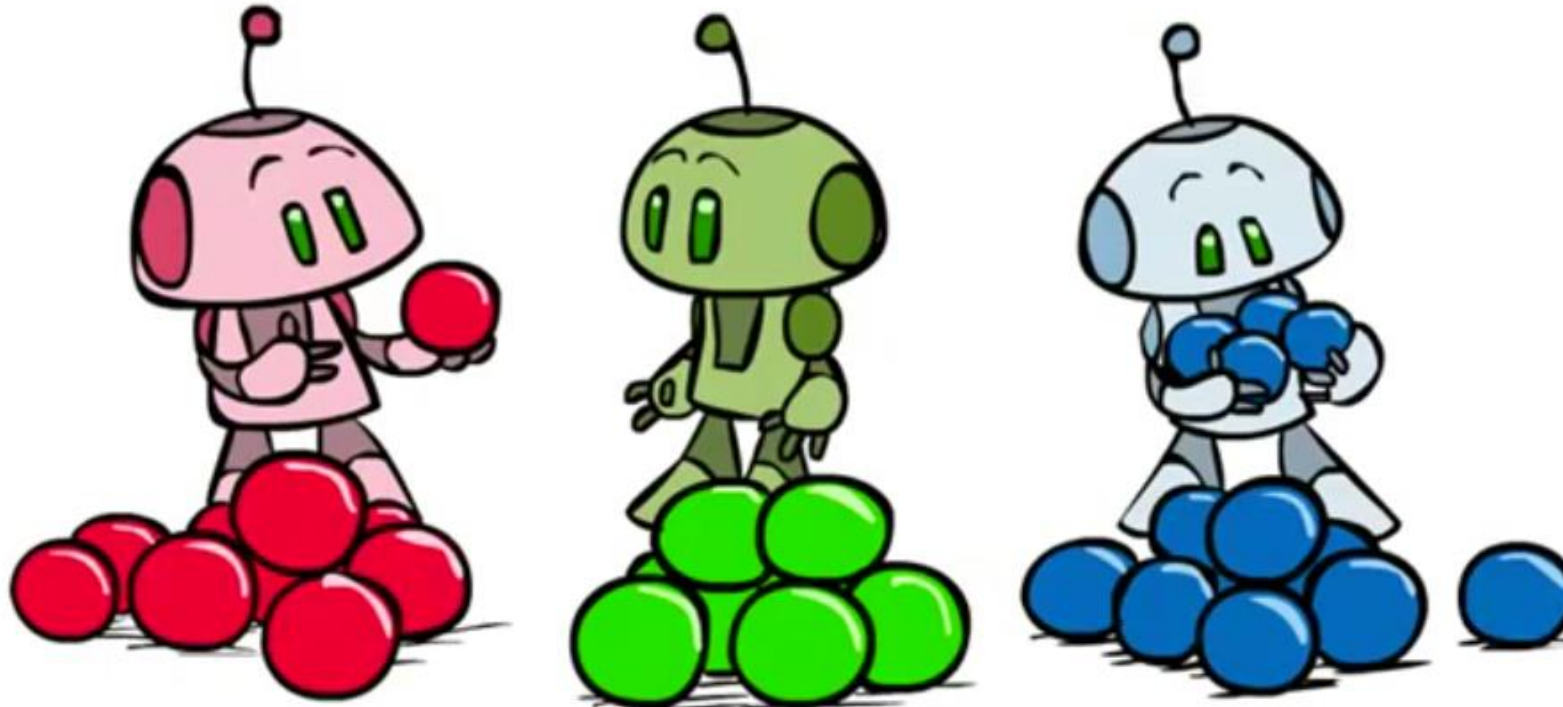


Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

تحلیل داده‌های ستاره‌شناسی (نحوه تشکیل کهکشان‌ها)

الگوریتم خوشه‌بندی K-means

۹



الگوریتم خوشه‌بندی K-means

۱۰

□ یک الگوریتم خوشه‌بندی تکرار شونده.

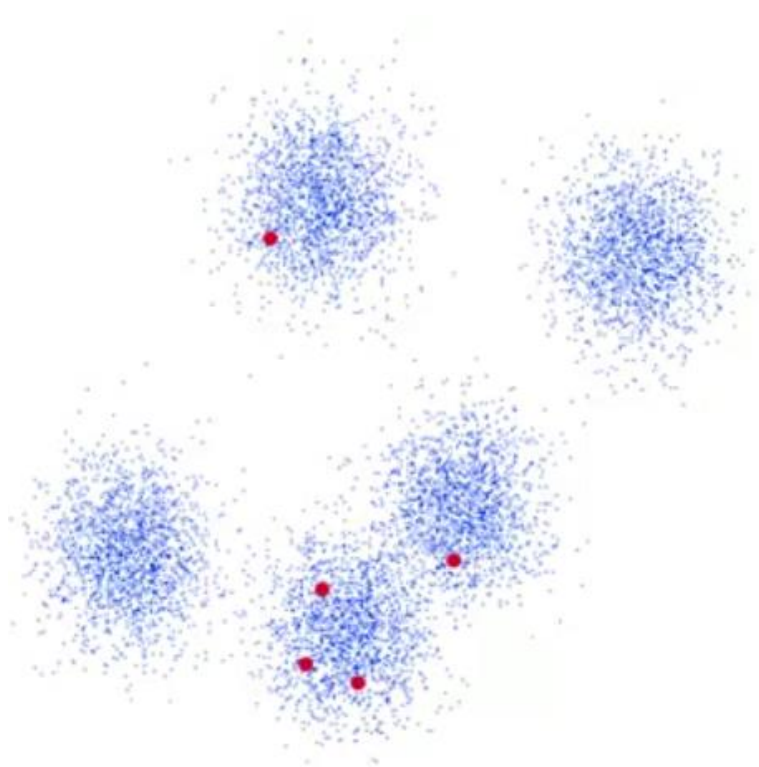
□ K نقطه را به صورت تصادفی به عنوان مراکز خوشه‌ها انتخاب کن.

□ مراحل زیر را تکرار کن:

■ هر داده را به یک خوشه با نزدیک‌ترین مرکز انتساب بده.

■ مرکز هر خوشه را با میانگین‌گیری از داده‌های انتساب یافته به آن خوشه به روز رسانی کن.

□ توقف: زمانی که در یک تکرار هیچ داده‌ای خوشه‌ی خود را عوض نکند.



الگوریتم خوشه‌بندی K-means

۱۱

□ یک الگوریتم خوشه‌بندی تکرار شونده.

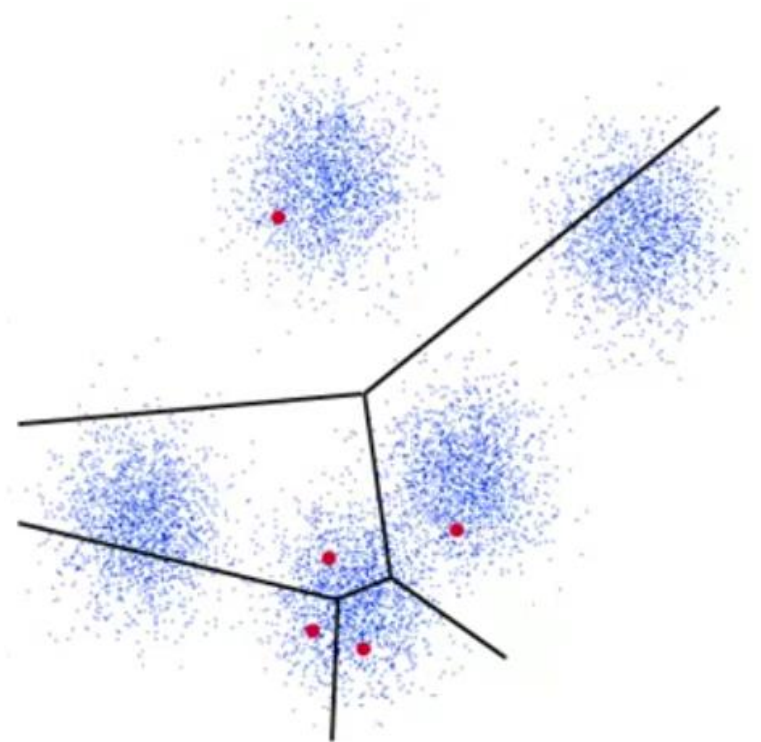
□ K نقطه را به صورت تصادفی به عنوان مراکز خوشه‌ها انتخاب کن.

□ مراحل زیر را تکرار کن:

■ هر داده را به یک خوشه با نزدیک‌ترین مرکز انتساب بده.

■ مرکز هر خوشه را با میانگین‌گیری از داده‌های انتساب یافته به آن خوشه به روز رسانی کن.

□ توقف: زمانی که در یک تکرار هیچ داده‌ای خوشه‌ی خود را عوض نکند.



الگوریتم خوشه‌بندی K-means

۱۲

□ یک الگوریتم خوشه‌بندی تکرار شونده.

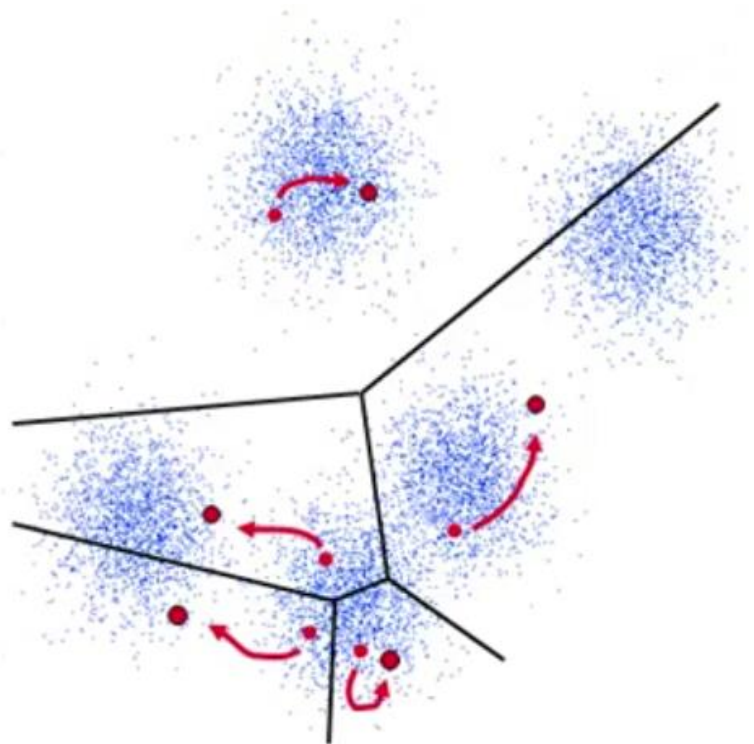
□ K نقطه را به صورت تصادفی به عنوان مراکز خوشه‌ها انتخاب کن.

□ مراحل زیر را تکرار کن:

■ هر داده را به یک خوشه با نزدیک‌ترین مرکز انتساب بده.

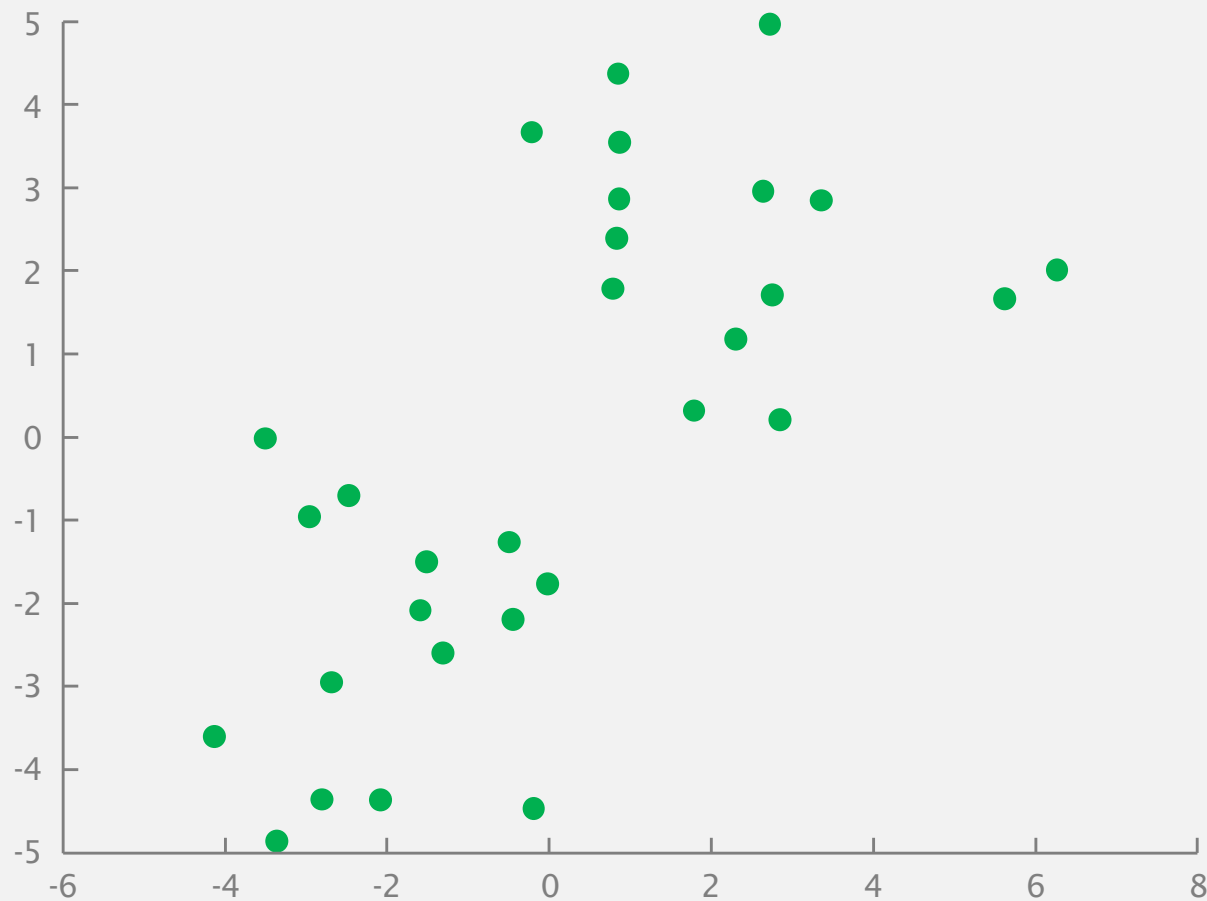
■ مرکز هر خوشه را با میانگین‌گیری از داده‌های انتساب یافته به آن خوشه به روز رسانی کن.

□ توقف: زمانی که در یک تکرار هیچ داده‌ای خوشه‌ی خود را عوض نکند.



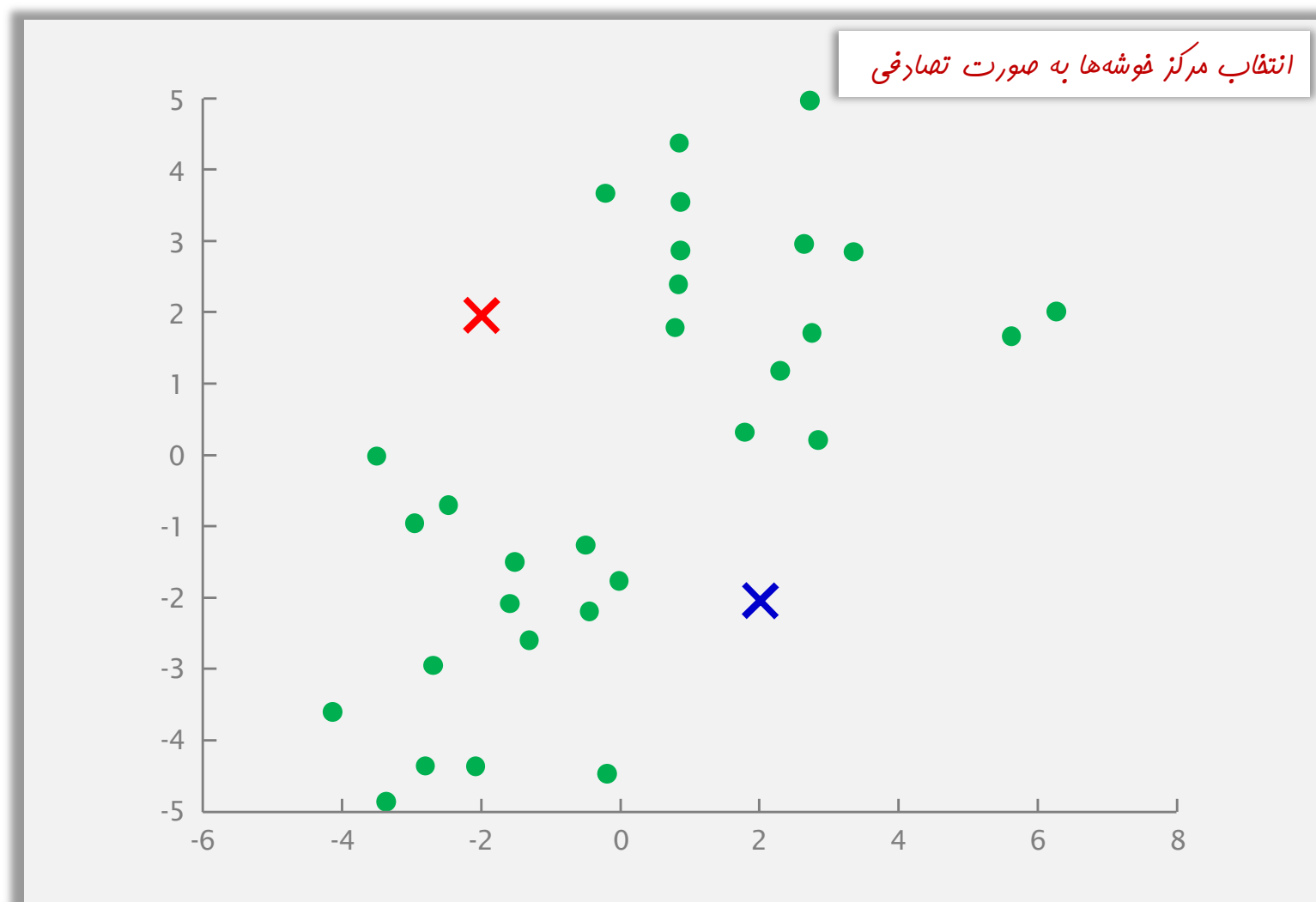
الگوریتم K-means: اجرای نمایشی

۱۳



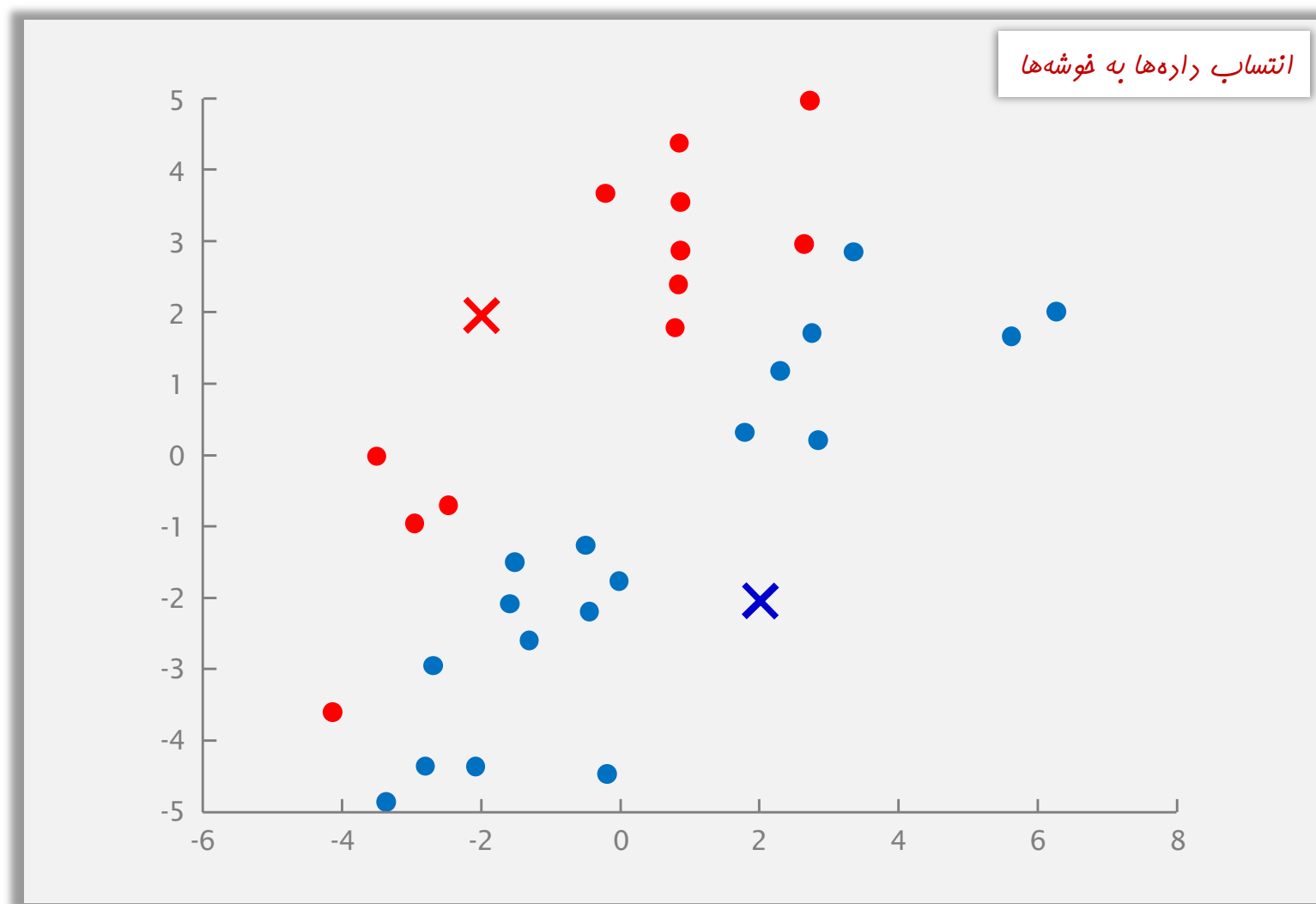
الگوریتم K-means: اجرای نمایشی

۱۴



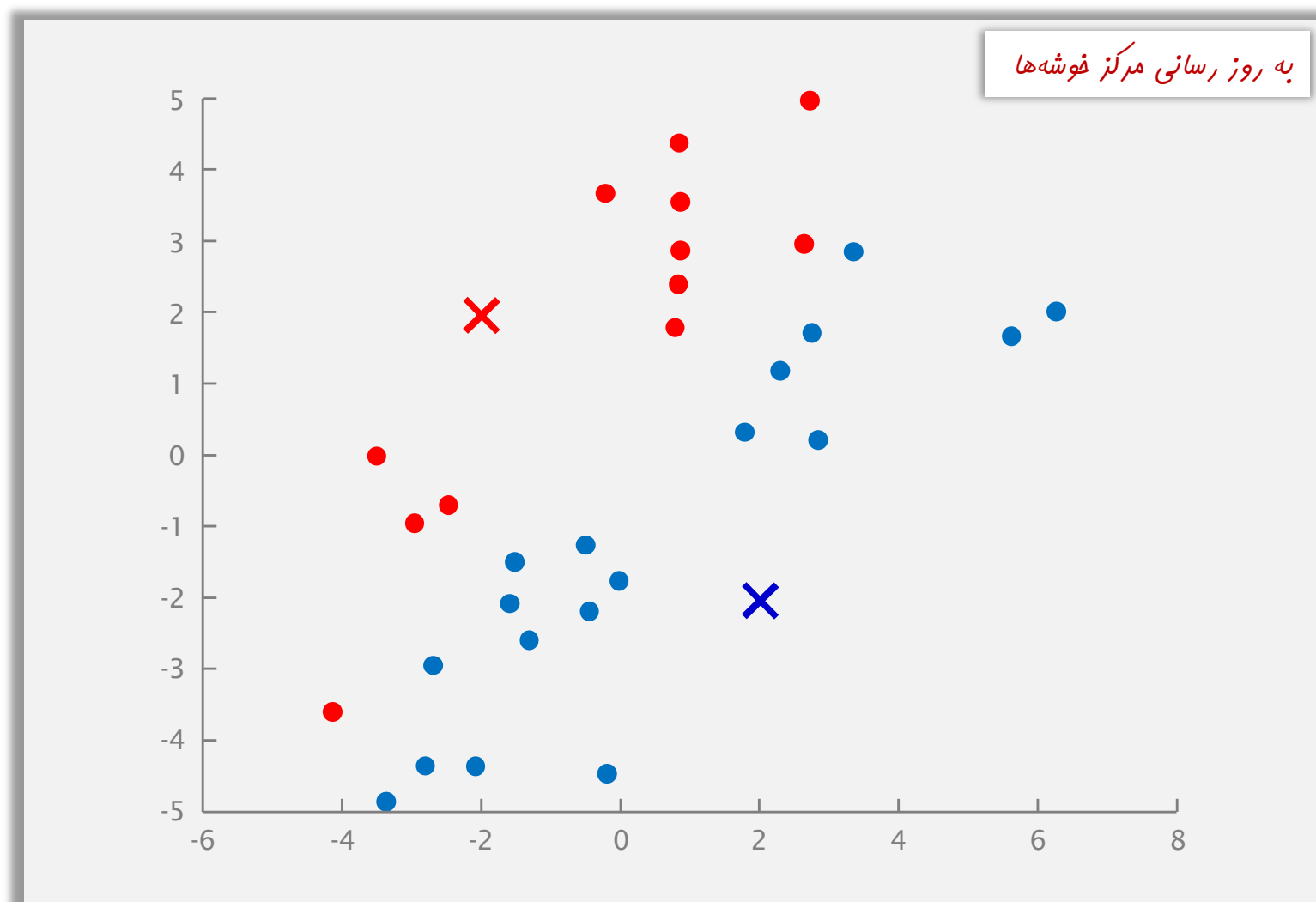
الگوریتم K-means: اجرای نمایشی

۱۵



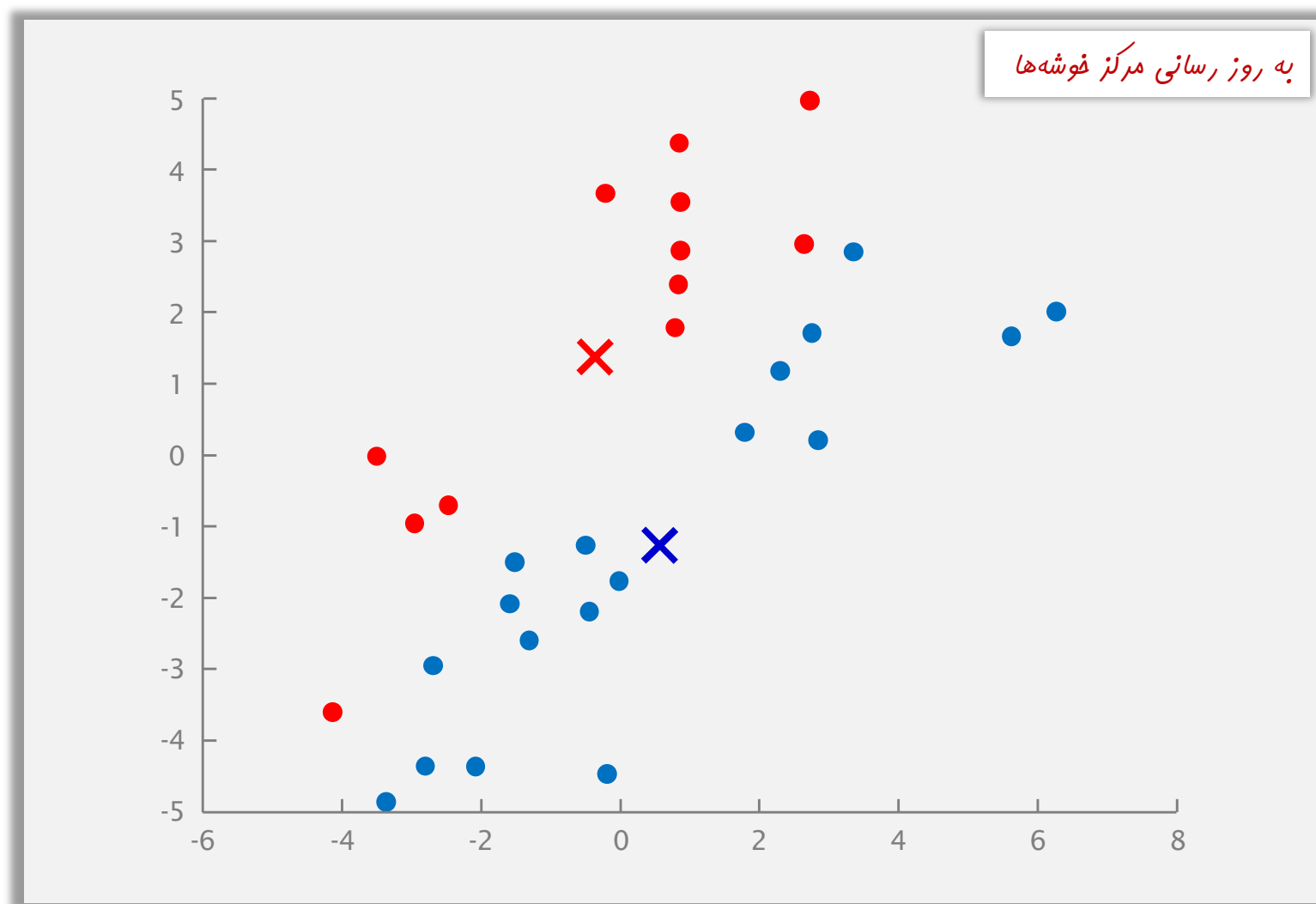
الگوریتم K-means: اجرای نمایشی

۱۶



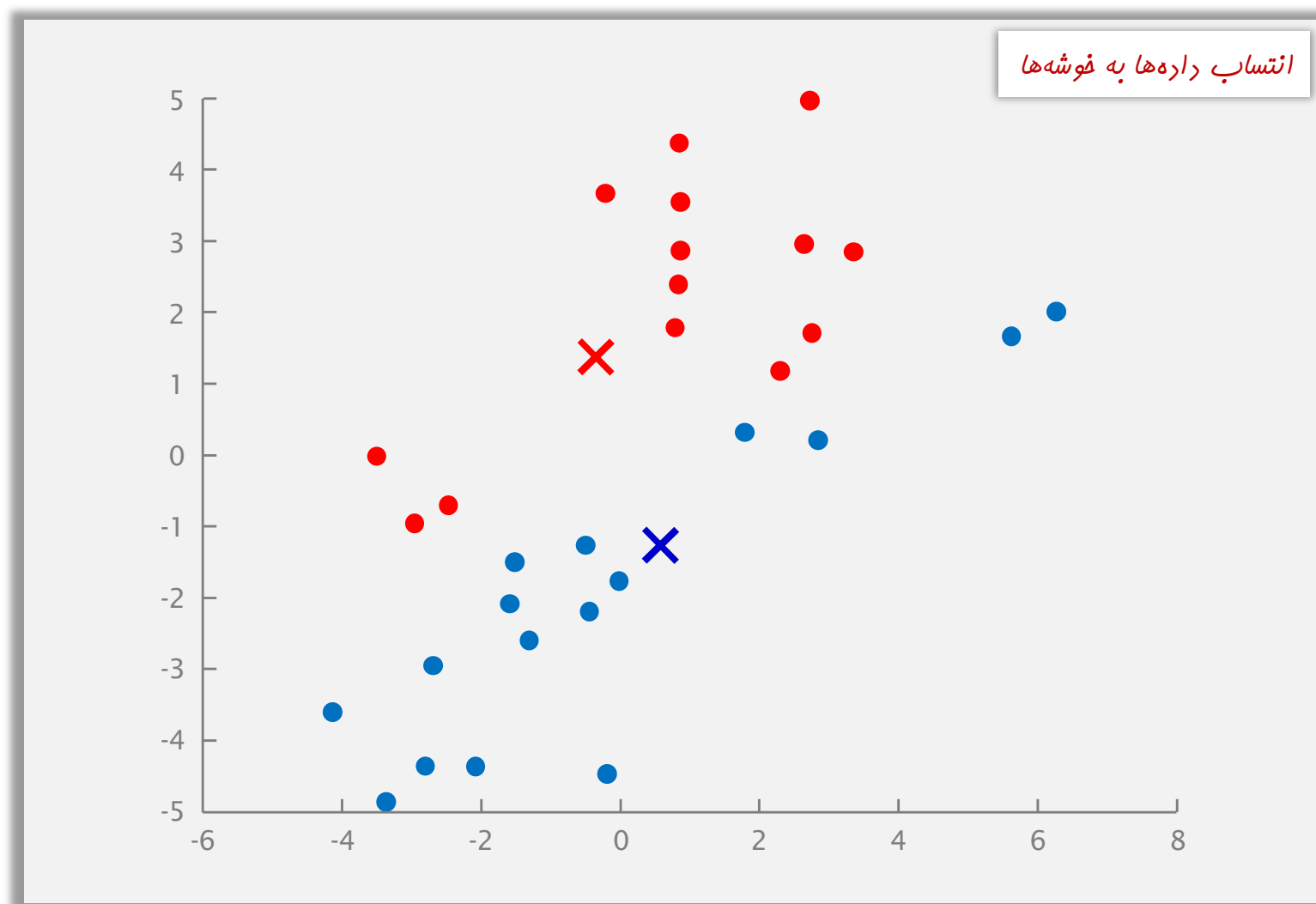
الگوریتم K-means: اجرای نمایشی

۱۷



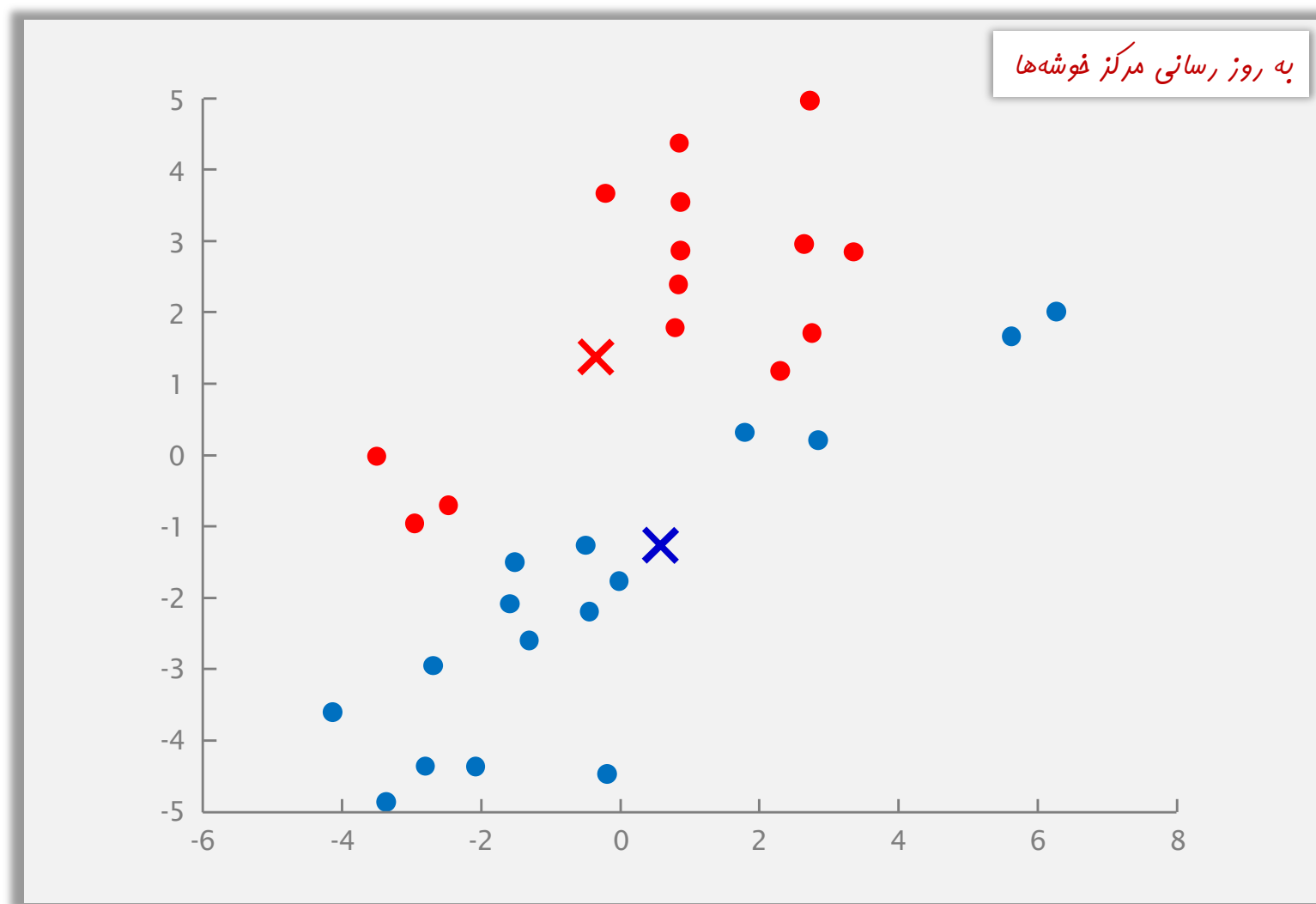
الگوریتم K-means: اجرای نمایشی

۱۸



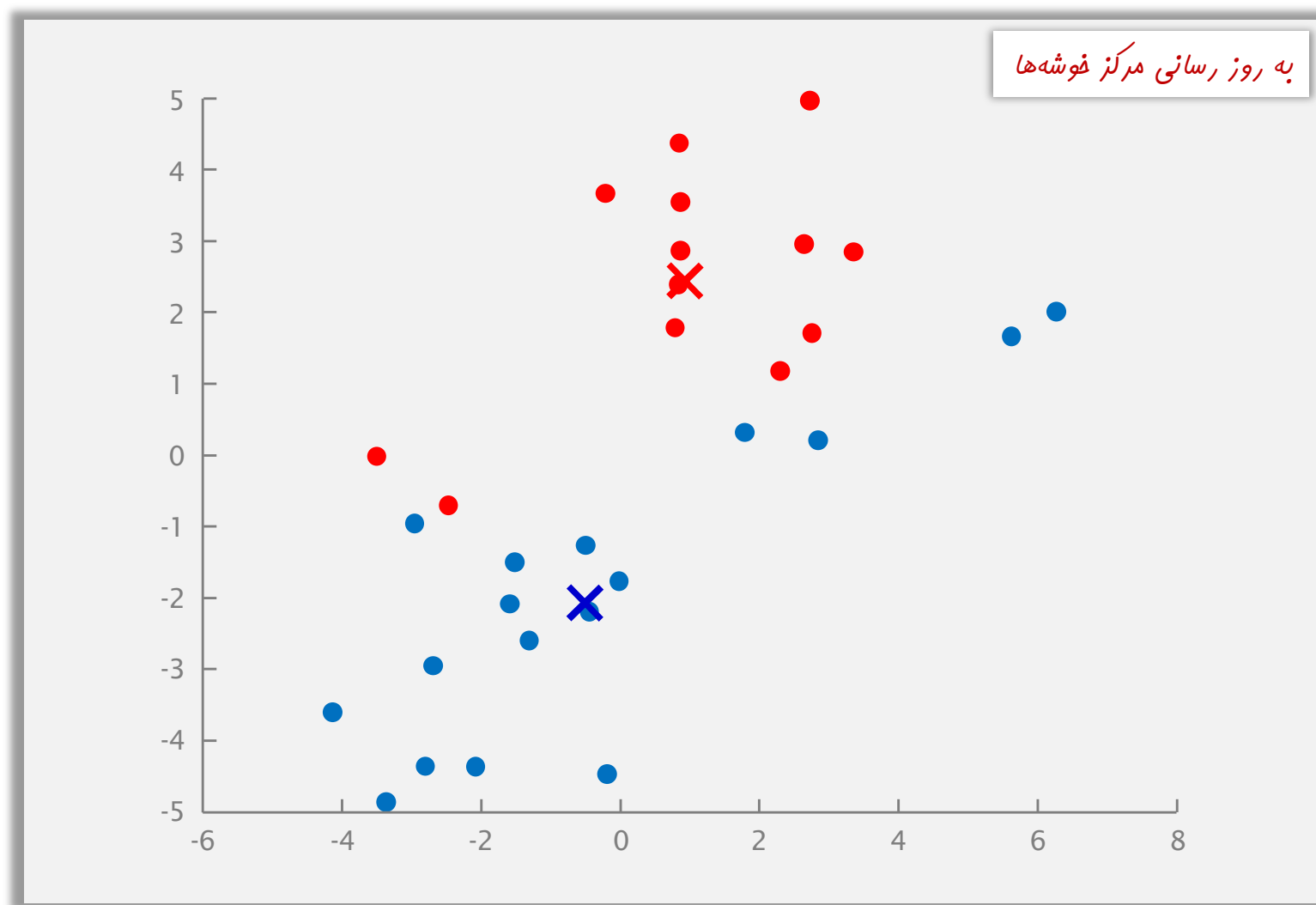
الگوریتم K-means: اجرای نمایشی

۱۹



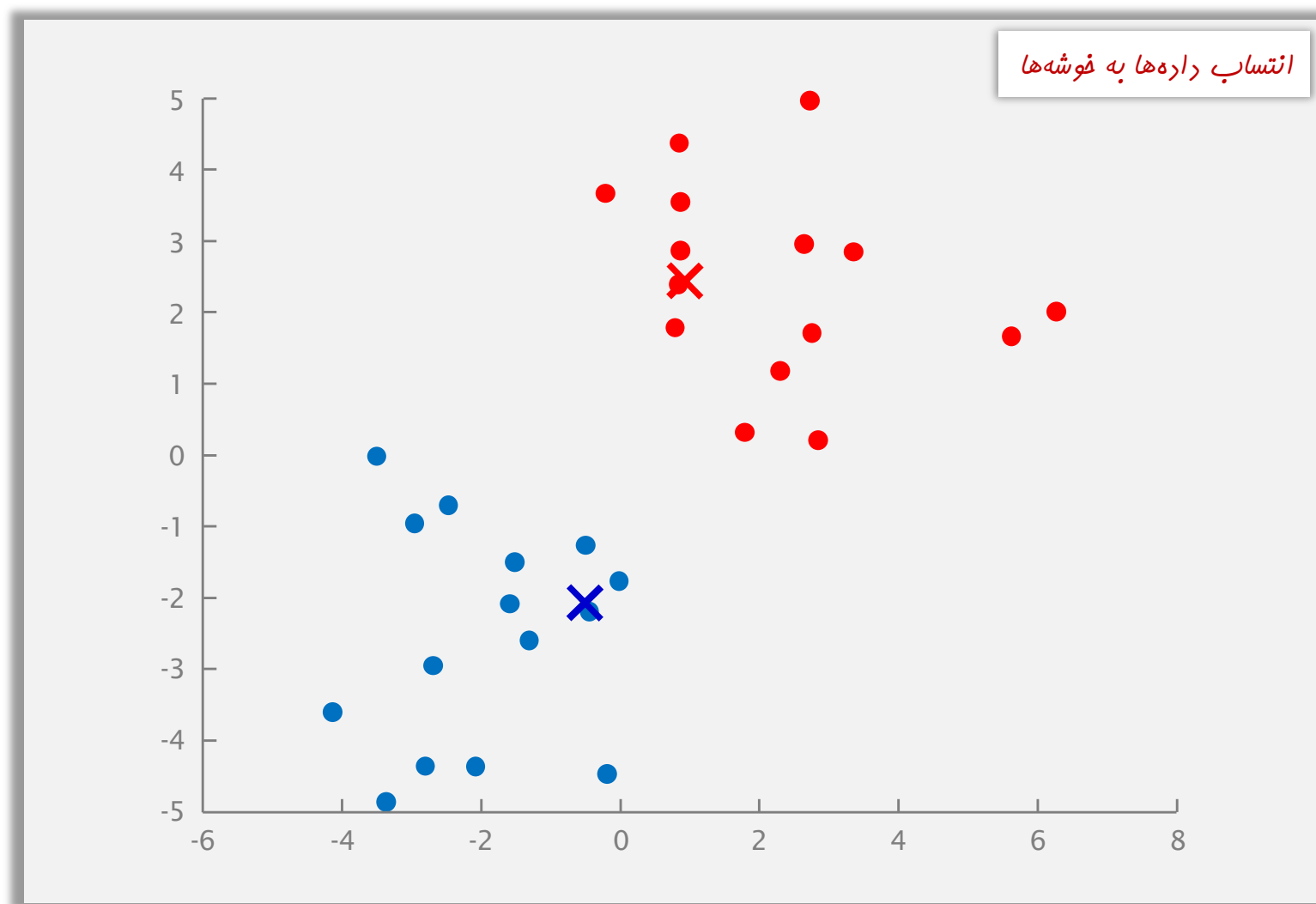
الگوریتم K-means: اجرای نمایشی

۲۰



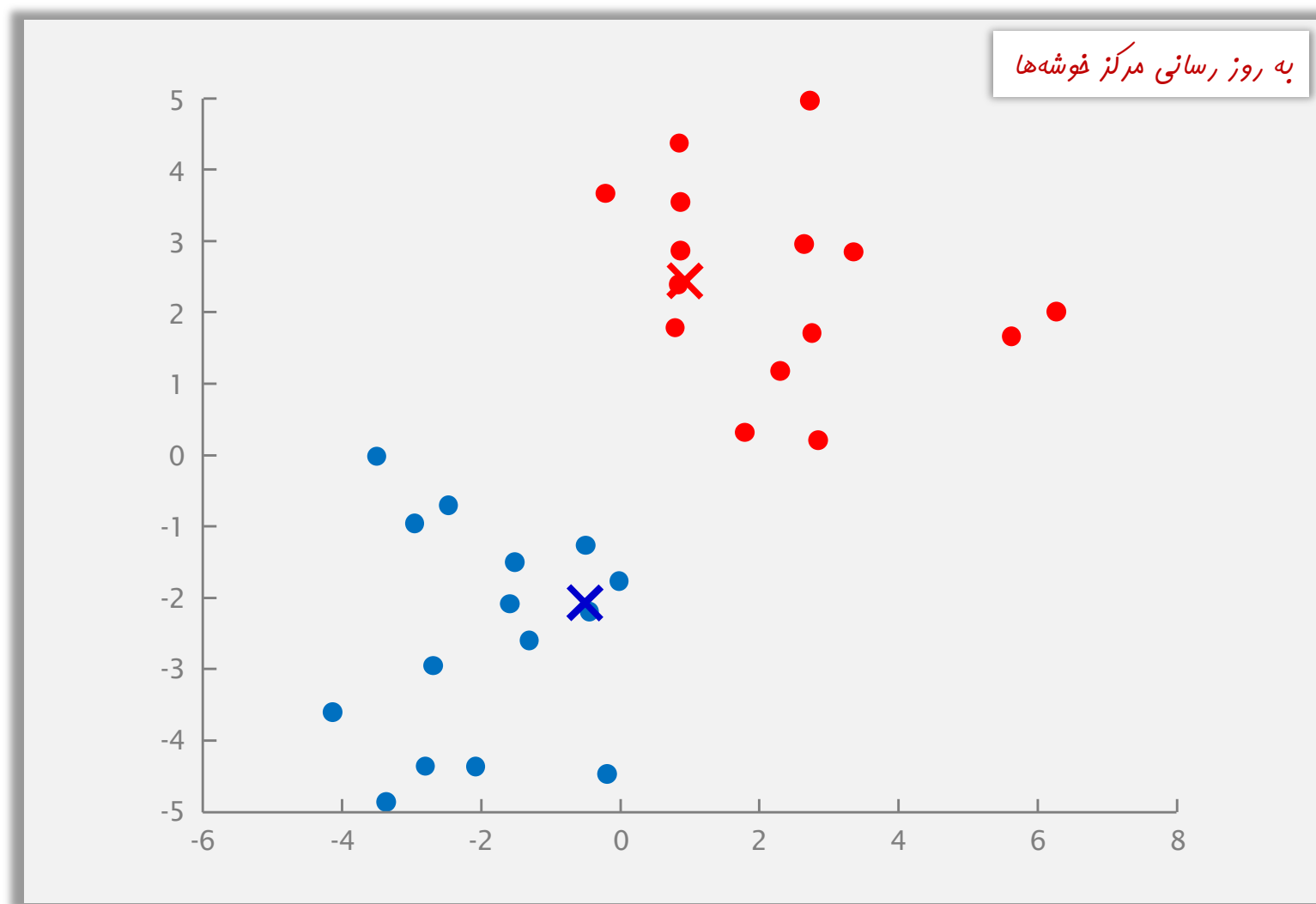
الگوریتم K-means: اجرای نمایشی

۲۱



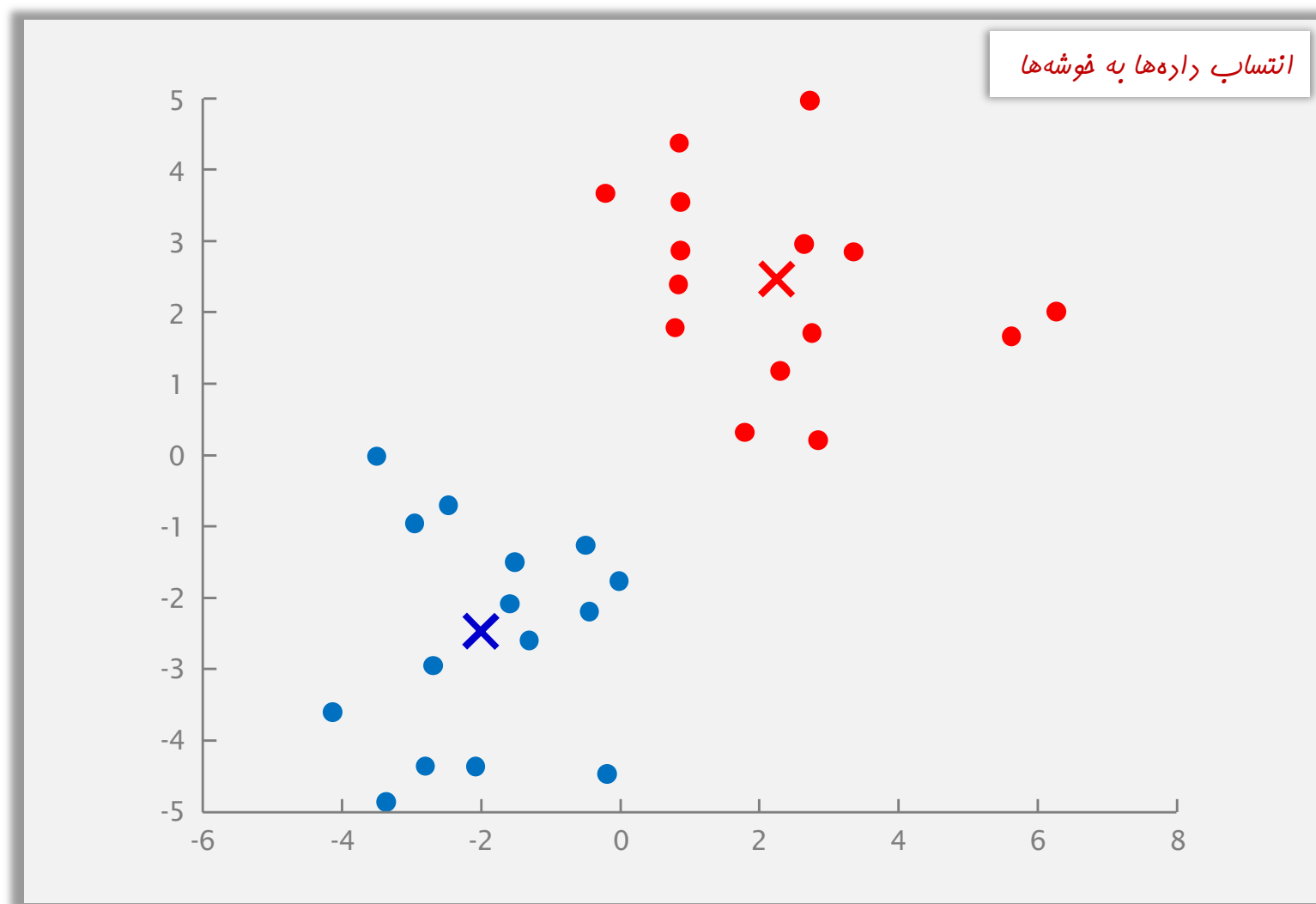
الگوریتم K-means: اجرای نمایشی

۲۲



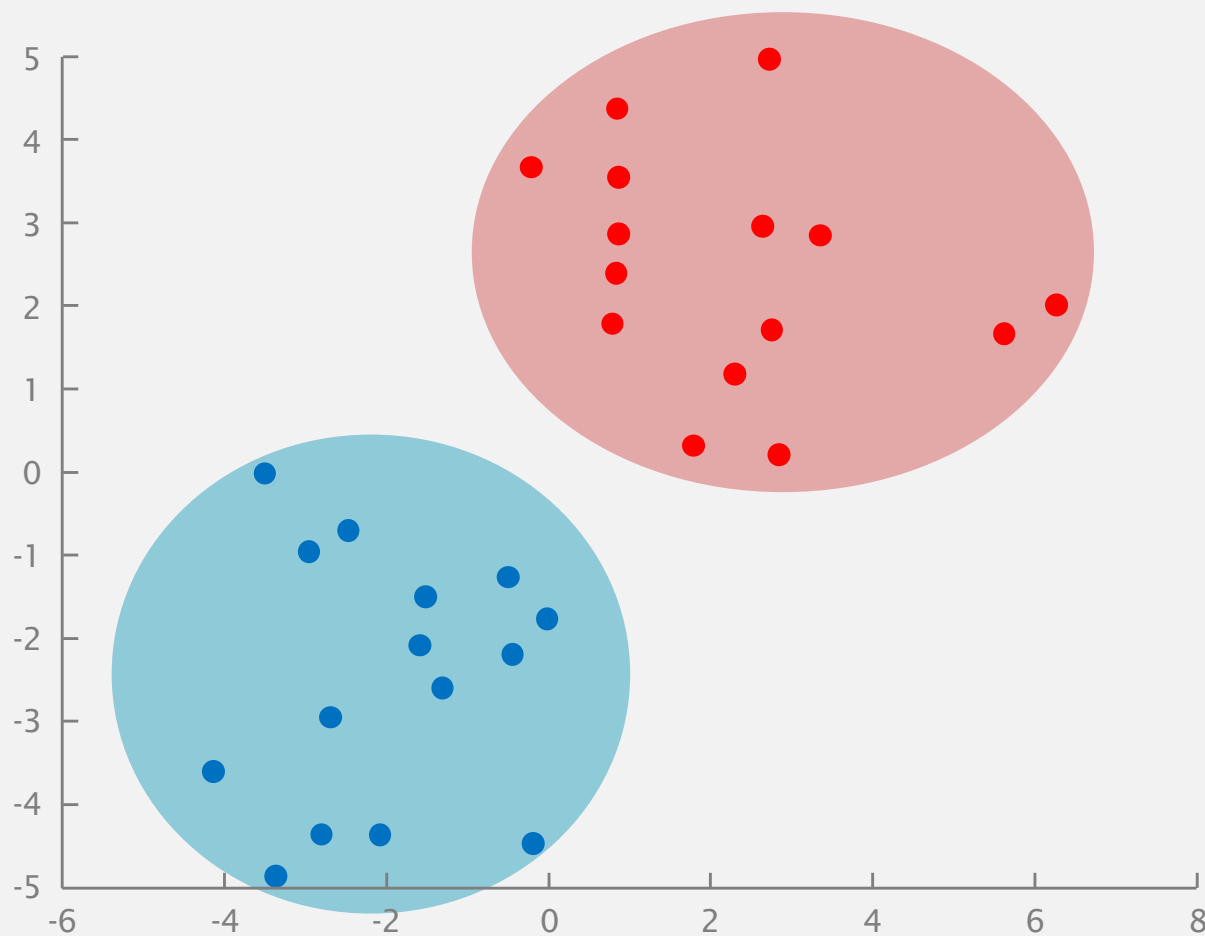
الگوریتم K-means: اجرای نمایشی

۲۳



الگوریتم K-means: اجرای نمایشی

۲۴



الگوریتم K-means

۲۵

□ ورودی.

□ تعداد خوشه‌ها: k

□ مجموعه‌ی آموزشی: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

□ توجه. در مجموعه‌ی آموزشی هیچ برچسبی برای داده‌ها تعیین نشده است.

□ توجه. در خوشه‌بندی نیازی به افزودن ویژگی $x_0 = 1$ نیست.

الگوریتم K-means

۲۶

randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

repeat

{

for $i = 1$ **to** m

انتساب داده‌ها به خوشه‌ها

$$c^{(i)} = \arg \min_k \|x^{(i)} - \mu_k\|$$

for $k = 1$ **to** K

به روز رسانی مرکز خوشه‌ها

μ_k = average of points assigned to cluster k

}

خوشه‌بندی: تابع هدف

تابع هدف

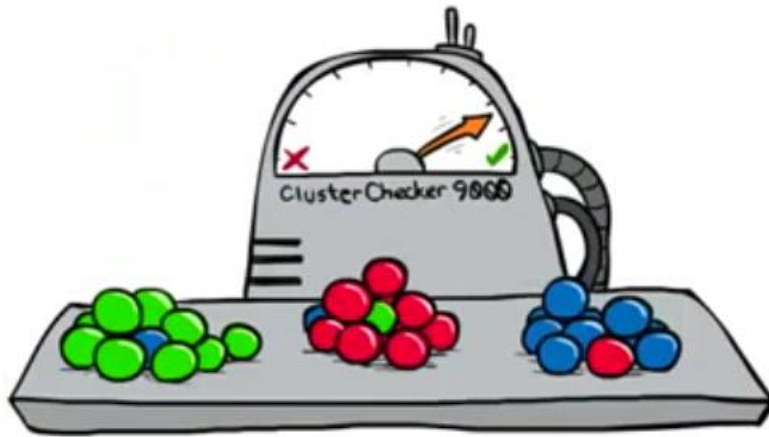
۲۸

□ نمادها.

□ μ_k : مرکز خوشه‌ی k

□ $c^{(i)}$: شماره‌ی خوشه‌ی اختصاص یافته به داده‌ی $x^{(i)}$

□ $\mu_{c^{(i)}}$: مرکز خوشه‌ی اختصاص یافته به داده‌ی $x^{(i)}$



□ تابع هدف.

$$J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

الگوریتم K-means

۲۹

randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

repeat

{

for $i = 1$ **to** m

$$c^{(i)} = \arg \min_k \|x^{(i)} - \mu_k\|$$

کمینه‌سازی تابع هدف
نسبت به پارامترهای $c^{(i)}$

for $k = 1$ **to** K

μ_k = average of points assigned to cluster k

کمینه‌سازی تابع هدف
نسبت به پارامترهای μ

}

مقداردهی اولیه به مراکز خوشه‌ها

الگوریتم K-means

۳۱

randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$

repeat

{

for $i = 1$ **to** m

$$c^{(i)} = \arg \min_k \|x^{(i)} - \mu_k\|$$

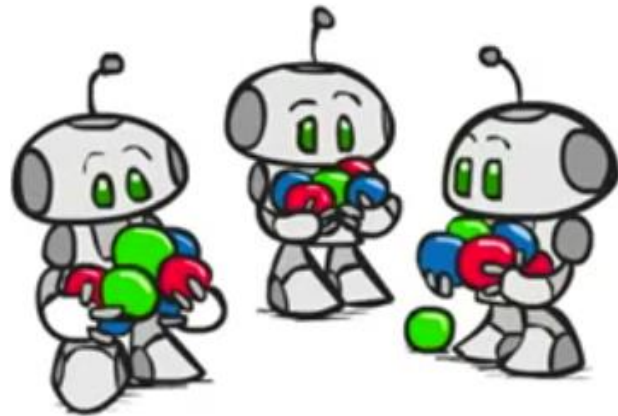
for $k = 1$ **to** K

μ_k = average of points assigned to cluster k

}

مقداردهی اولیه به مرکز خوشه‌ها

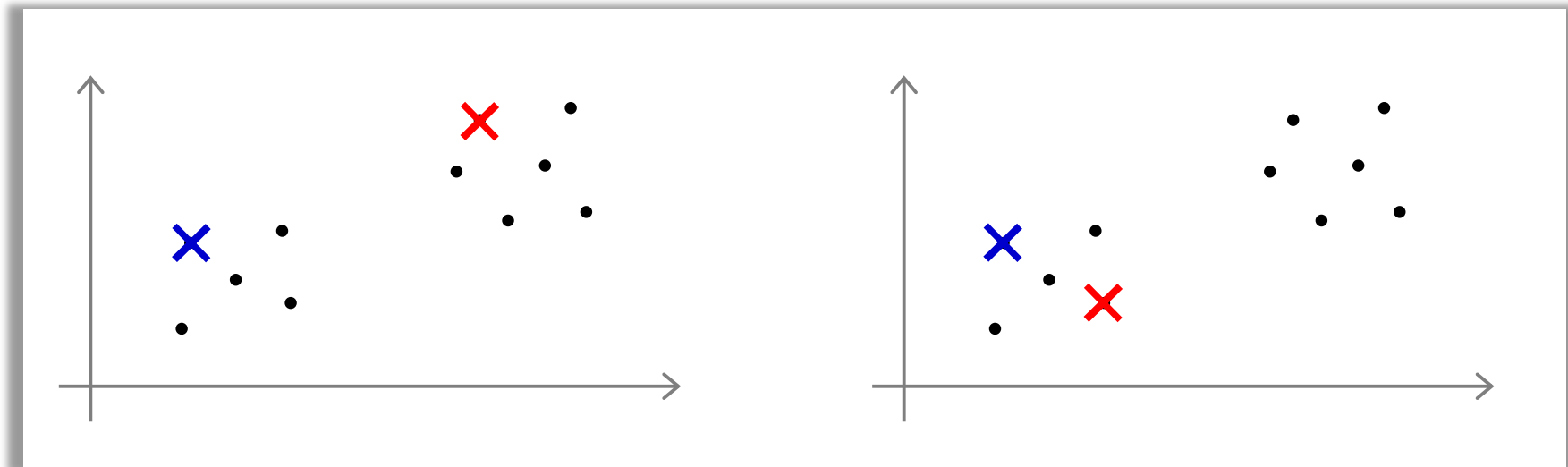
۳۲



□ مقداردهی اولیه. $[K \leq m]$

□ انتخاب K نمونه‌ی آموزشی به صورت تصادفی

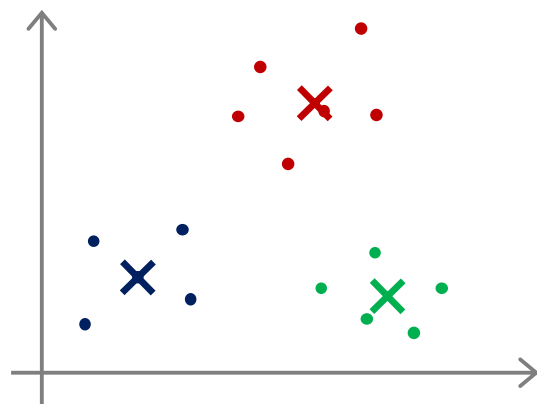
□ انتساب مراکز خوشه‌ها به K نمونه‌ی انتخاب شده



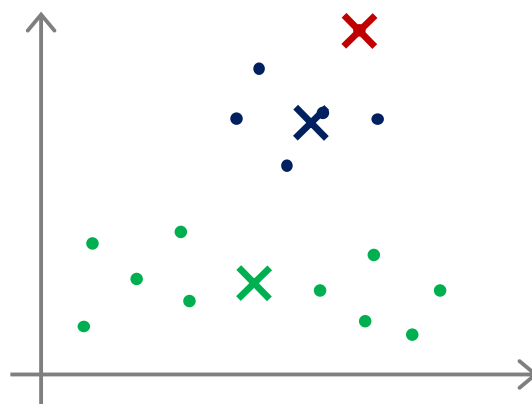
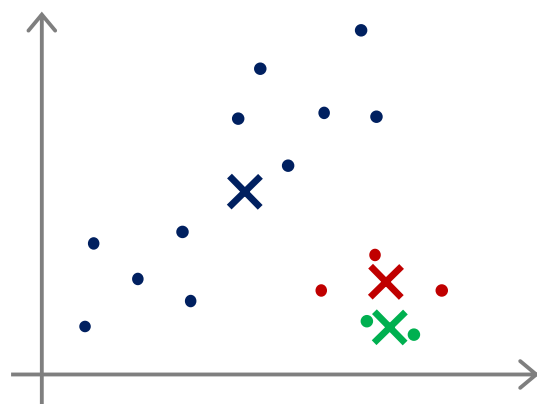
بهینه‌ی محلی

۳۳

□ بهینه‌ی سراسری.



□ بهینه‌ی محلی.



اجتناب از بهینه‌های محلی

۳۴

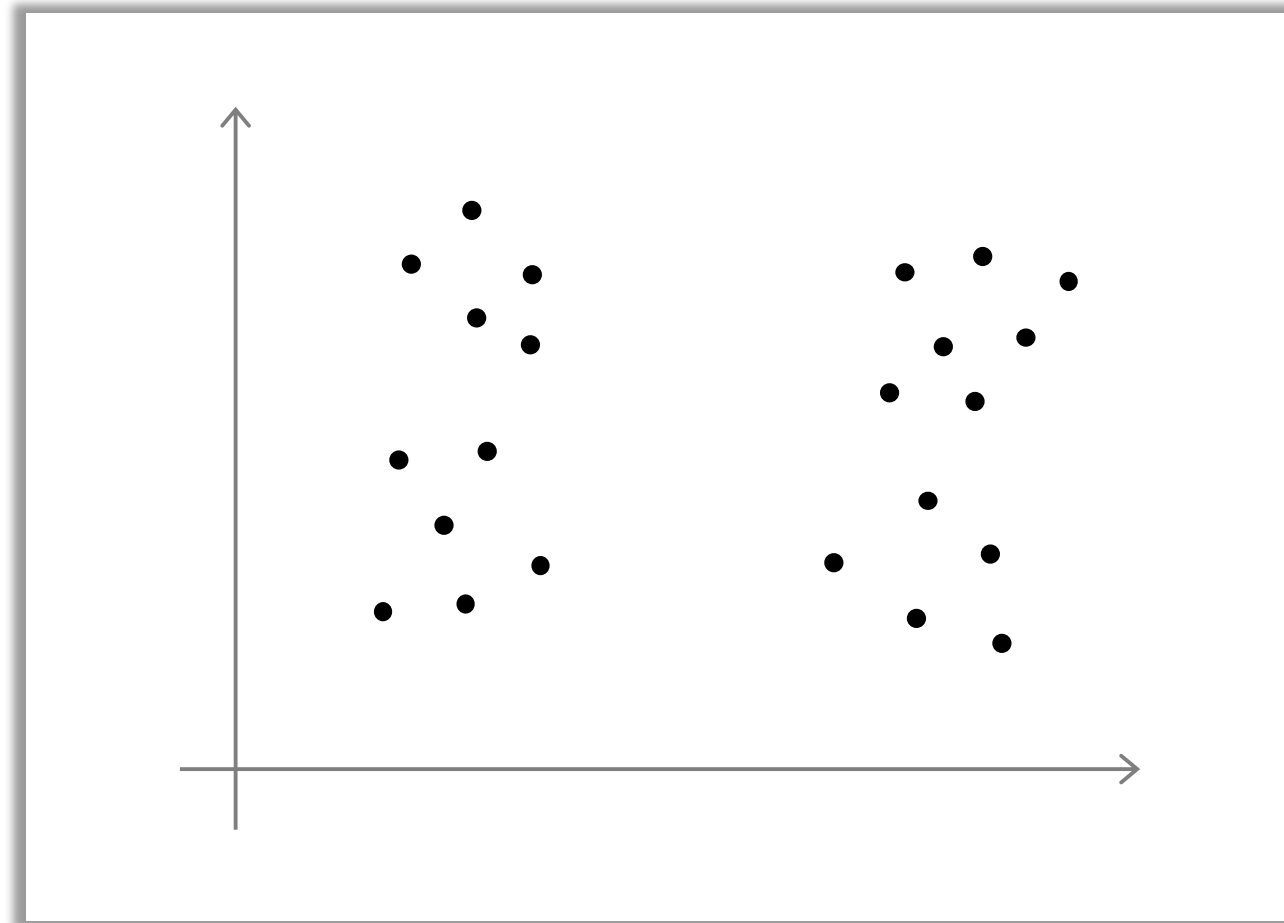
```
for  $t = 1$  to  $MAX$ 
{
    randomly initialize cluster centroids  $\mu_1, \mu_2, \dots, \mu_k$ 
    run K-means to get  $c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k$ 
    compute cost function  $J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k)$ 
}
pick clustering with minimum cost
```

تعیین تعداد خوشه‌ها



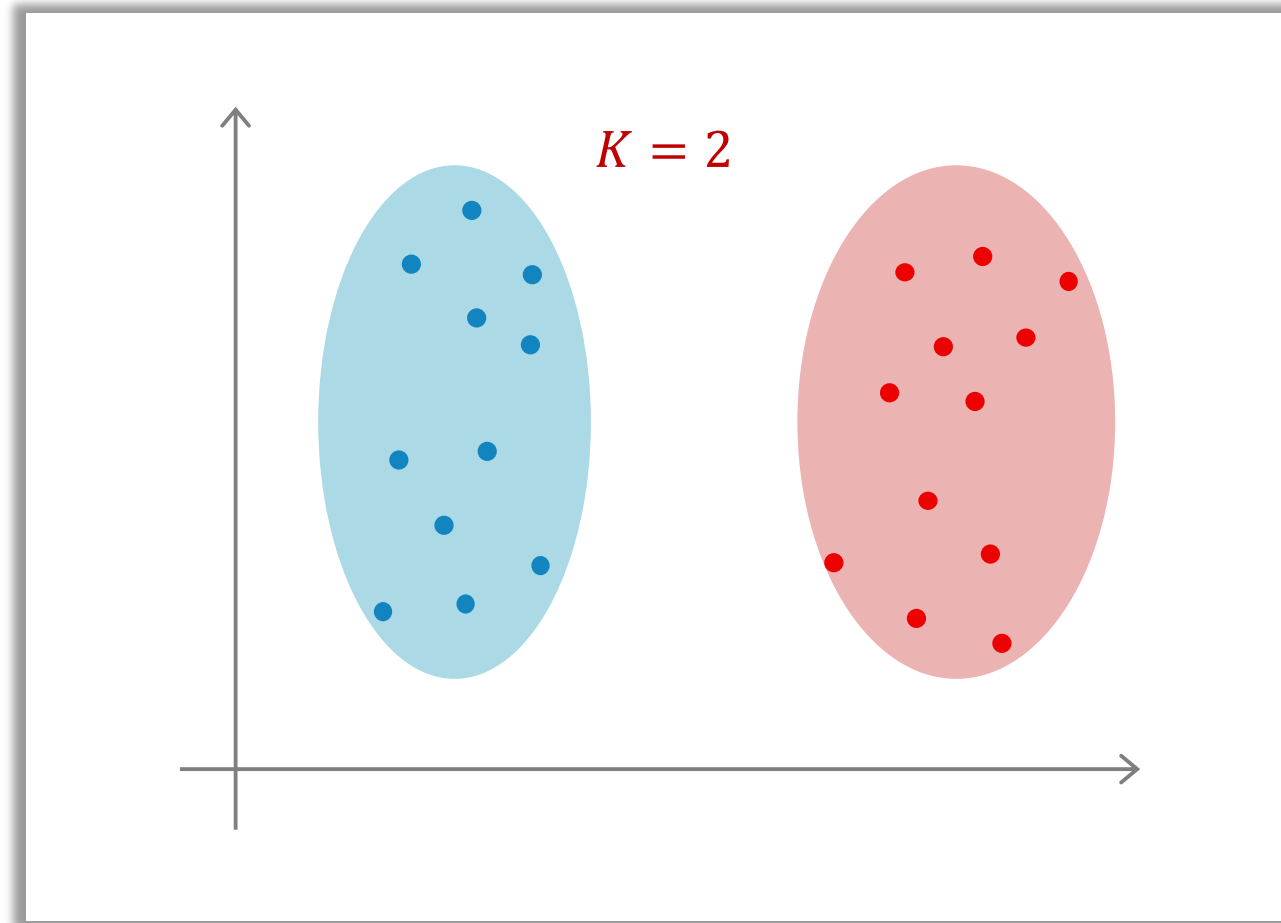
مقدار مناسب برای K کدام است؟

۳۶



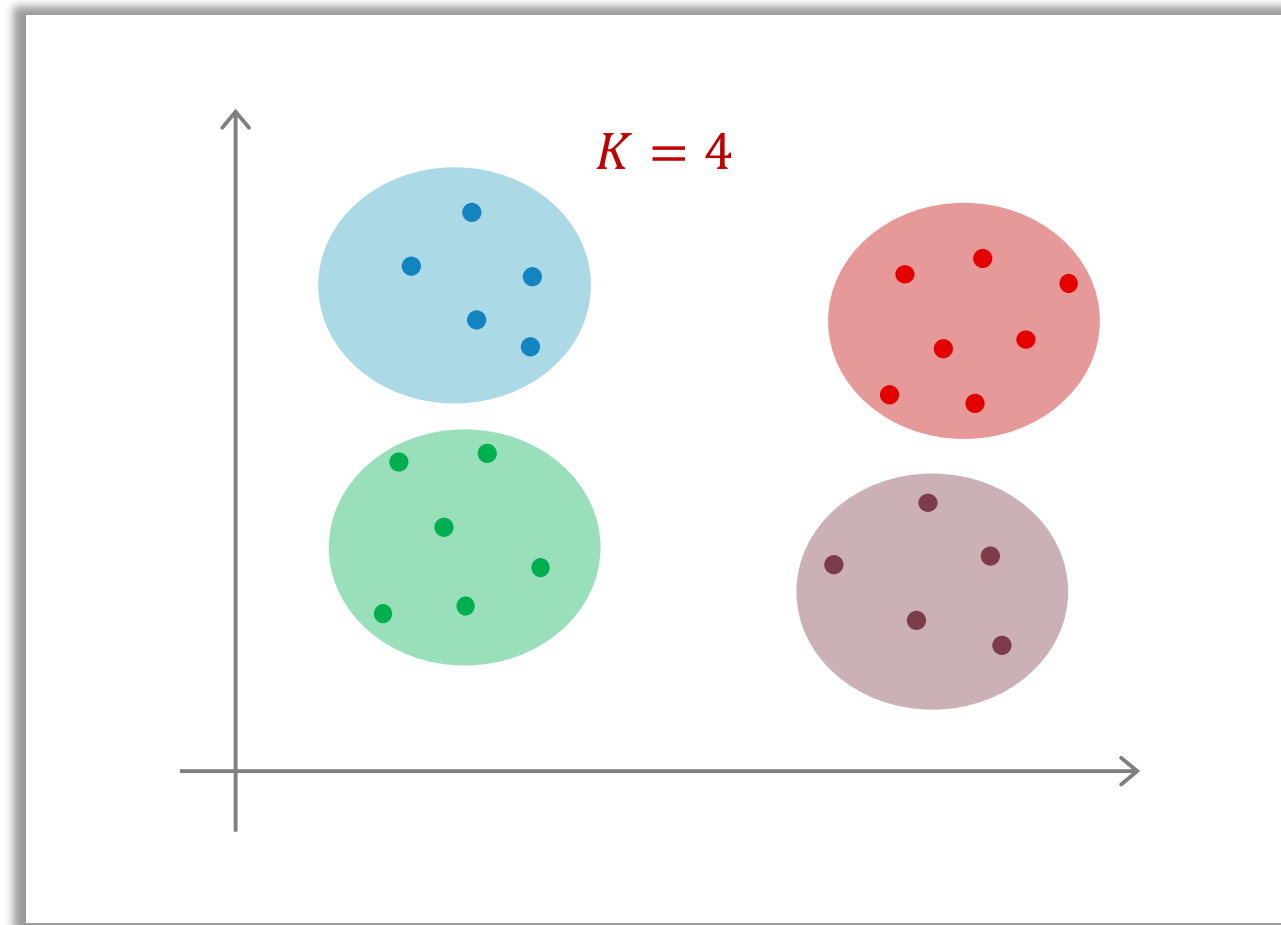
مقدار مناسب برای K کدام است؟

۳۷



مقدار مناسب برای K کدام است؟

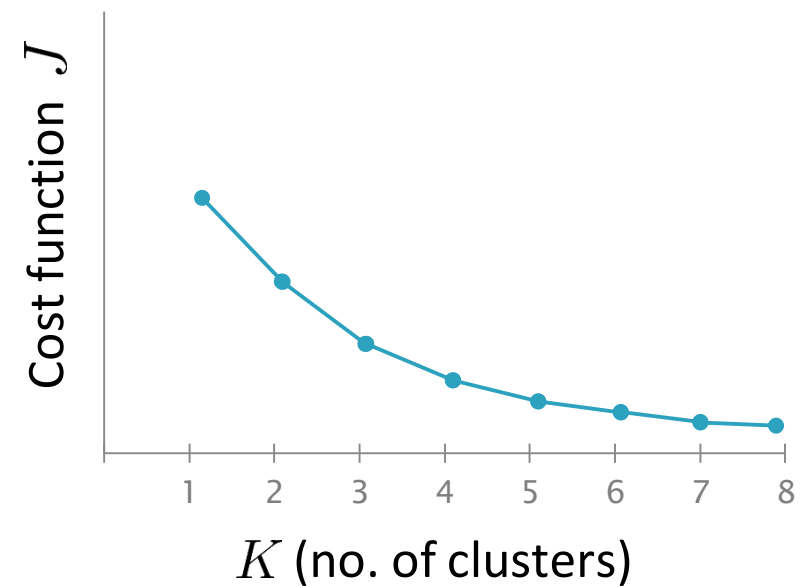
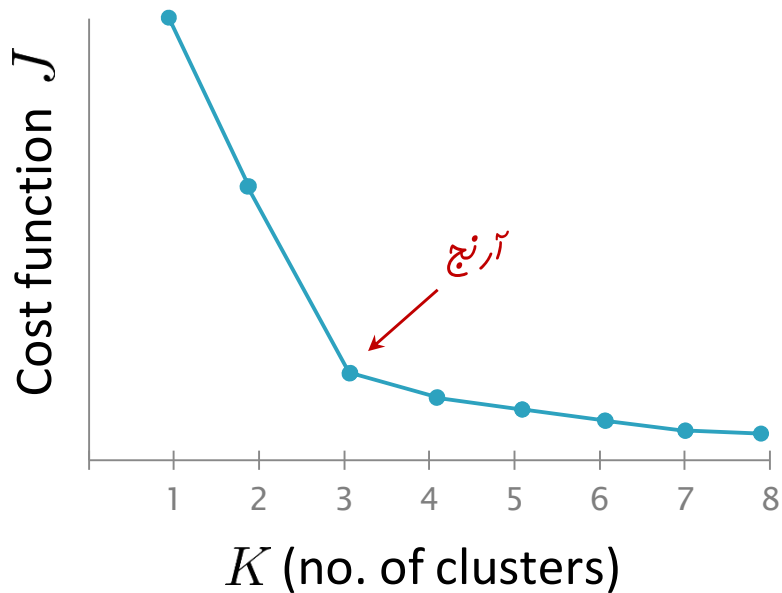
۳۸



تعیین تعداد مناسب خوشه‌ها

۳۹

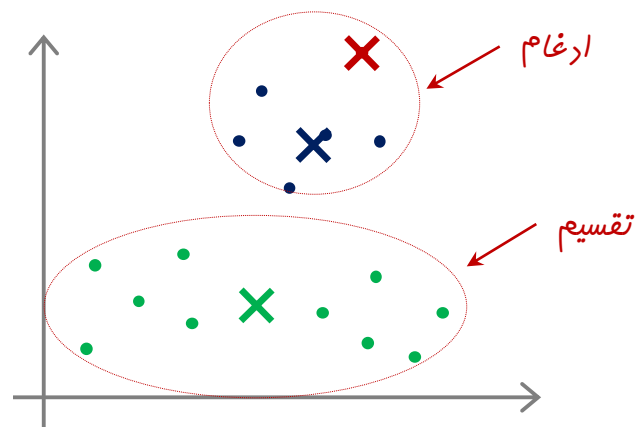
□ روش «آرنج».



بهبود خوشه‌بندی

بهبود خوشه‌بندی با پس‌پردازش خوشه‌ها

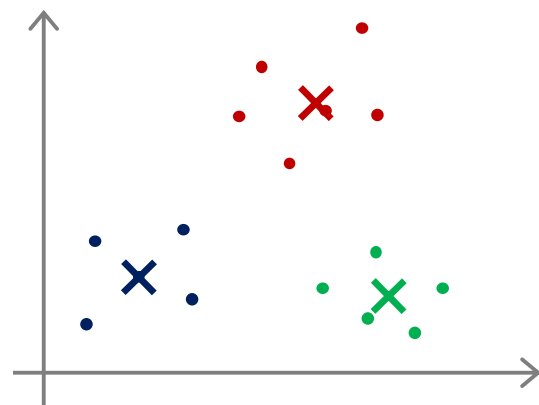
۴۱



□ تقسیم.

□ تقسیم یک خوشه با بیشترین خطا به دو خوشه

با اجرای K-means بر روی داده‌های
این خوشه با مقدار $K = 2$



□ ادغام.

□ ادغام نزدیک‌ترین دو خوشه

□ ادغام دو خوشه با حداقل افزایش در مجموع خطا

الگوریتم K-means دو بخشی ساز

۴۲

□ الگوریتم دو بخشی ساز.

□ با یک خوشه شامل تمامی داده‌ها شروع کن.

□ هر بار یک خوشه را انتخاب کن:

■ خوشه‌ی انتخاب شده را به وسیله‌ی الگوریتم K-means به دو خوشه تقسیم کن.

■ مجموع خطای خوشه‌بندی را محاسبه کن.

■ خوشه‌بندی با کمترین خطا را انتخاب کن.

□ عمل فوق را تا زمان رسیدن به تعداد خوشه‌های مورد نظر تکرار کن.

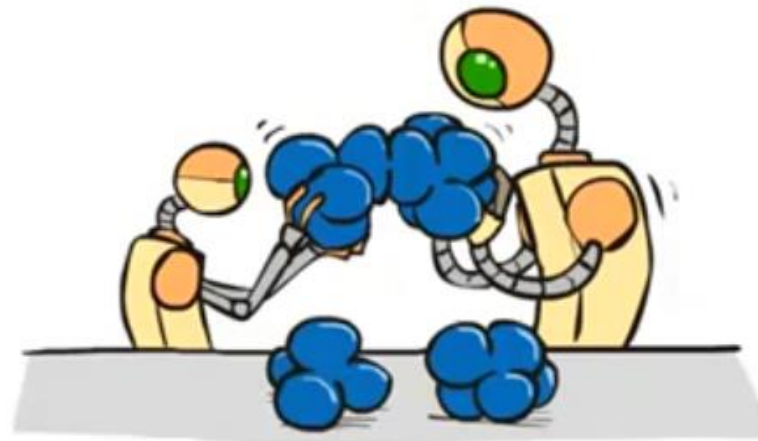
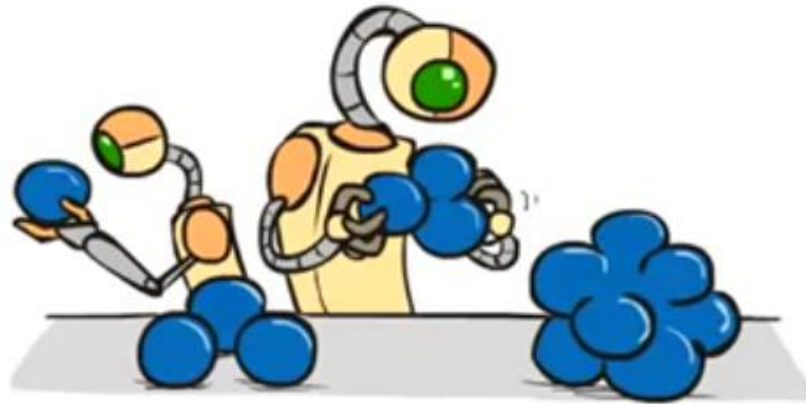
الگوریتم K-means دو بخشی ساز

۴۳

Start with all the points in one cluster
while the number of clusters is less than K
 measure the total error
 for every cluster
 perform K-means clustering with $k = 2$ on the given cluster
 measure the total error after splitting
 choose the cluster split that gives the lowest error

خوشه‌بندی سلسله‌مراتبی

۴۴



خوشه‌بندی سلسله‌مراتبی

۴۵

□ خوشه‌بندی سلسله‌مراتبی.

□ ابتدا داده‌های بسیار شبیه را ادغام کن.

□ به تدریج با ادغام خوشه‌های کوچک‌تر، خوشه‌های بزرگ‌تری ایجاد کن.

□ الگوریتم.

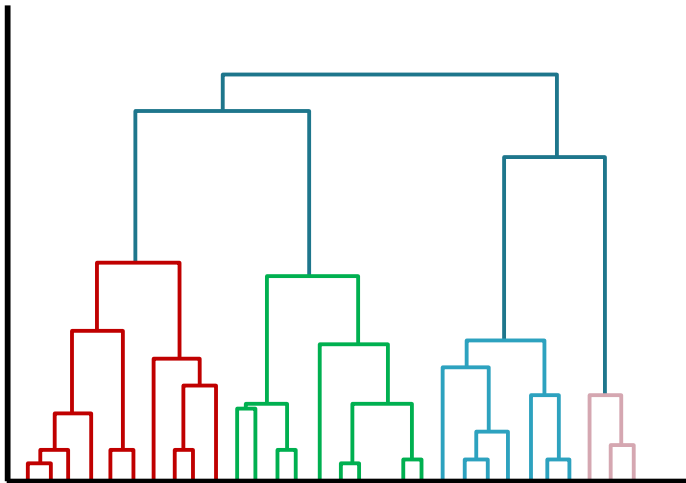
□ در ابتدا هر داده بیانگر یک خوشه است.

□ مراحل زیر را تکرار کن:

■ هر بار **نزدیک‌ترین** دو خوشه را انتخاب کن.

■ آن دو خوشه را در یک خوشه جدید ادغام کن.

■ توقف: زمانی که تنها یک خوشه باقی مانده باشد.

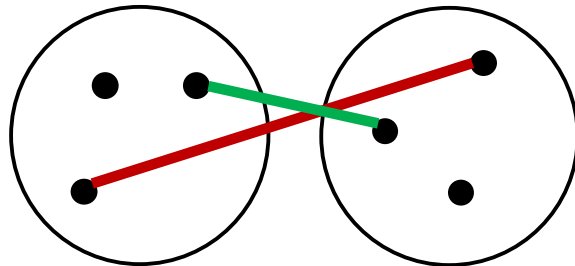


□ ایجاد یک **درخت نگاره** شامل یک طیف گسترده از خوشه‌بندی‌ها.

خوشه‌بندی سلسله‌مراتبی

۴۶

□ س. چگونه می‌توان نزدیک‌ترین دو خوشه را تعریف نمود؟



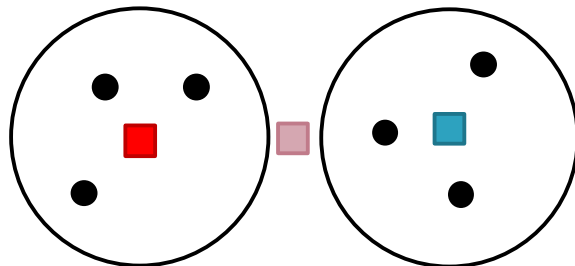
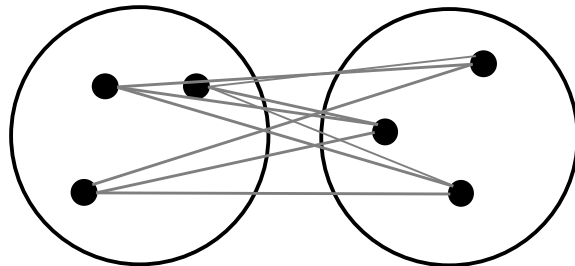
□ معیارهای تعیین شباهت خوشه‌ها.

□ نزدیک‌ترین زوج (خوشه‌بندی تک-پیوندی)

□ دورترین زوج (خوشه‌بندی تمام-پیوندی)

□ میانگین فاصله همه زوج‌ها

□ روش «وارد» (کمترین پراکندگی، مانند k-means)



□ معیارهای مختلف باعث ایجاد خوشه‌بندی‌های متفاوتی می‌شوند.

□ یادگیری بدون نظارت. یافتن ساختار در داده‌ها

□ خوشه‌بندی. گروه‌بندی داده‌های مشابه

□ الگوریتم خوشه‌بندی K-means

■ پیاده‌سازی آسان

■ برای مجموعه داده‌های بسیار بزرگ کند

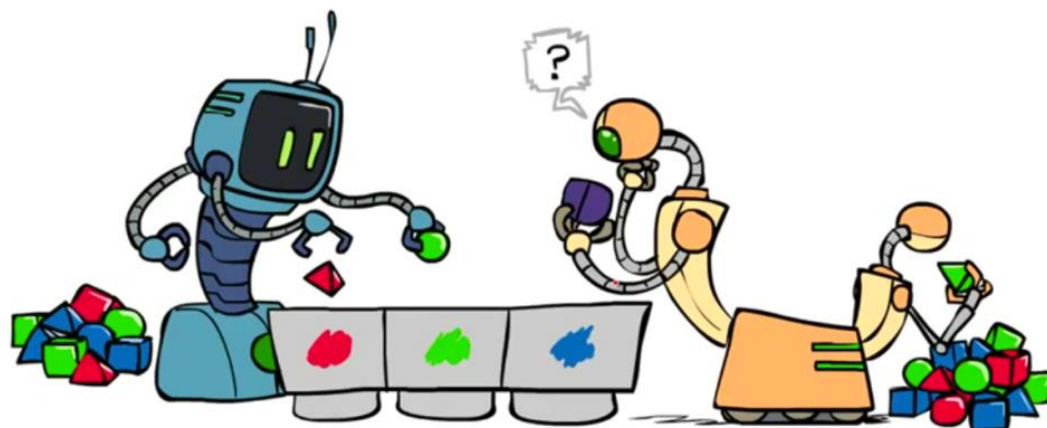
■ امکان گیر کردن در بهینه‌ی محلی

□ پس‌پردازش خوشه‌ها: تقسیم و ادغام خوشه‌ها

□ الگوریتم K-means دو بخشی‌ساز

■ خوشه‌بندی بهتر نسبت به الگوریتم K-means

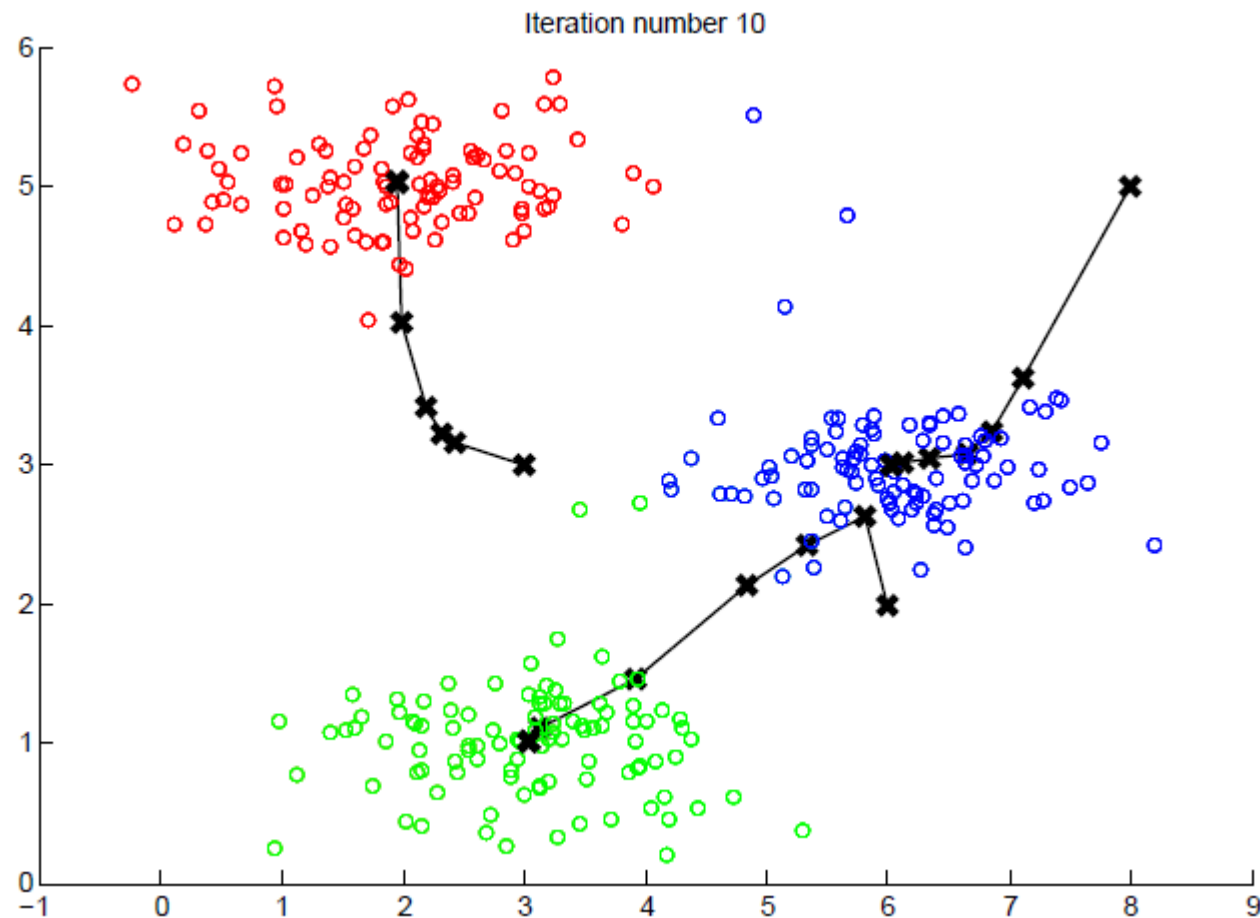
□ الگوریتم‌های خوشه‌بندی سلسله‌مراتبی



تمرین‌ها

تمرین ۱: پیاده‌سازی الگوریتم K-means

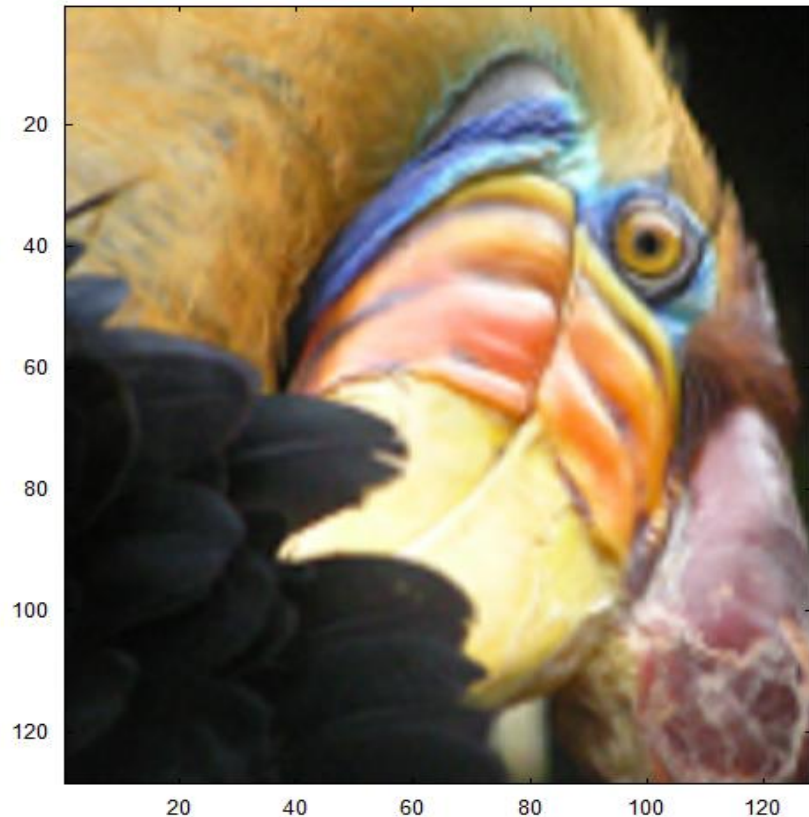
۴۹



تمرین ۲: فشرده‌سازی تصویر با استفاده از K-means

۵۰

تصویر اصلی



تصویر فشرده شده (۱۶ رنگ)

