یادگیری بدون نظارت: خوشه‌بندی

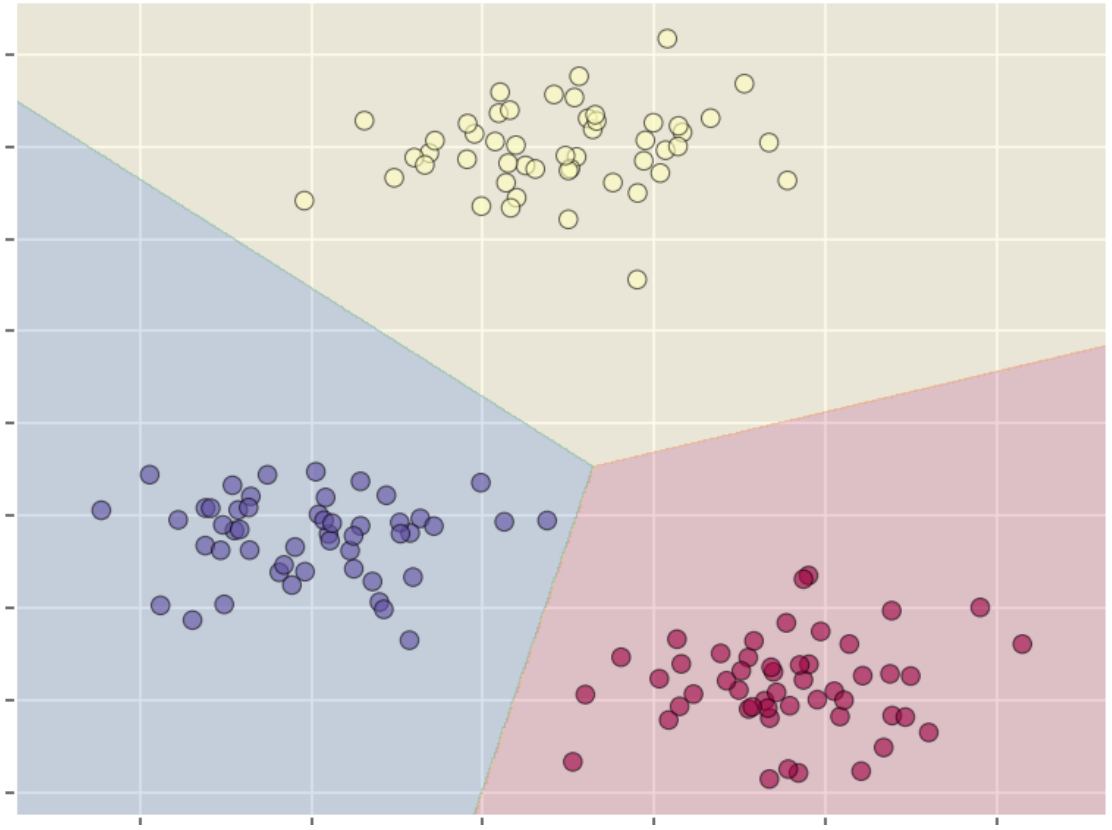
سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

فهرست مطالب

یادآوری: یادگیری نظارت شده

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



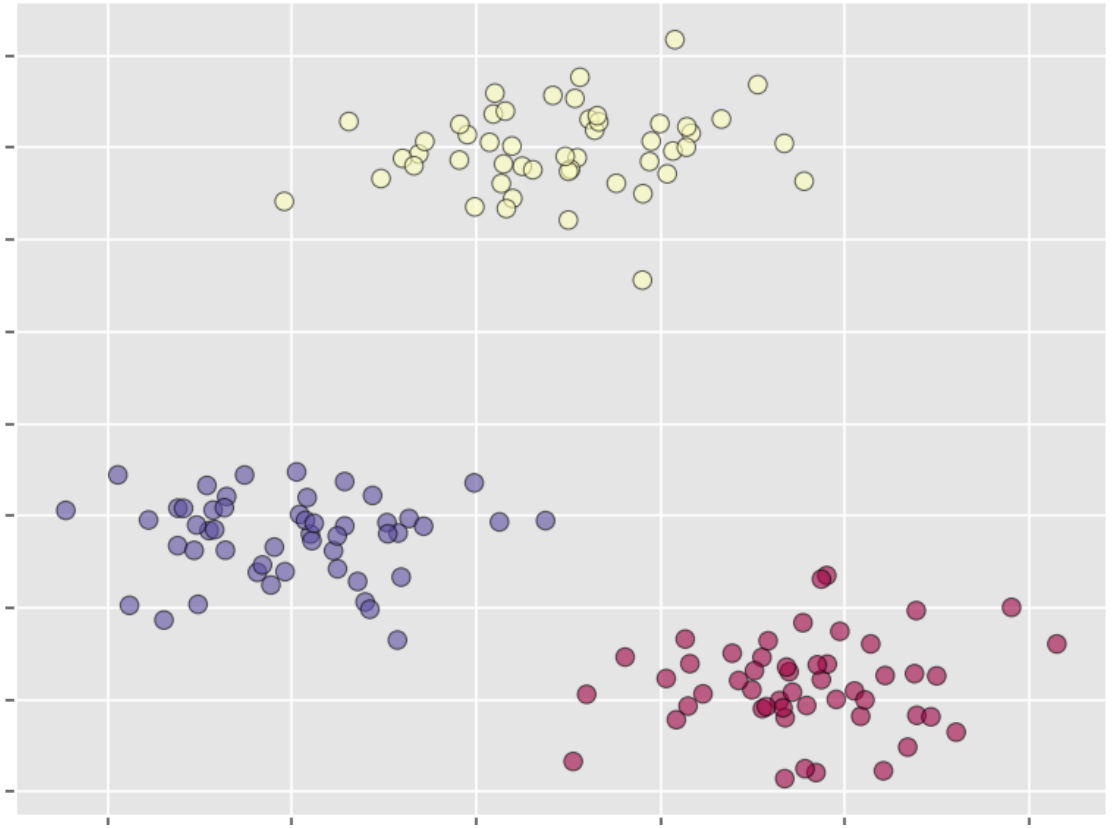
$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$$

مجموعه آموزشی

- انواع یادگیری نظارت شده.
  - رگرسیون: تخمین یک کمیت پیوسته
  - دسته‌بندی: تخمین یک کمیت گسسته

یادگیری بدون نظارت

- یادگیری بدون نظارت. عدم آگاهی از پاسخ‌های درست.



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

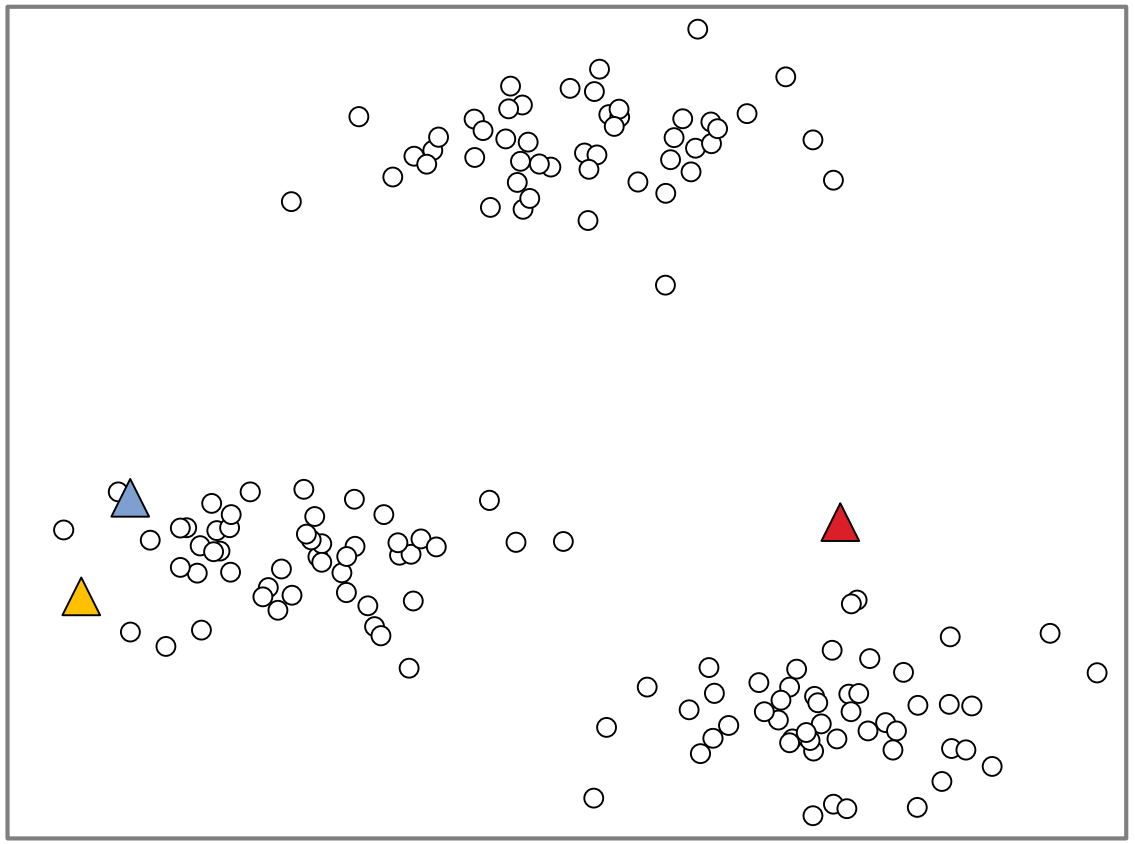
مجموعه آموزشی

- هدف. تشخیص ساختار در داده‌های ورودی

کاربرد خوشه‌بندی: گروه‌بندی اخبار مرتبط

الگوریتم خوشه‌بندی K-means

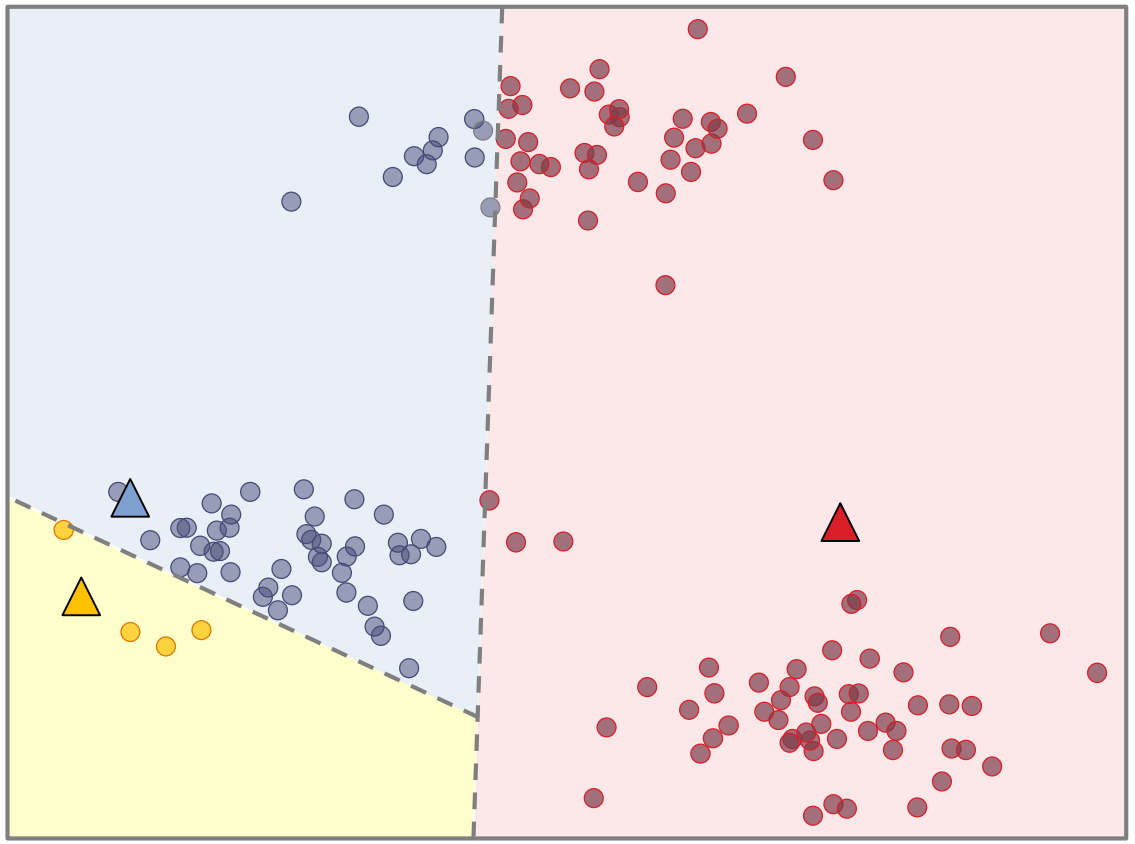
- یک الگوریتم خوشه‌بندی تکرار شونده.



- 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 خوشه خود را عوض نکند.

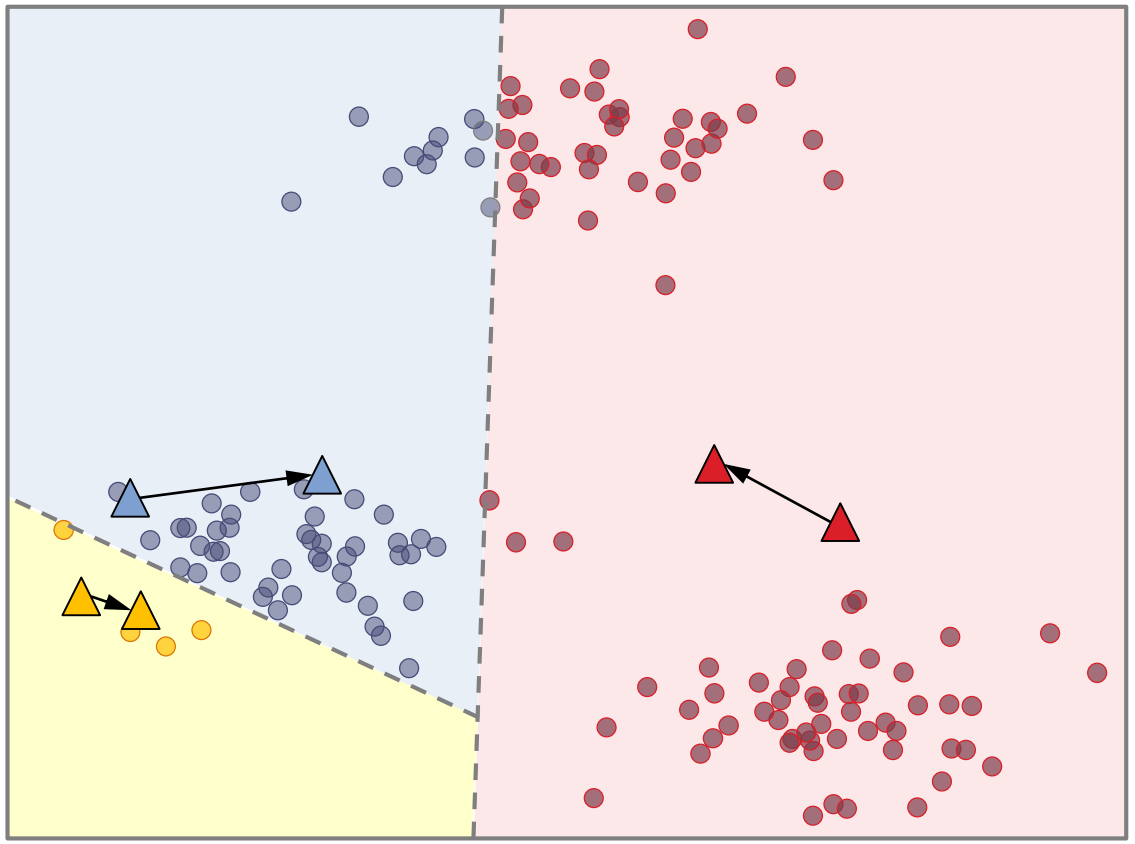
الگوریتم خوشه‌بندی K-means

- 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 خوشه خود را عوض نکند.

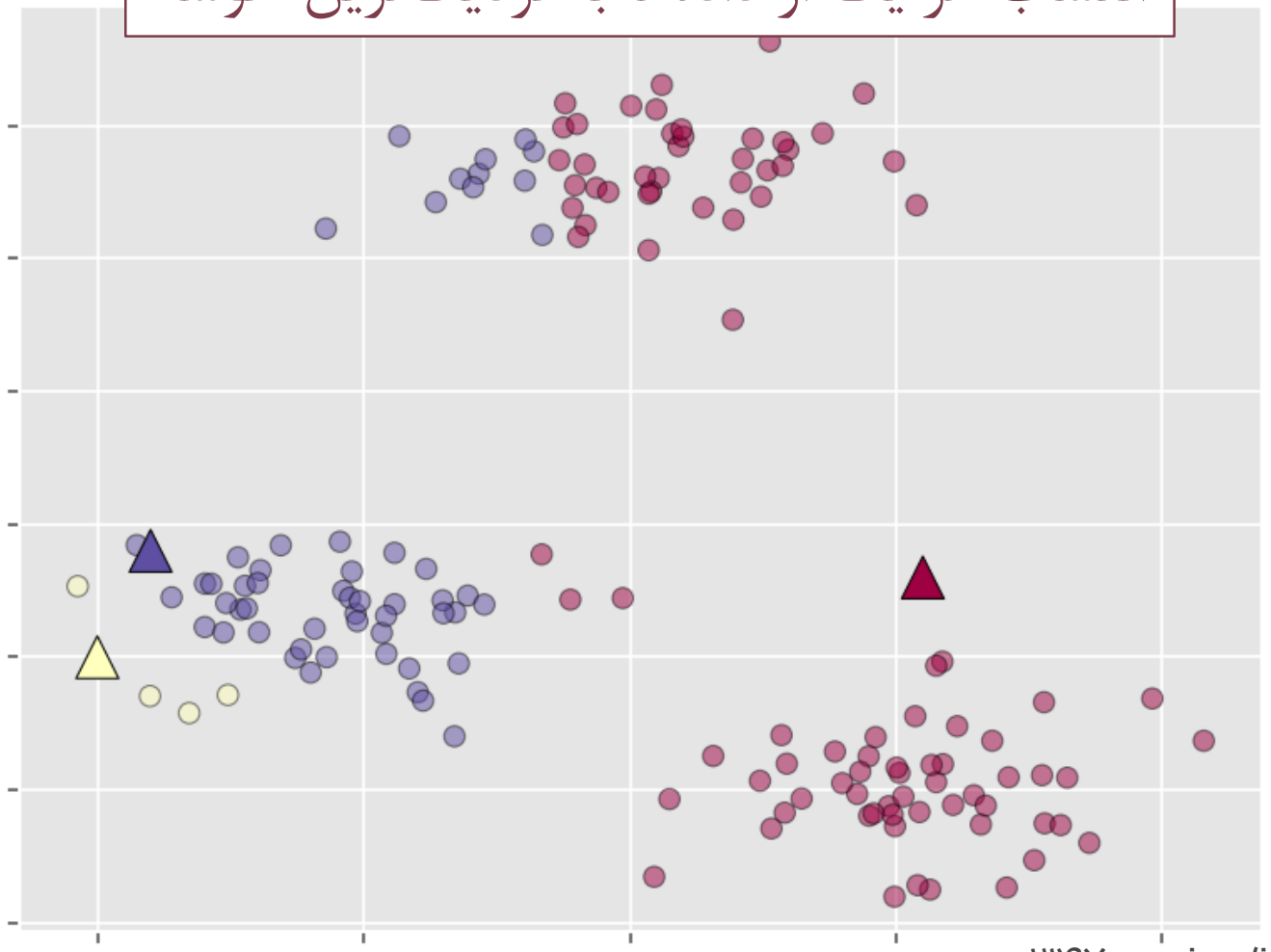
الگوریتم خوشه‌بندی K-means



- 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 خوشه خود را عوض نکند.

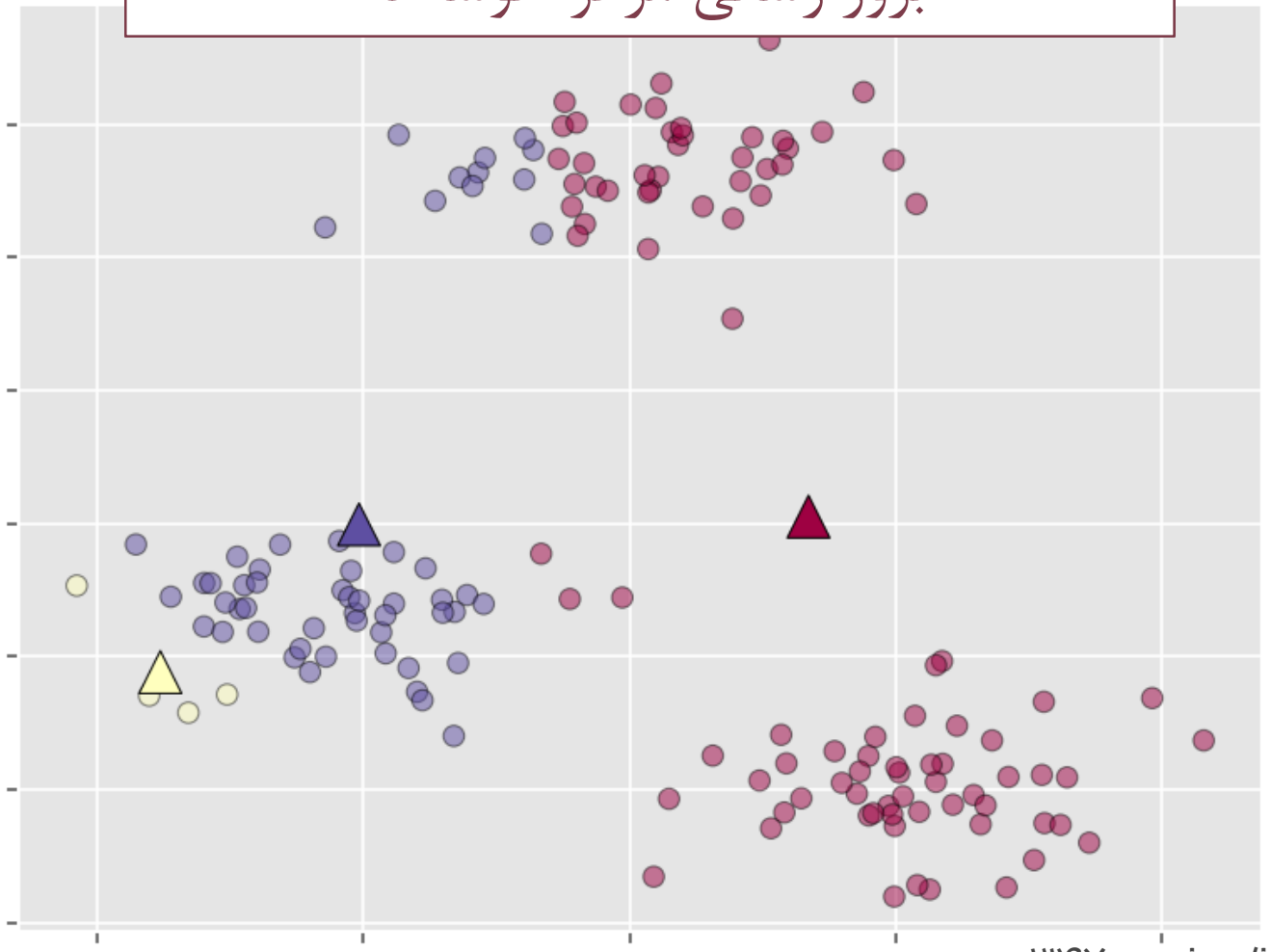
خوشه‌بندی: اجرای نمایشی

انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه



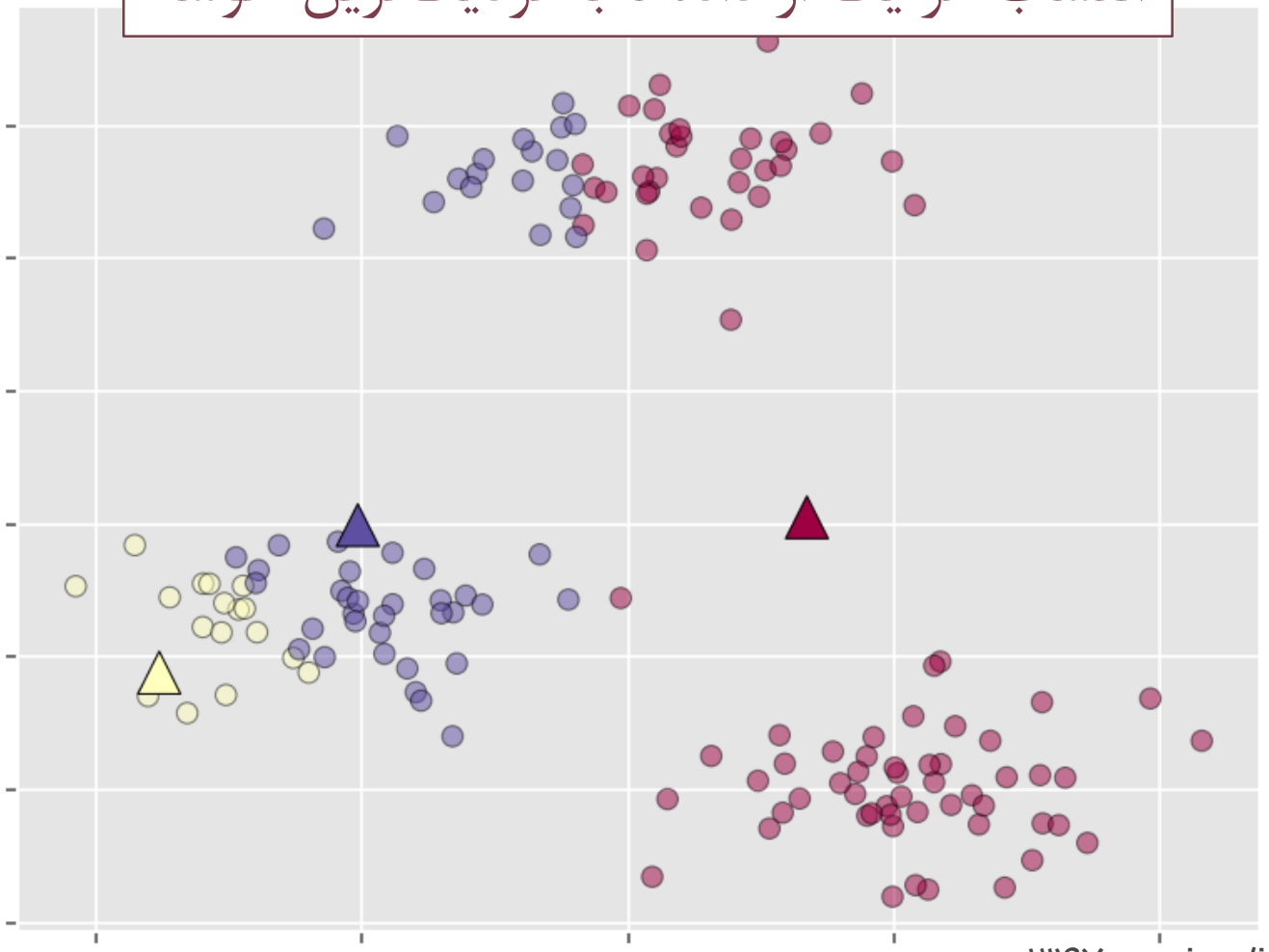
خوشه‌بندی: اجرای نمایشی

بروز رسانی مرکز خوشه‌ها



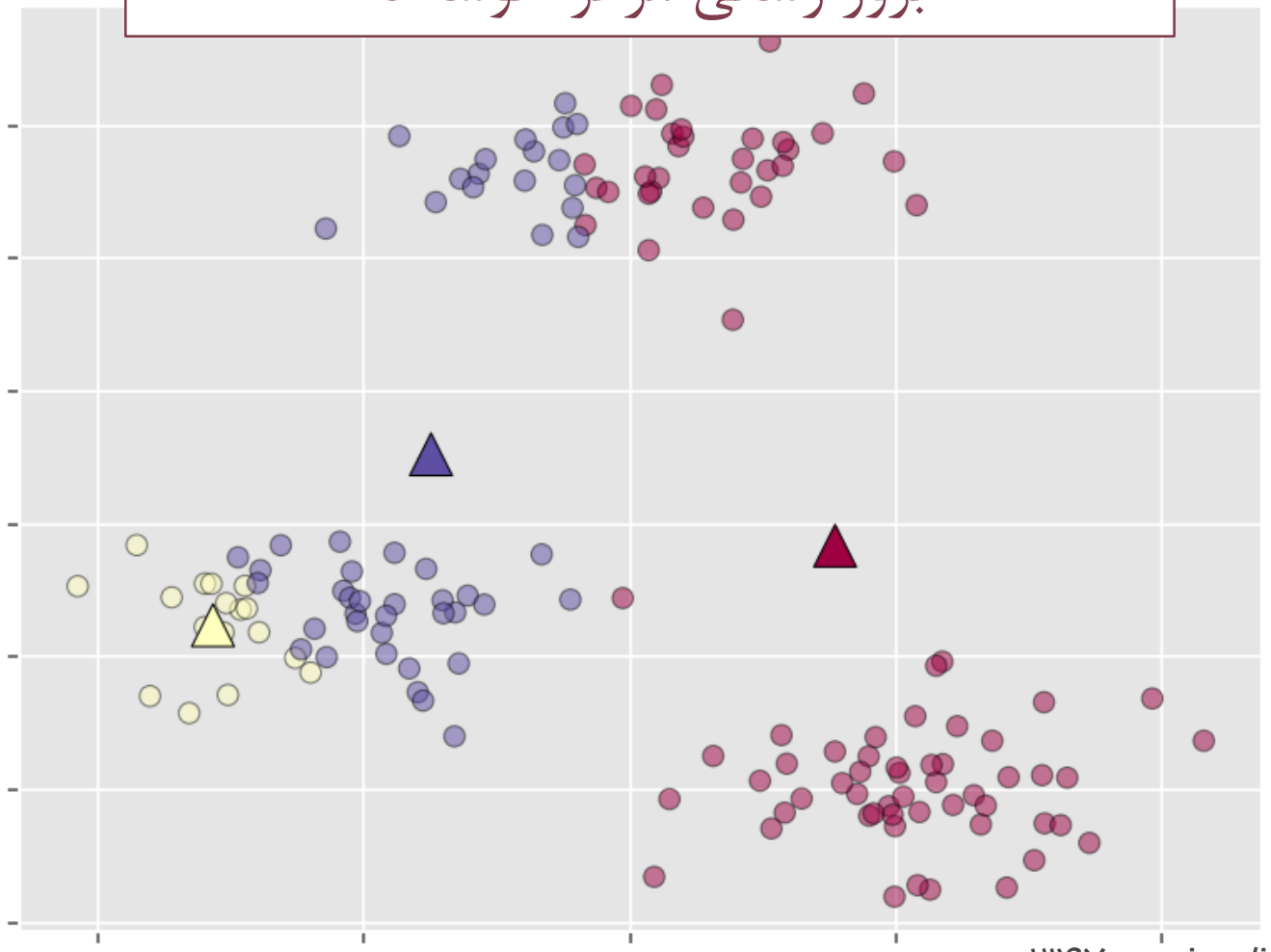
خوشه‌بندی: اجرای نمایشی

انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه



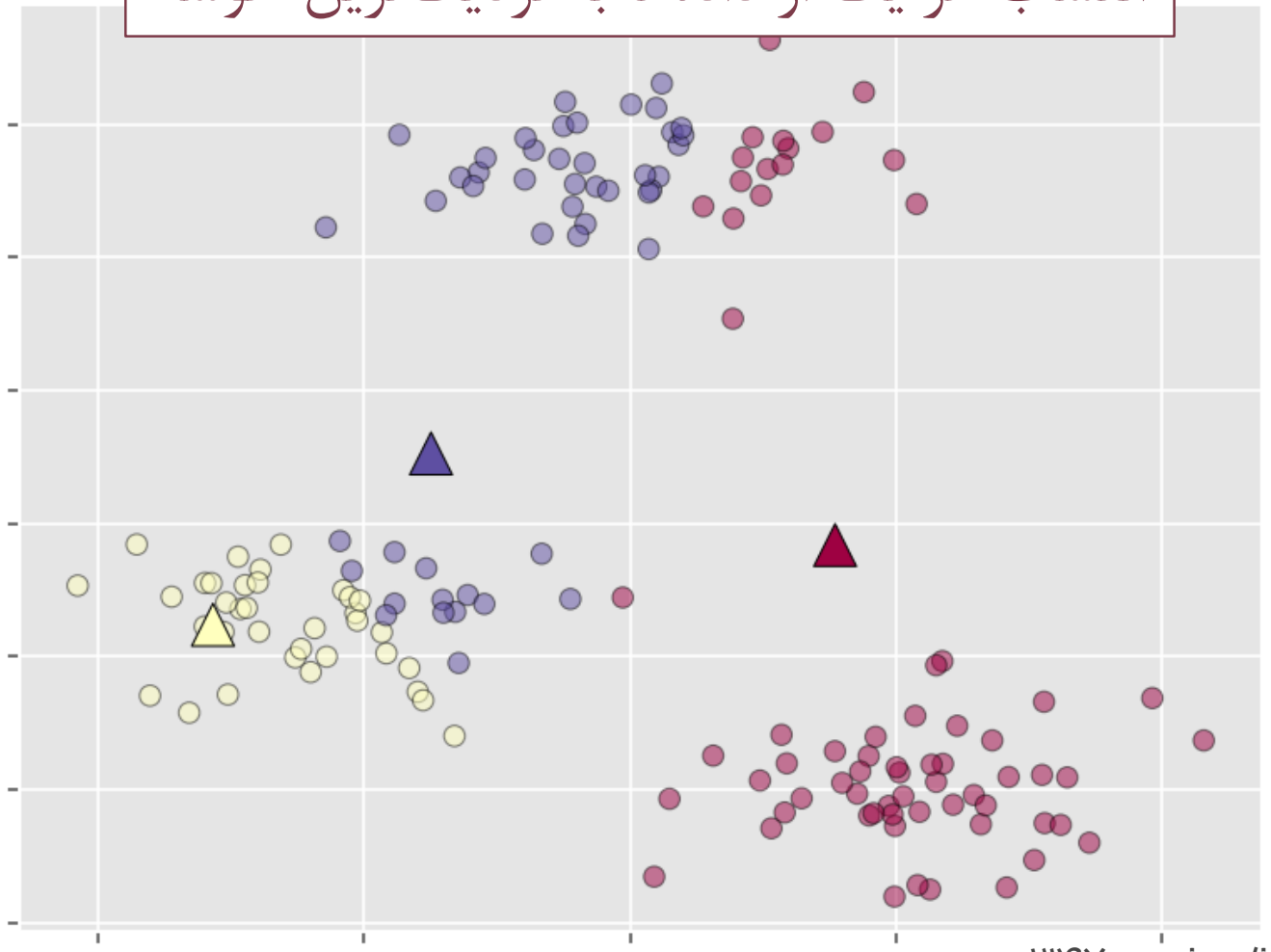
خوشه‌بندی: اجرای نمایشی

بروز رسانی مرکز خوشه‌ها



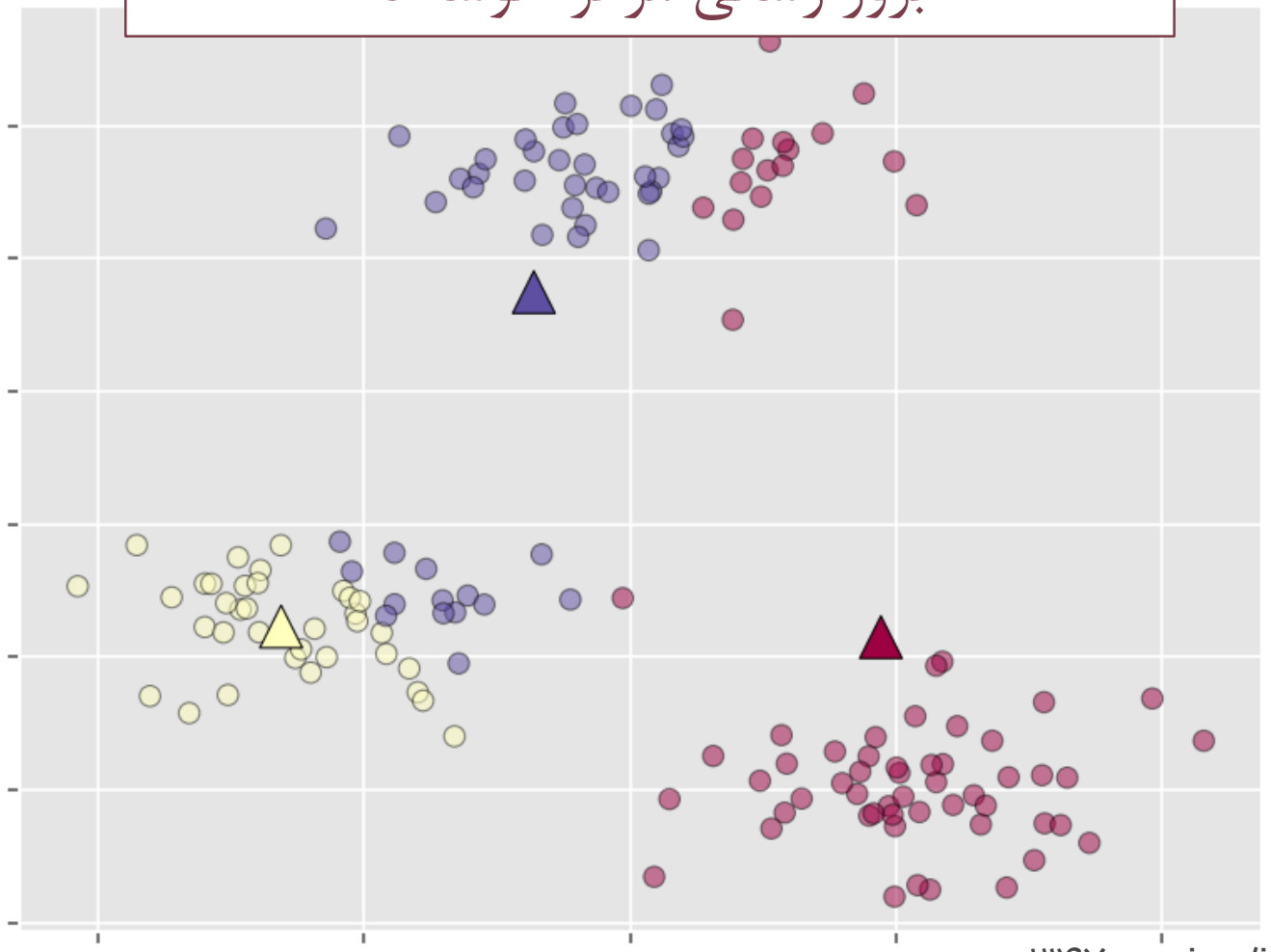
خوشه‌بندی: اجرای نمایشی

انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه



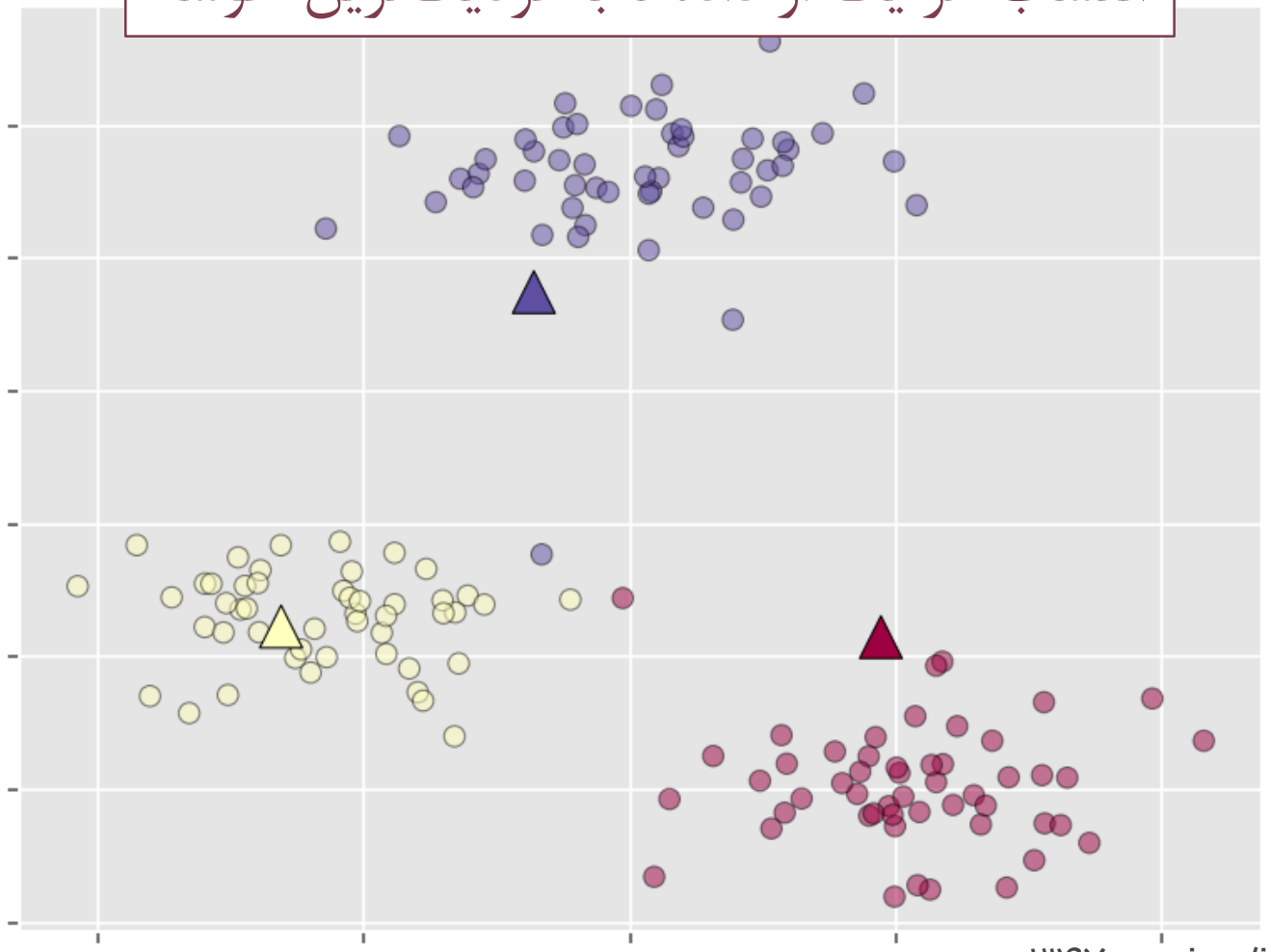
خوشه‌بندی: اجرای نمایشی

بروز رسانی مرکز خوشه‌ها



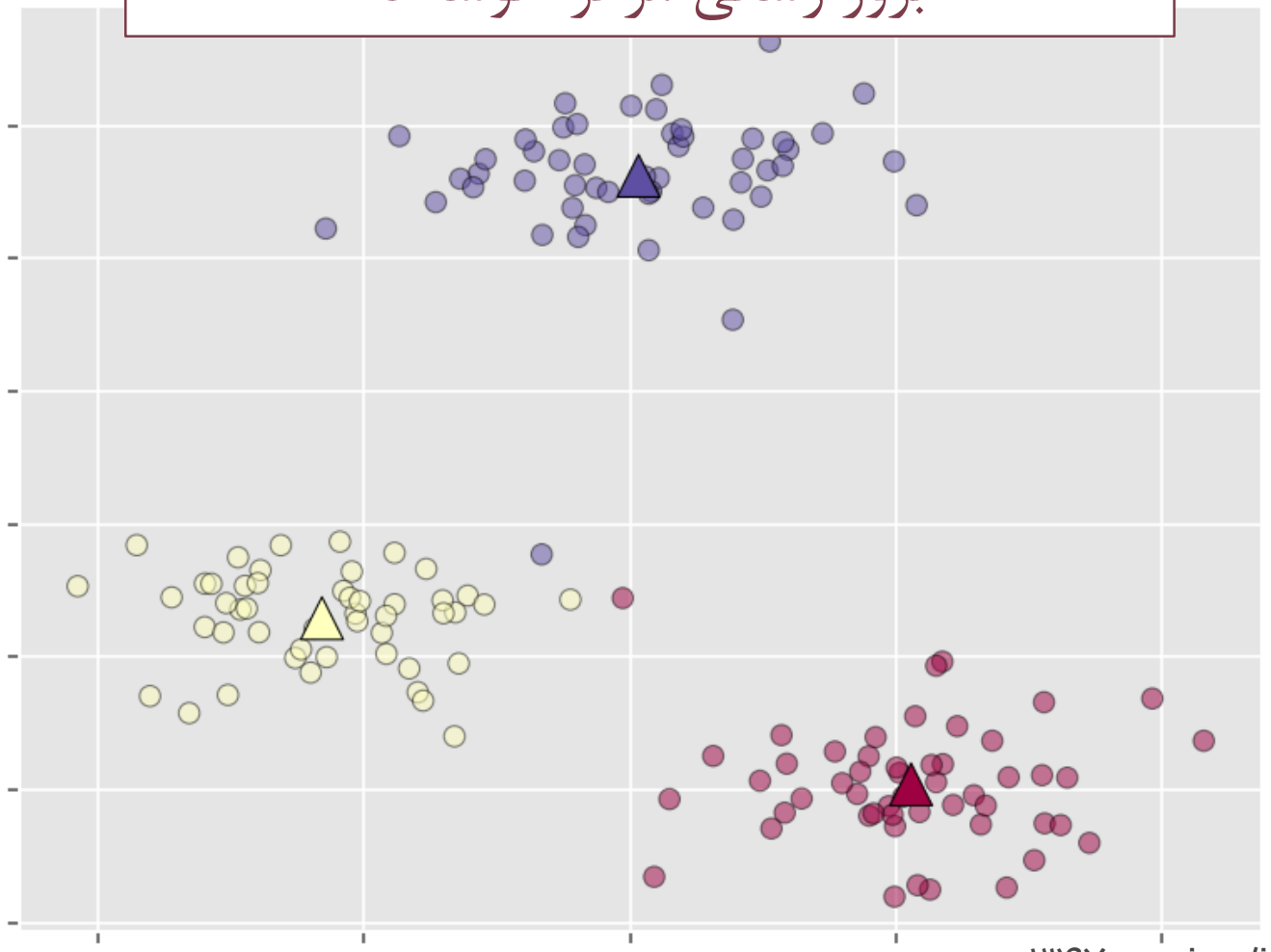
خوشه‌بندی: اجرای نمایشی

انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه



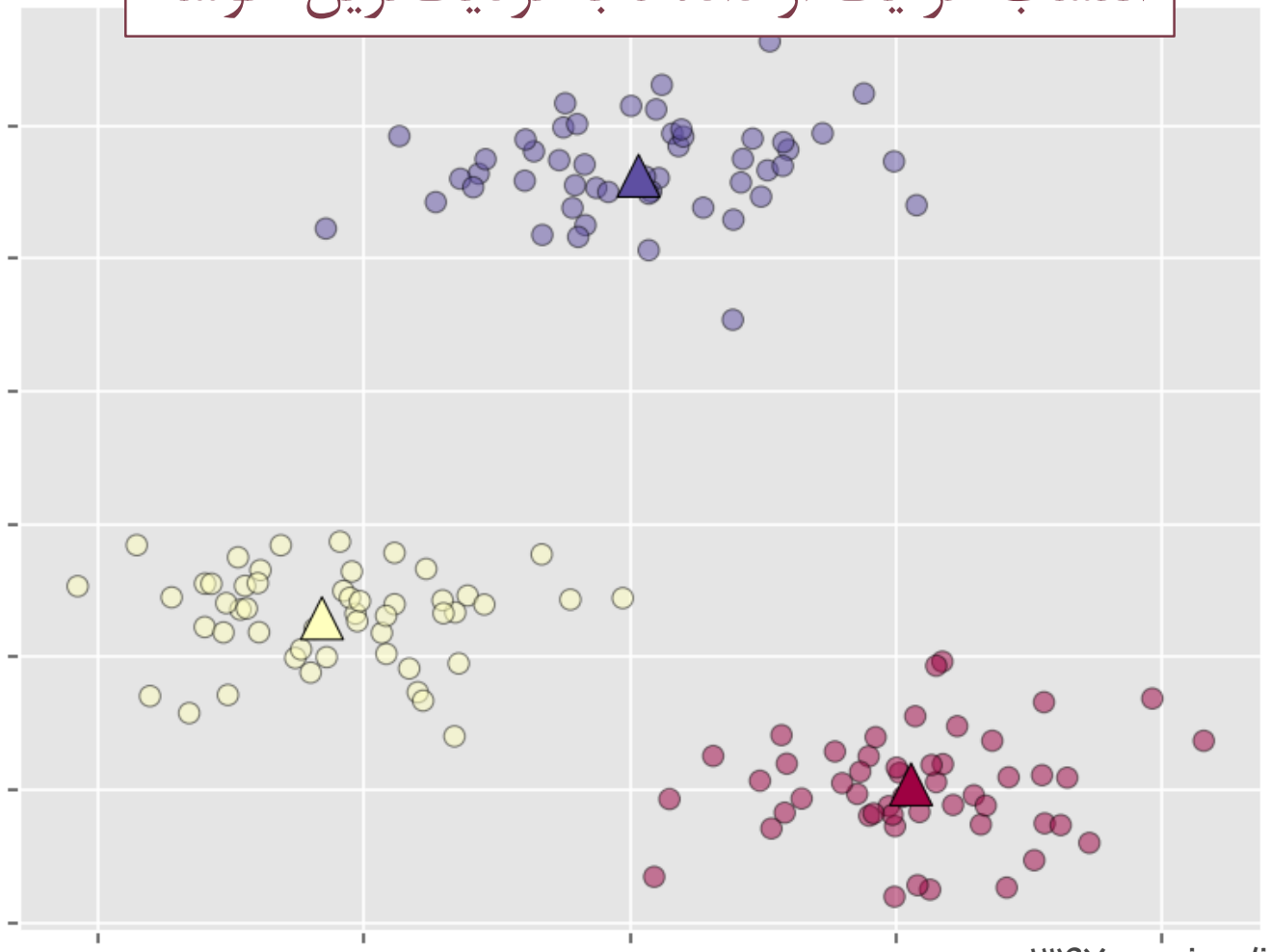
خوشه‌بندی: اجرای نمایشی

بروز رسانی مرکز خوشه‌ها



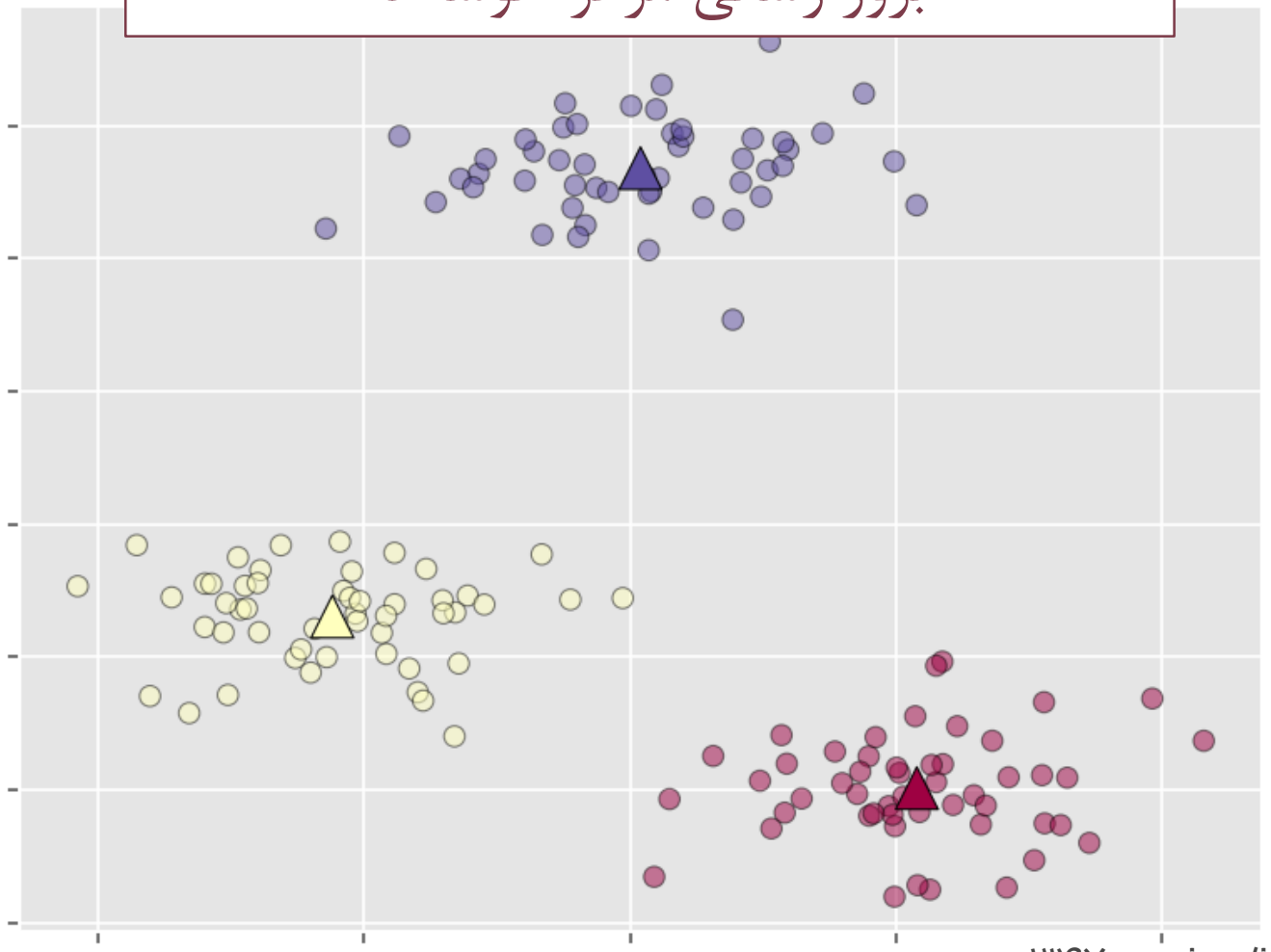
خوشه‌بندی: اجرای نمایشی

انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه



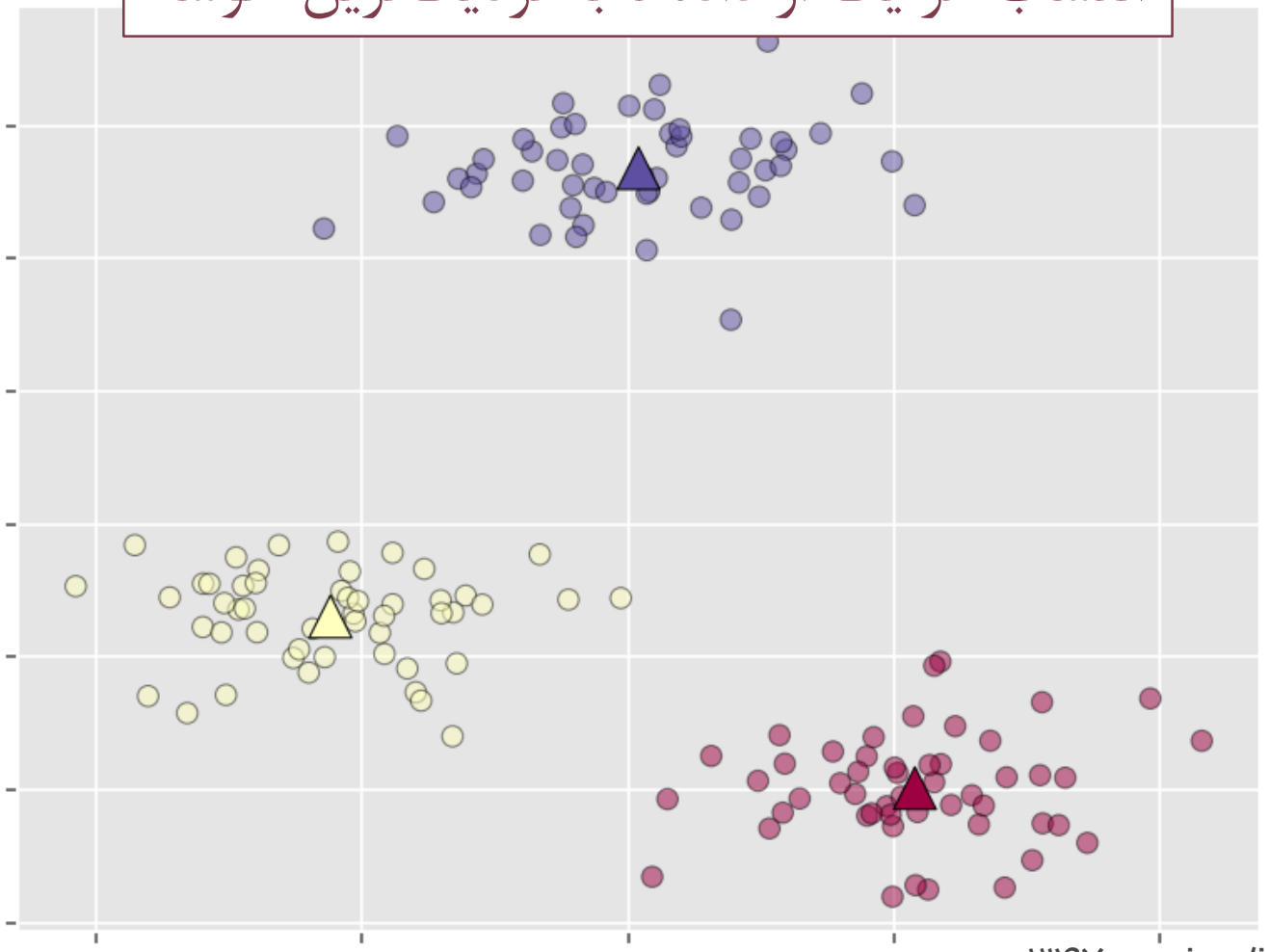
خوشه‌بندی: اجرای نمایشی

بروز رسانی مرکز خوشه‌ها



خوشه‌بندی: اجرای نمایشی

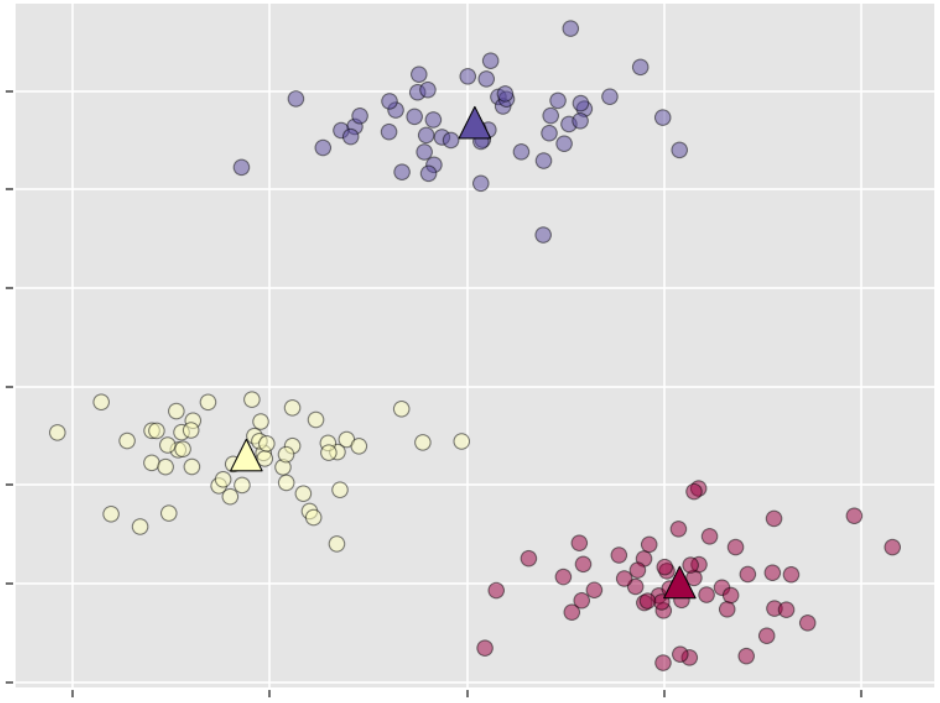
انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه



همگرایی:

خوشه هیچ یک از داده‌ها تغییر نکرد

الگوریتم K-means



- ورودی‌ها.
  - تعداد خوشه‌ها: K
  - مجموعه آموزشی:

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

- توجه. در مجموعه آموزشی، هیچ برچسبی برای داده‌ها تعیین نشده است.
- توجه. در خوشه‌بندی نیازی به افزودن ویژگی $x_0 = 1$ نیست.

الگوریتم K-means

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

repeat

{

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

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

انتساب داده‌ها به خوشه‌ها

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

μ_k = average of points assigned to cluster k

بروز رسانی مرکز خوشه‌ها

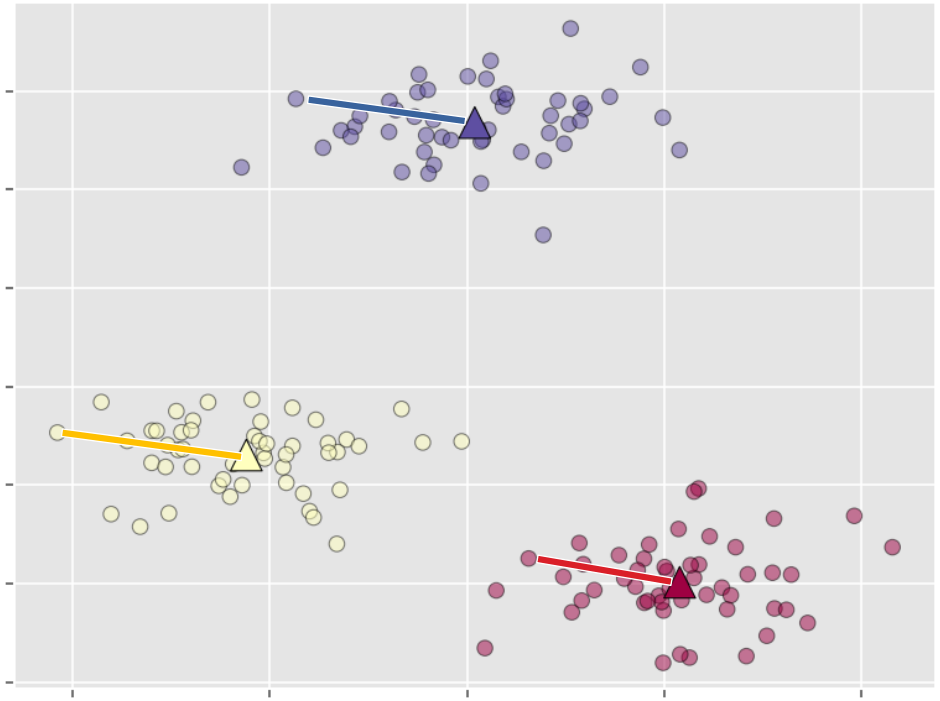
}

الگوریتم K-means


```
centroids = np.random.random((K, n))

while True:

    for i in range(m):
        c[i] = np.argmin(np.linalg.norm(X[i] - centroids, axis=1))

    for k in range(K):
        centroids[k] = np.mean(X[c == k], axis=0)
```

خوشه‌بندی: تابع هدف

تابع هدف



- نمادها.
  - μ_k : مرکز خوشه k
  - $c^{(i)}$: شماره خوشه اختصاص یافته به داده $x^{(i)}$
  - $\mu_{c^{(i)}}$: مرکز خوشه اختصاص یافته به داده $x^{(i)}$

- تابع هدف.

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

الگوریتم K-means

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

repeat

{

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

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

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

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

repeat

{

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

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

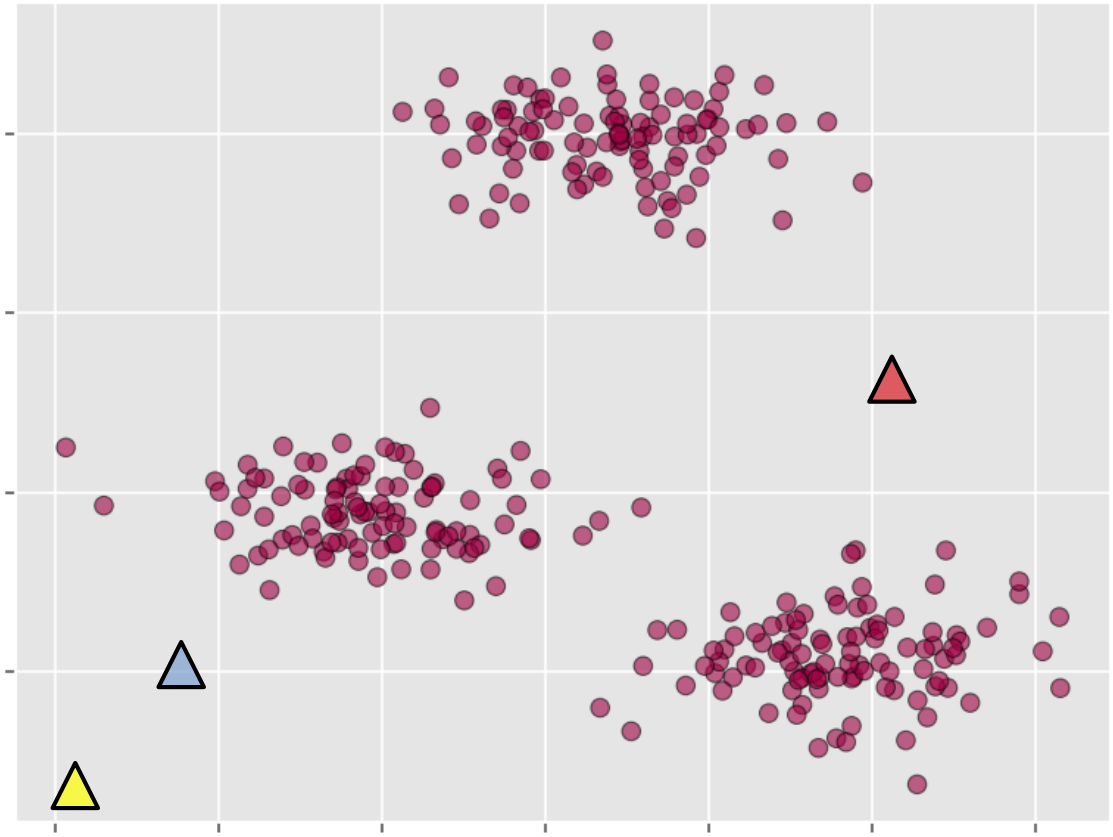
for $k = 1$ **to** K

μ_k= average of points assigned to cluster k

}

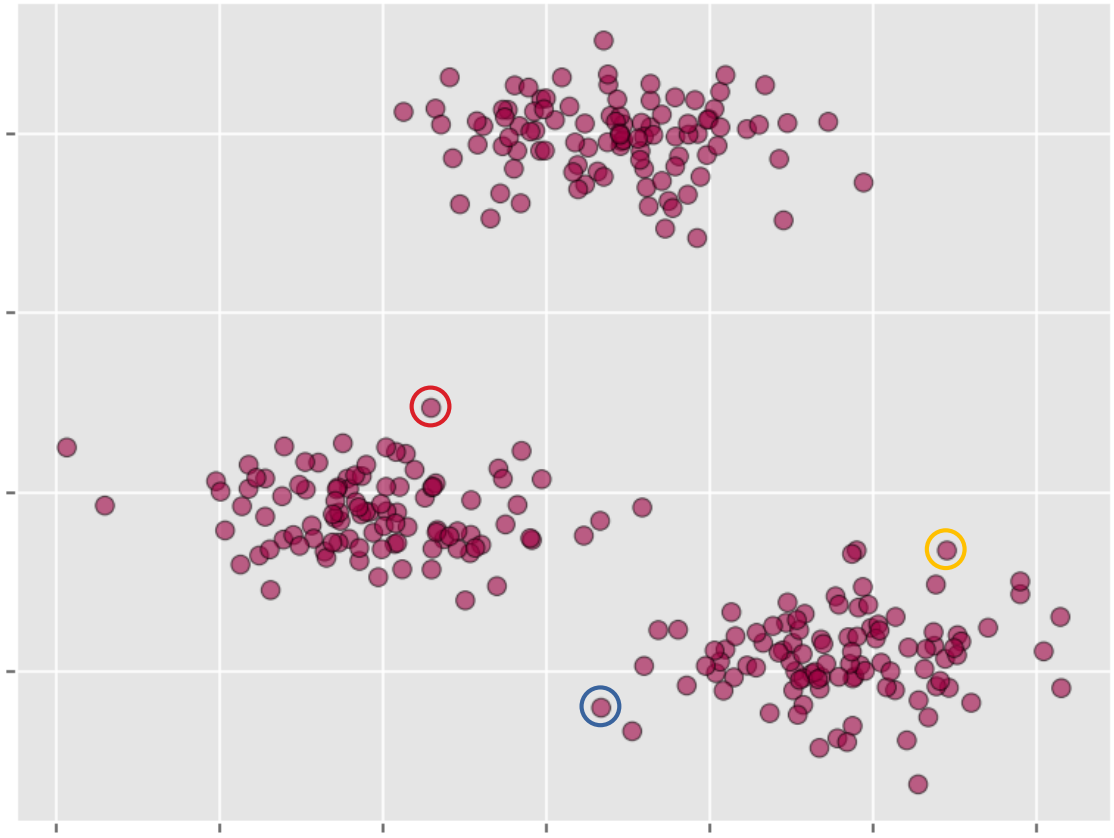
مقداردهی اولیه به مرکز خوشه‌ها

- مقداردهی اولیه. $[K \leq m]$
  - انتخاب K نقطه به صورت تصادفی
  - انتساب مراکز خوشه‌ها به K نقطه انتخاب شده



ممکن است یک مرکز به گونه‌ای انتخاب گردد که هیچ داده‌ای به آن تعلق نگیرد.

مقداردهی اولیه به مرکز خوشه‌ها: روش بهتر

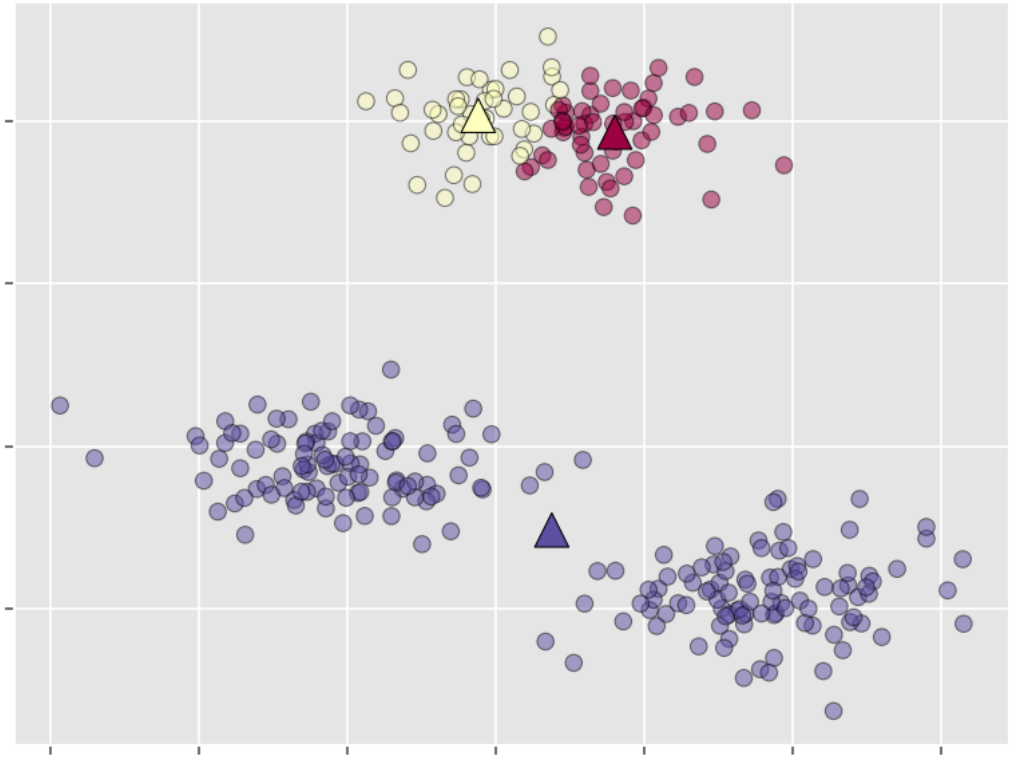


- مقداردهی اولیه. $[K \leq m]$
  - انتخاب K نمونه آموزشی به صورت تصادفی
  - انتساب مراکز خوشه‌ها به K نمونه انتخاب شده

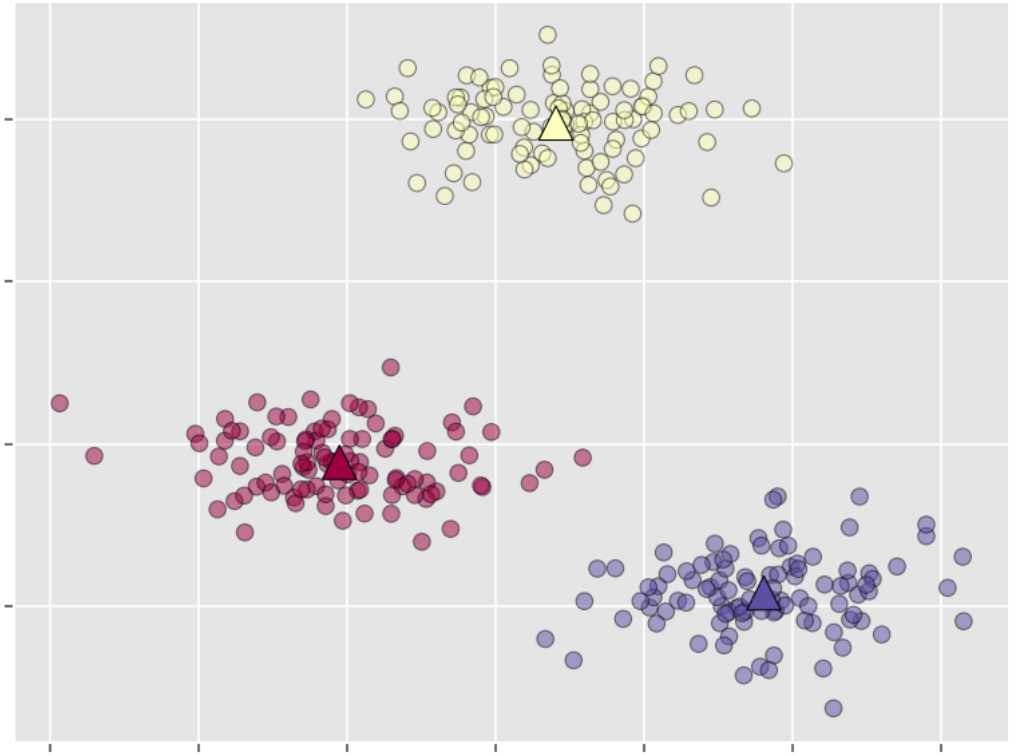
```
C = np.random.permutation(X)[:K]
```

بهینه محلی و بهینه سراسری

بهینه محلی



بهینه سراسری



اجتناب از بهینه‌های محلی

for $t = 1$ **to** MAX

{

randomly initialize cluster centroids $\mu_1, \mu_2, \dots, \mu_k$

run K-means to get $c^{(1)}, c^{(2)}, \dots c^{(m)}, \mu_1, \mu_2, \dots, \mu_k$

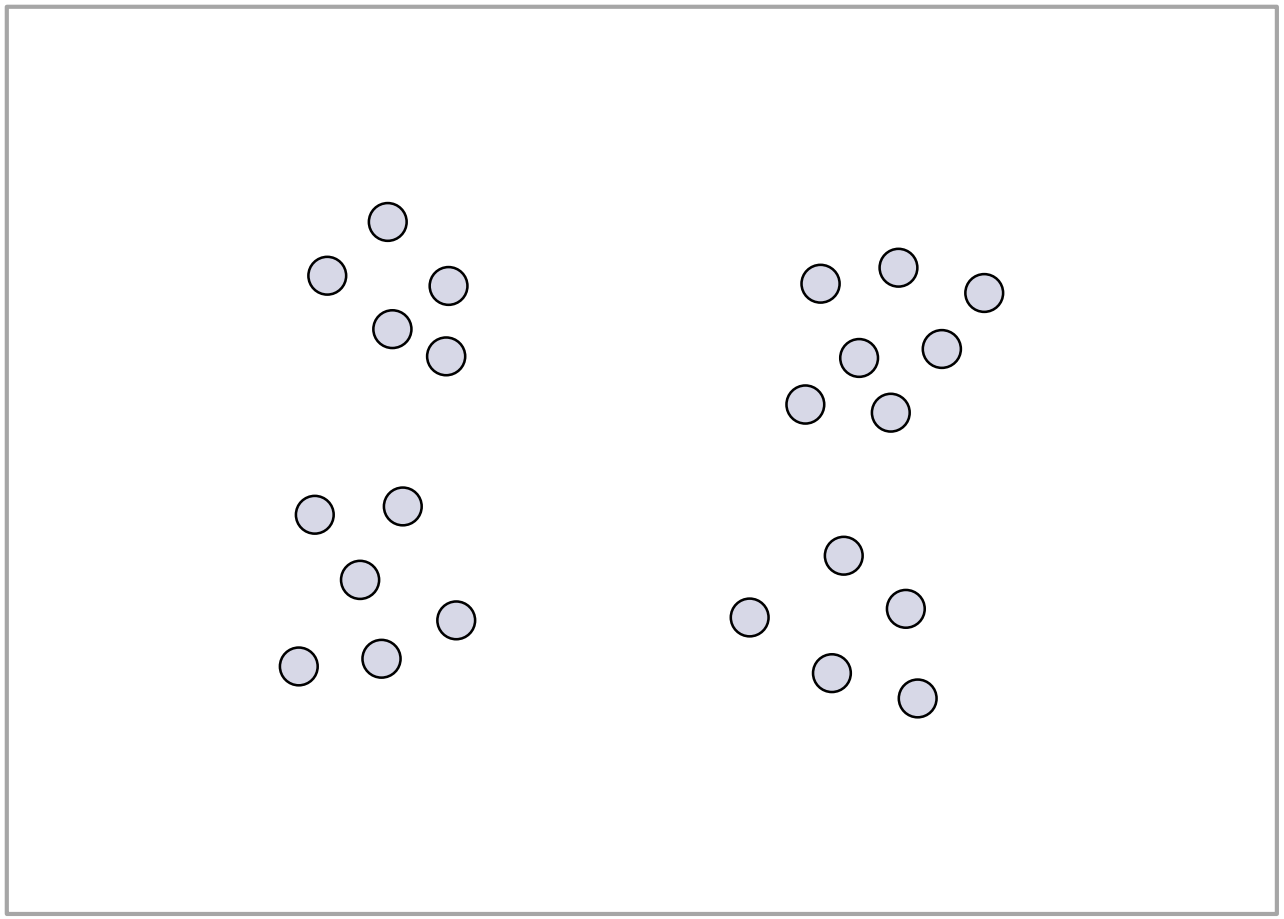
compute cost function $J(c^{(1)}, c^{(2)}, \dots c^{(m)}, \mu_1, \mu_2, \dots, \mu_k)$

}

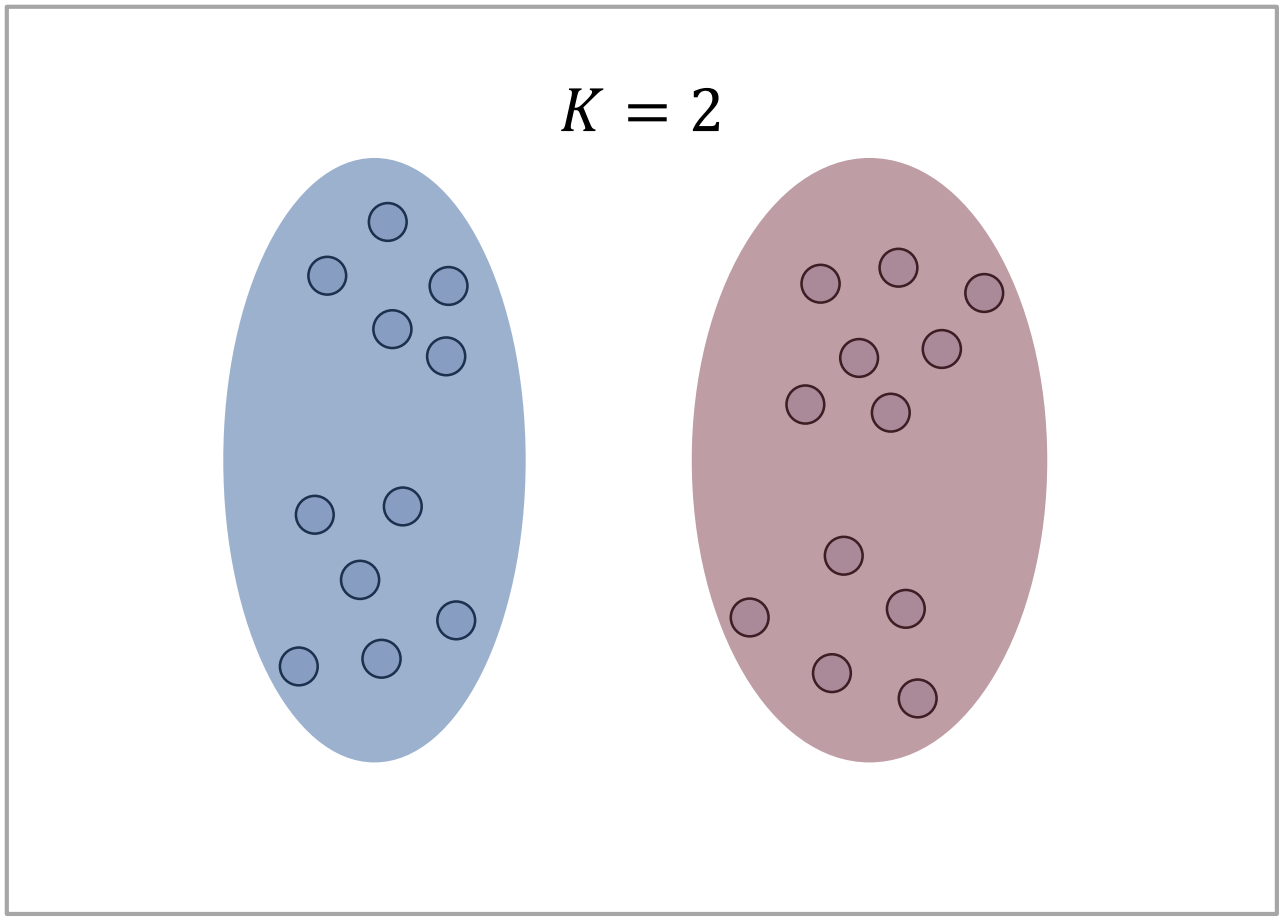
pick clustering with minimum cost

تعیین تعداد خوشه‌ها

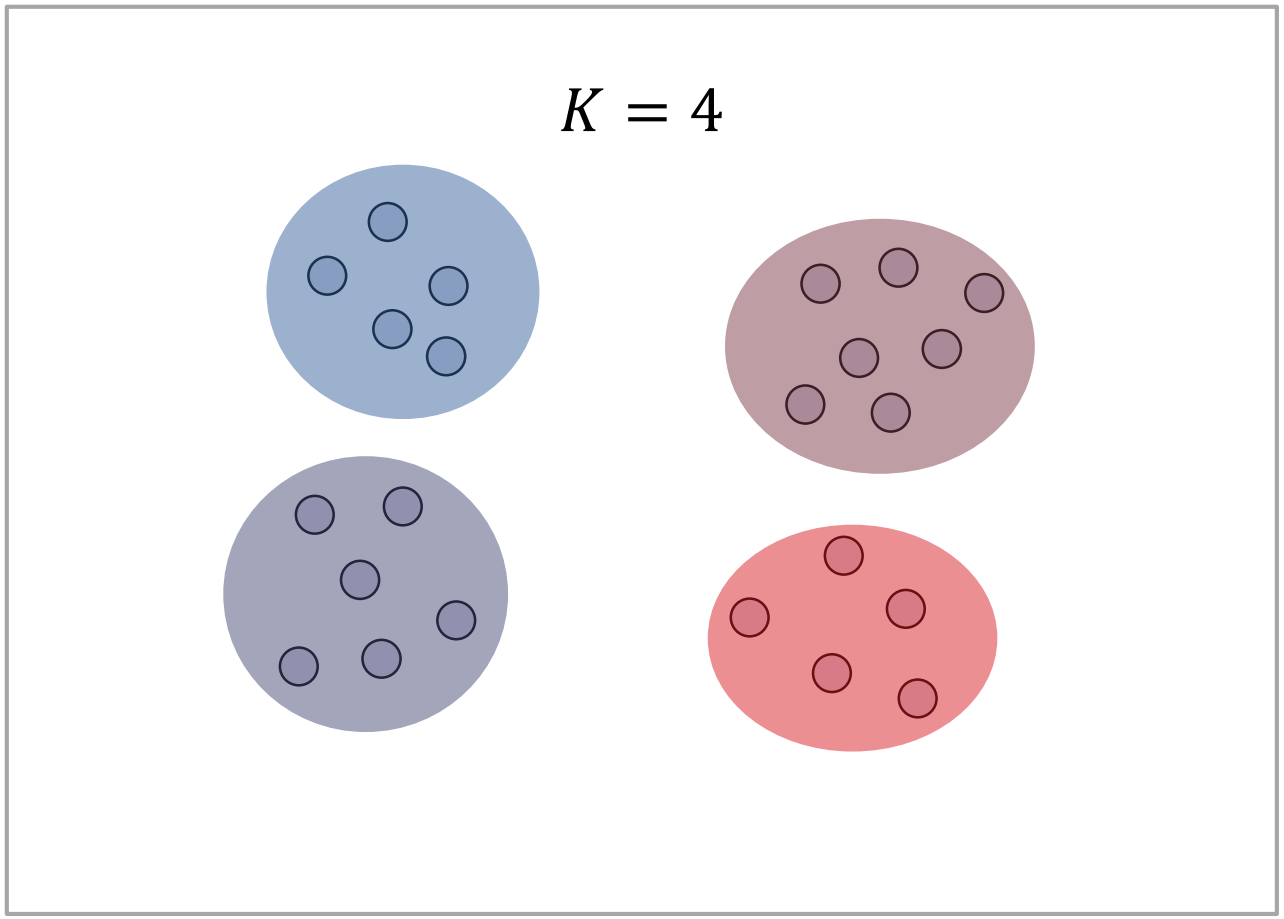
مقدار مناسب برای K کدام است؟



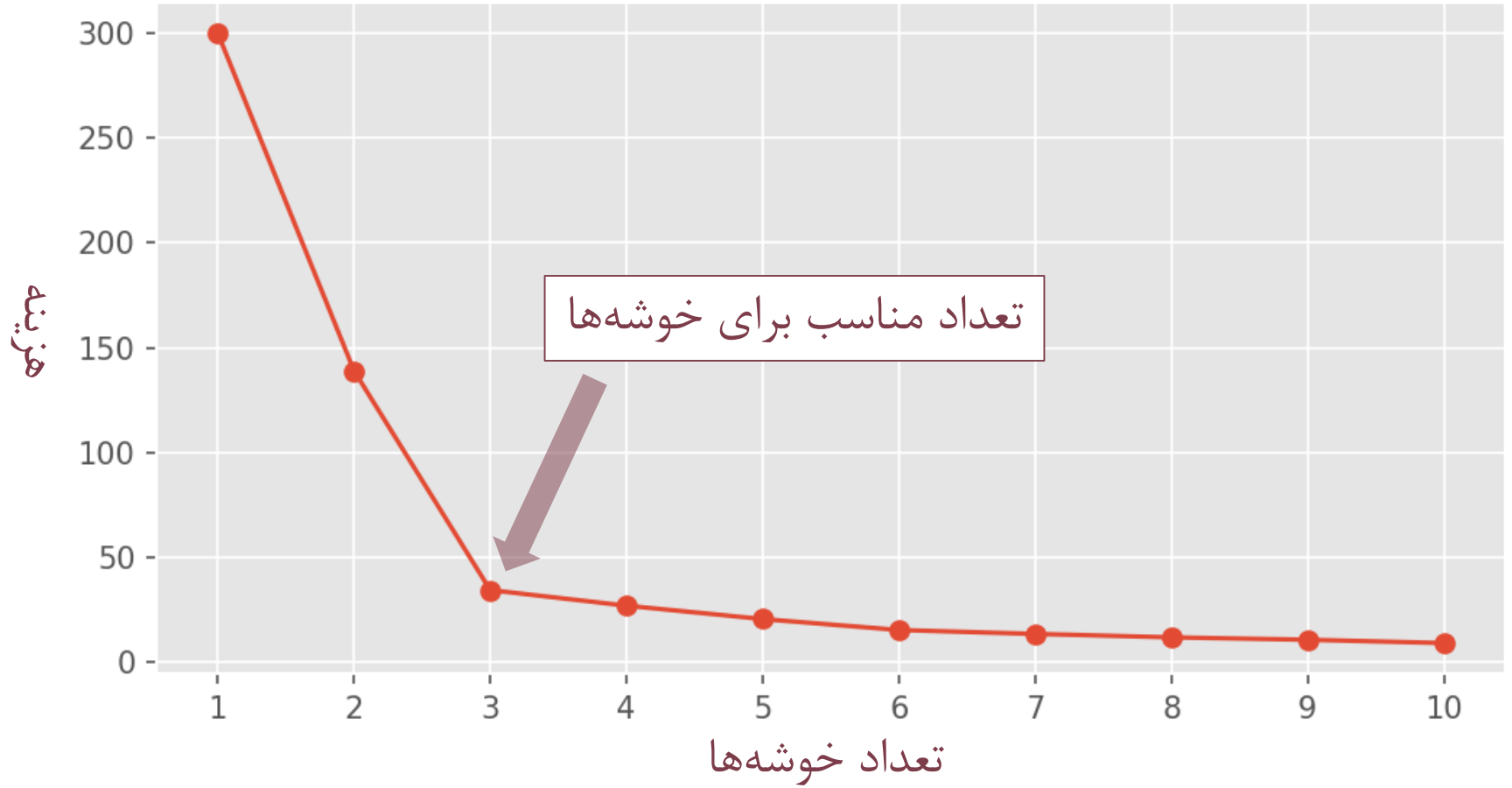
مقدار مناسب برای K کدام است؟



مقدار مناسب برای K کدام است؟

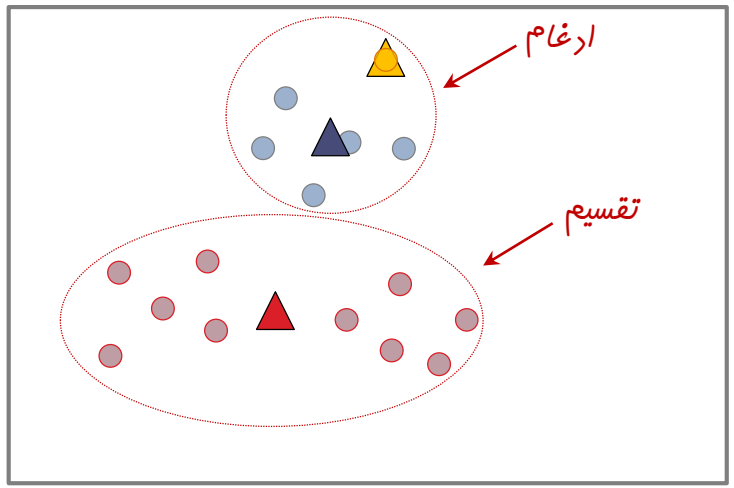


تعیین تعداد مناسب خوشه‌ها: روش آرنج



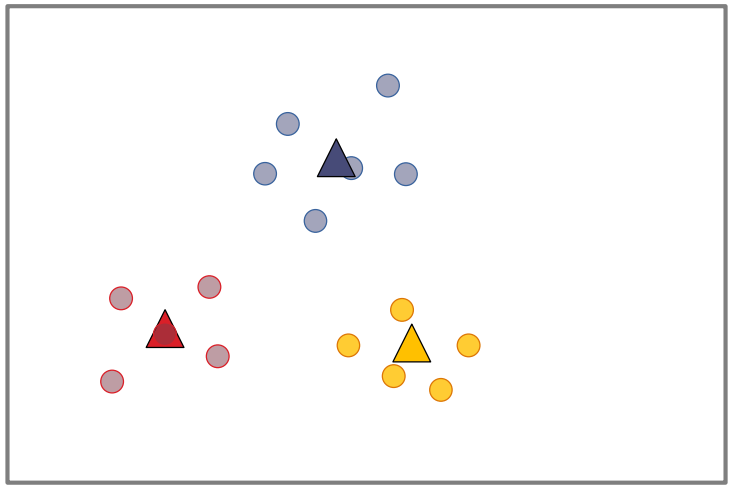
بهبود خوشه‌بندی

بهبود خوشه‌بندی با پس‌پردازش خوشه‌ها



- 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 بخشی‌ساز

- الگوریتم دو بخشی‌ساز.

  - با یک خوشه شامل تمامی داده‌ها شروع کن.

  - هر بار یک خوشه را انتخاب کن:
    - خوشه انتخاب شده را به وسیله الگوریتم K-means به دو خوشه تقسیم کن.
    - مجموع خطای خوشه‌بندی را محاسبه کن.
    - خوشه‌بندی با کمترین خطا را انتخاب کن.

  - عمل بالا را تا زمان رسیدن به تعداد خوشه‌های مورد نظر تکرار کن.

الگوریتم K-means دو بخشی‌ساز

Start with all the points in one cluster

while the number of clusters is less than K

    measure the total error

    for every cluster

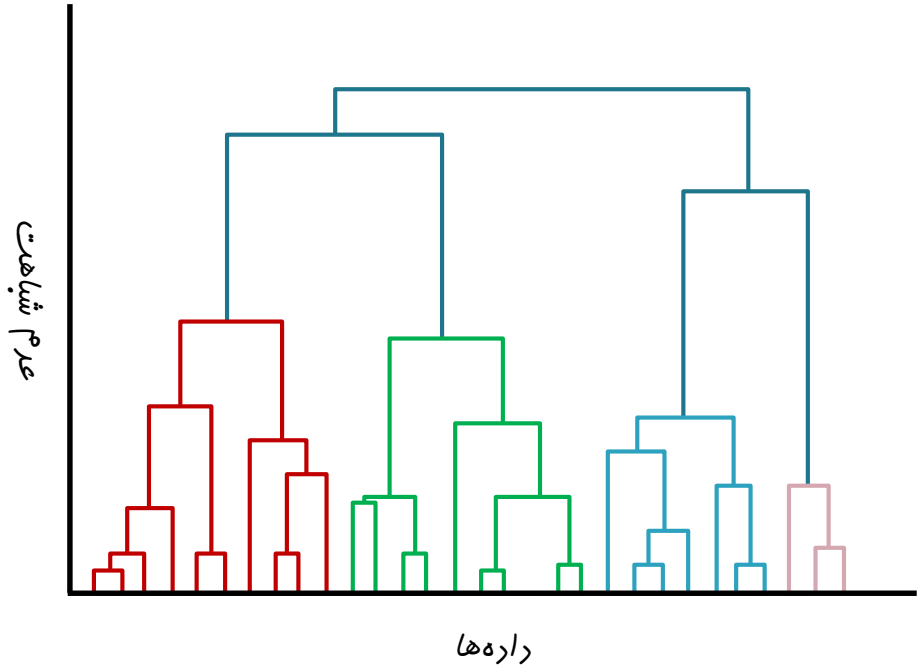
        perform K-means clustering with $k = 2$ on the given cluster

        measure the total error after splitting

    choose the cluster split that gives the lowest error

خوشه‌بندی سلسله‌مراتبی

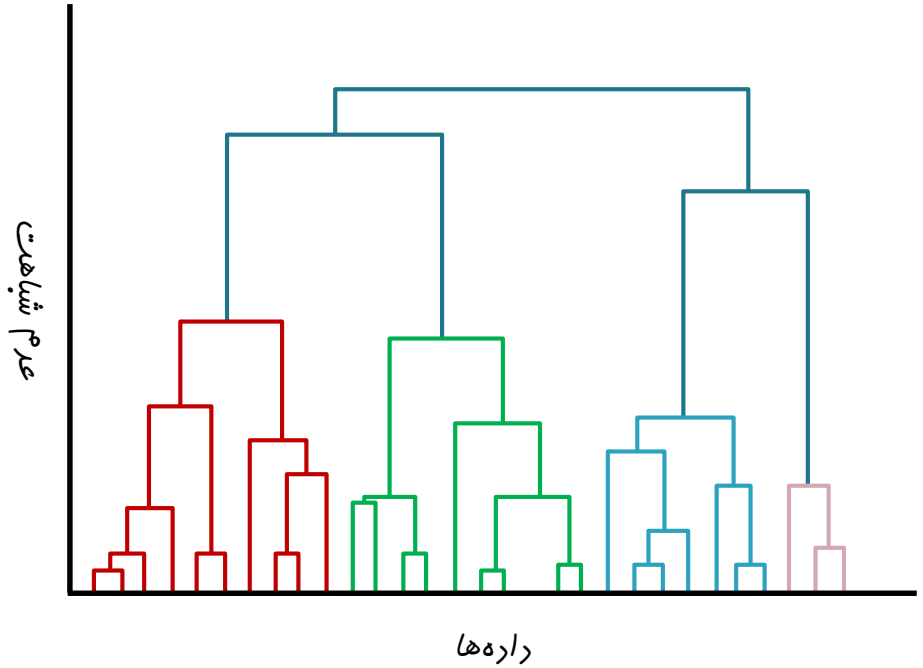
- خوشه‌بندی سلسله‌مراتبی.
  - ابتدا داده‌های بسیار شبیه را ادغام کن.
  - به تدریج با ادغام خوشه‌های کوچک‌تر، خوشه‌های بزرگ‌تری ایجاد کن.



- 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

- 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.

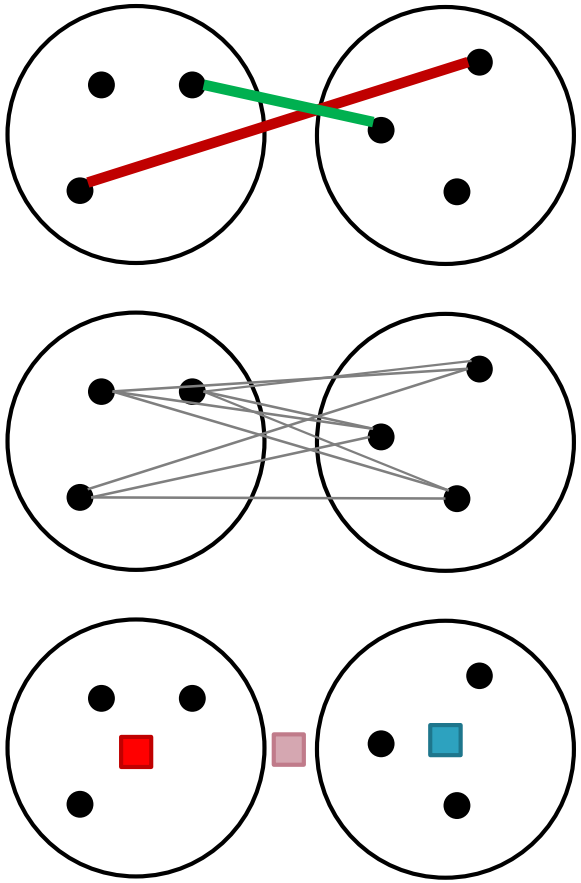
- ایجاد یک درخت نگاره شامل یک طیف گسترده از خوشه‌بندی‌ها.

خوشه‌بندی سلسله‌مراتبی

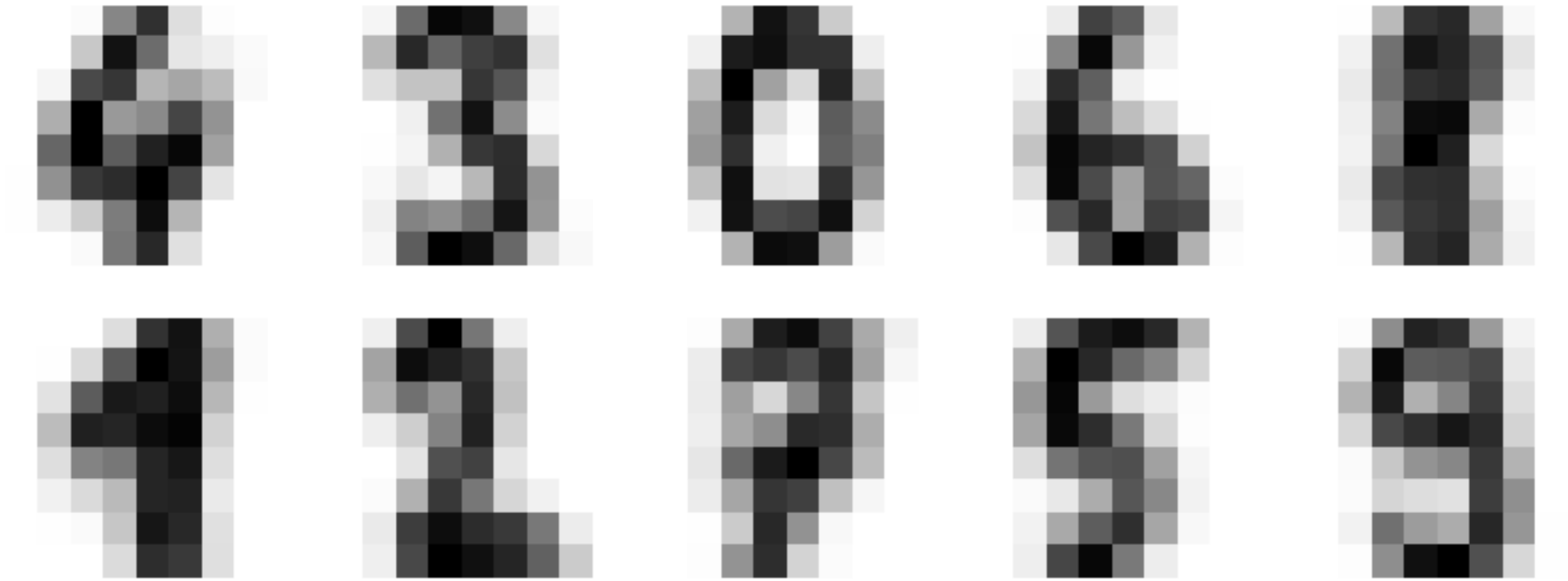
- چگونه می‌توان نزدیک‌ترین دو خوشه را تعریف نمود؟

- 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)

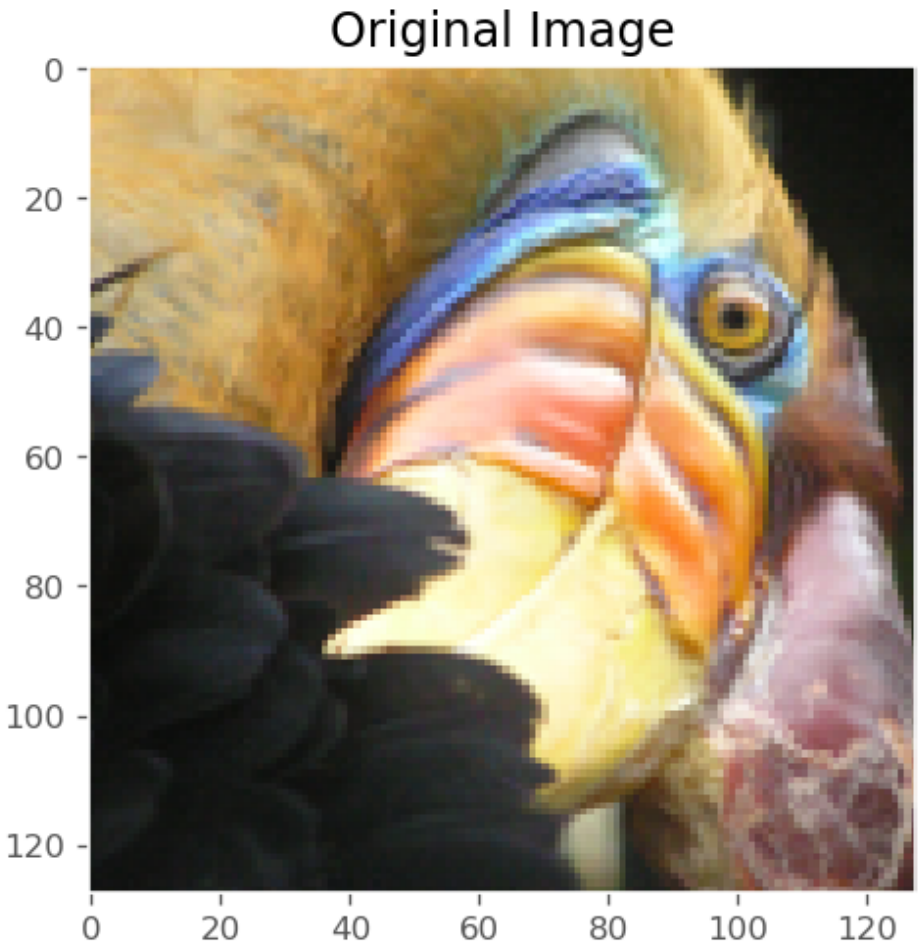
- معیارهای مختلف باعث ایجاد خوشه‌بندی‌های متفاوتی می‌شوند.

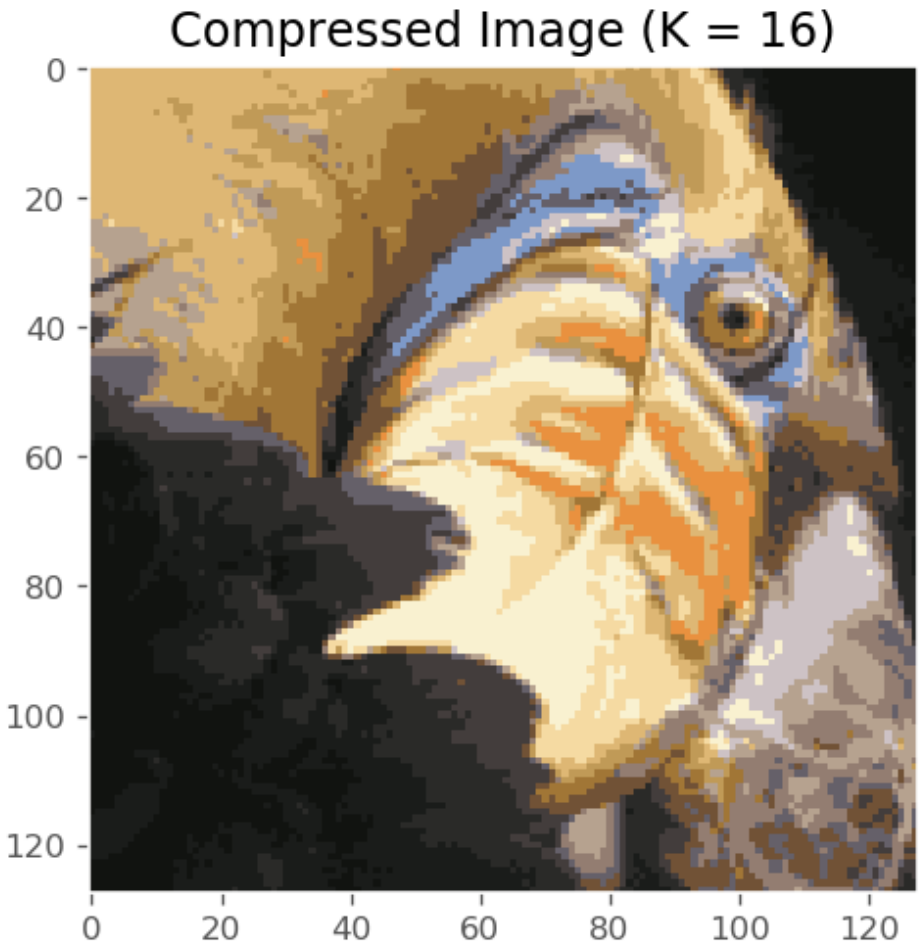


خوشه‌بندی ارقام

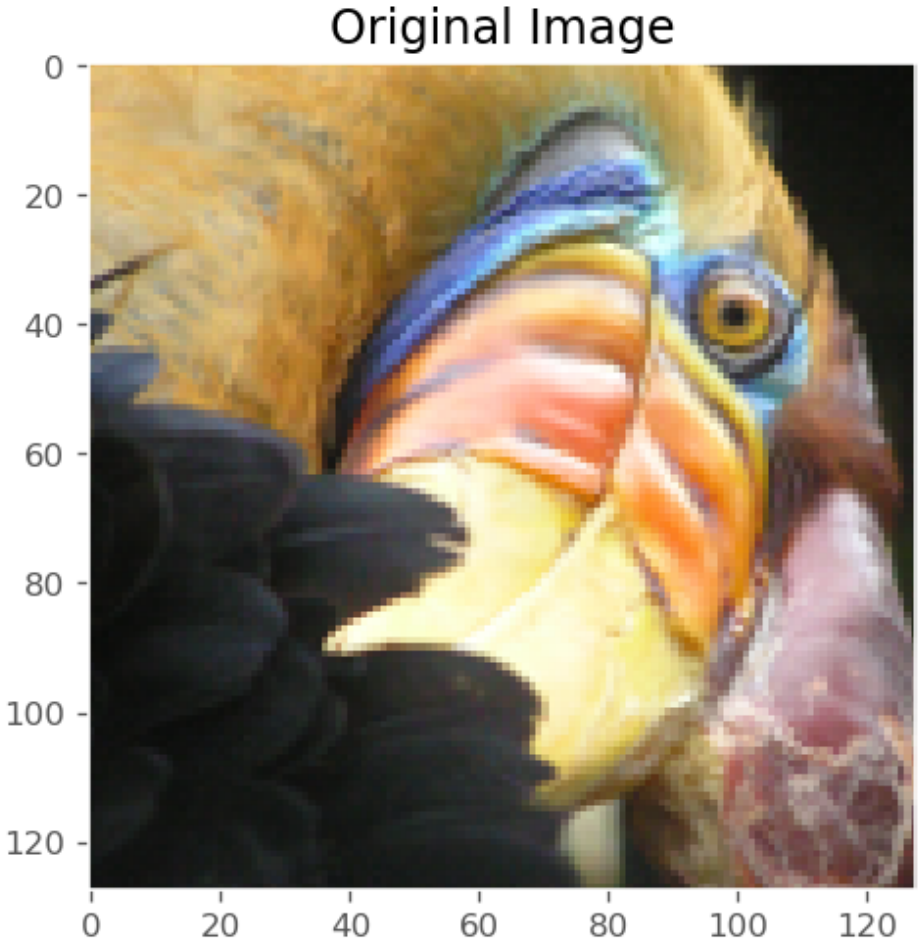


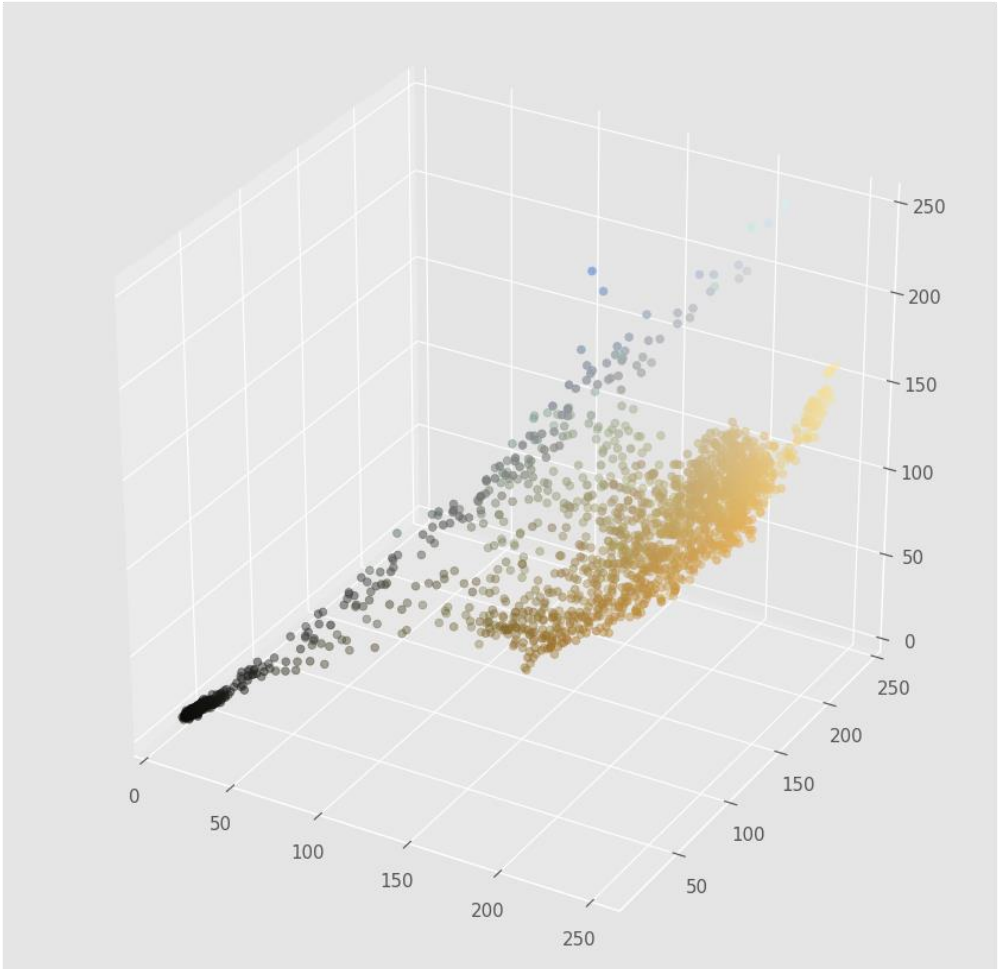
فشرده‌سازی تصویر





فشرده‌سازی تصویر





خلاصه

- 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 بهینه محلی
  - پس‌پردازش خوشه‌ها: تقسیم و ادغام خوشه‌ها
  - الگوریتم K-means دو بخشی‌ساز
    - خوشه‌بندی بهتر نسبت به الگوریتم K-means
  - الگوریتم‌های خوشه‌بندی سلسله‌مراتبی