Statistical Pattern Recognition

CLUSTERING

Expectation–maximization algorithm

شناسایی آماری الگو
بخش هشتم
(۴۱-۱۰-۷۱۱-۰۱)

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
بهار ۱۳۹۶
احمد محمودی ازناوه

http://faculties.sbu.ac.ir/~a_mahmoudi/

فهرست مطالب

- روش‌های نیمه‌پارامتری
- ترکیب چند توزیع
- K-means
- الگوریتم امید ریاضی-بیشینه کردن(EM)
- خوشه‌بندی سلسله‌مراتبی

Semiparametric Density Estimation

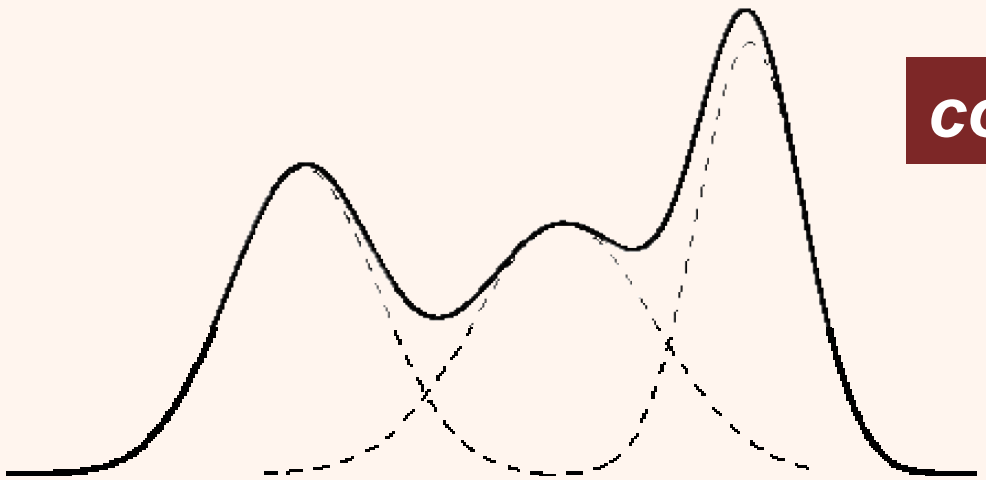
- «روش‌های پارامتری»: داده‌ها از یک توزیع تصادفی استخراج شده‌اند(مانند $p(\boldsymbol{x} \mid C_i)$).
  - مزیت این دسته از روش‌ها این است که تنها یافتن پارامترهای مدل کفایت می‌کند.
  - استفاده از روش‌های پارامتری، می‌تواند باعث ایجاد بایاس شود.
  - در برخی کاربردها، داده‌های یک دسته دارای یک توزیع یکسان نیستند، مانند دست‌نوشته‌های مختلف یا تلفظ‌های مختلف
- «روش‌های نیمه‌پارامتری»: در این حالت برای هر دسته، خوشه‌ها(گروه‌ها)ی مختلفی در نظر گرفته می‌شود که هر کدام از یک توزیع پیروی می‌کنند.
- «روش‌های ناپارامتری»: هیچ‌گونه مدلی در نظر گرفته نمی‌شود، داده‌ها خود را توصیف می‌کنند.

Mixture Densities

component densities **mixture proportions**

$$p(\mathbf{x}) = \sum_{i=1}^{k} \overbrace{p(\mathbf{x} \mid G_i)}^{\text{component densities}} \overbrace{P(G_i)}^{\text{mixture proportions}}$$

components/groups/clusters



در صورتی که خوشه‌ها دارای توزیع گاوسی باشند:

$$p(x|G_i) \sim N(\mu_i, \Sigma_i)$$

با در اختیار داشتن مجموعه‌ی آموزشی $X=\{x^t\}_t$، پارامترهای که در طی فرآیند آموزش تخمین زده می‌شوند:

$$\Phi = \{P(G_i), \mu_i, \Sigma_i\}^k_{i=1}$$

مفهوم «دسته» در مقایسه با «خوشه»

Classes vs. Clusters

Classfiction

- Supervised: $X = \{\boldsymbol{x}^t, \boldsymbol{r}^t\}_t$
- Classes C_i $i=1,\ldots,K$

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x} \mid C_i) P(C_i)$$

where $p(\boldsymbol{x} \mid C_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

- $\Phi = \{P(C_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{K}$

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

- Unsupervised: $X = \{\boldsymbol{x}^t\}_t$
- Clusters G_i $i=1,\ldots,k$

$$p(\mathbf{x}) = \sum_{i=1}^{k} p(\mathbf{x} \mid G_i) P(G_i)$$

where $p(\boldsymbol{x} \mid G_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

- $\Phi = \{P(G_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{k}$

Labels r^t_i ?

Clustering

k-Means Clustering

- هدف یافتن گروه‌های مشابه از بین داده‌های برچسب‌نخورده است.
  - یافتن k «بردار مرجع» (reference vector) است که به بهترین نحو داده‌ها را نمایش دهند.

prototypes /codebook vectors /codewords

Reference vectors, m_j, $j = 1,\ldots,k$

- بعد از مشخص شدن بردارهای مرجع، نمونه‌ها در خوشه‌ی نزدیک‌ترین بردار مرجع قرار می‌گیرند:

$$\left\|\mathbf{x}^t - \mathbf{m}_i\right\| = \min_j \left\|\mathbf{x}^t - \mathbf{m}_j\right\|$$

- بدین‌ترتیب می‌توان به جای داده‌های از بردار مرجع متناظر آن استفاده کرد.

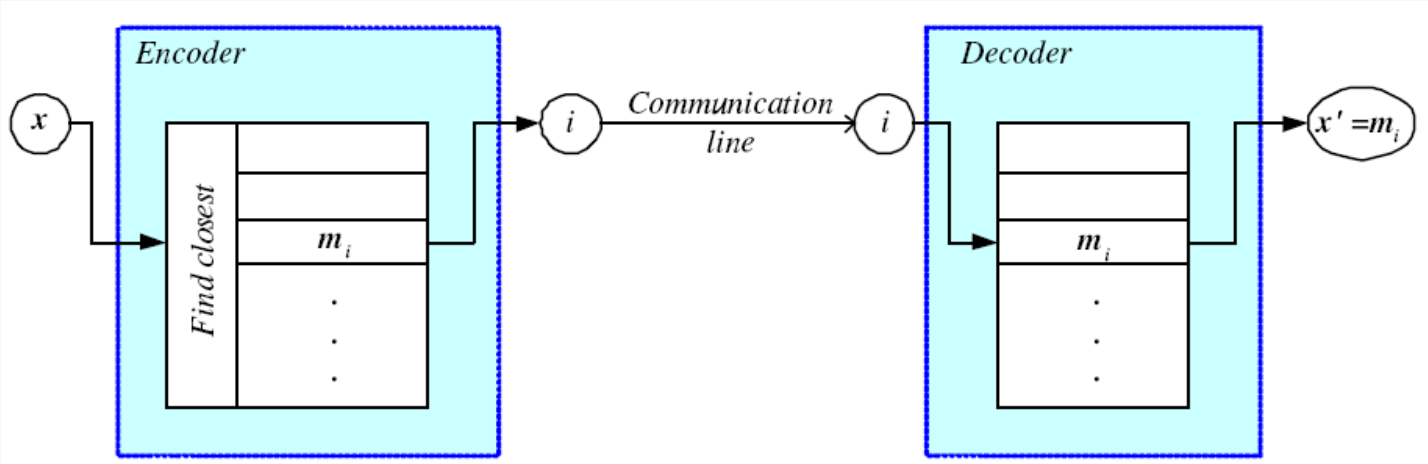
k-Means Clustering

- در این صورت «خطای بازسازی» به صورت زیر محاسبه می‌شود:

$$E\left(\{\mathbf{m}_i\}_{i=1}^{k} \middle| \mathcal{X}\right) = \sum_t \sum_i b_i^t \left\|\mathbf{x}^t - \mathbf{m}_i\right\|$$

Reconstruction error

$$b_i^t = \begin{cases} 1 & \text{if } \left\|\mathbf{x}^t - \mathbf{m}_i\right\| = \min_j \left\|\mathbf{x}^t - \mathbf{m}_j\right\| \\ 0 & \text{otherwise} \end{cases}$$



Encoding/Decoding

k-Means Clustering

- بهترین بردارهای مرجع، موجب می‌شوند تا خطای بازسازی مینیمم شود.

$$E\left(\{\mathbf{m}_i\}_{i=1}^{k} \middle| \mathcal{X}\right) = \sum_t \sum_i b_i^t \left\| \mathbf{x}^t - \mathbf{m}_i \right\|$$

- این رابطه افزون بر m_i به برچسب‌ها b_i^t هم بستگی دارد، از این رو نمی‌توان برای آن راه حل تحلیلی یافت.

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$

Repeat

For all $\boldsymbol{x}^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

For all $\boldsymbol{m}_i, i = 1, \ldots, k$

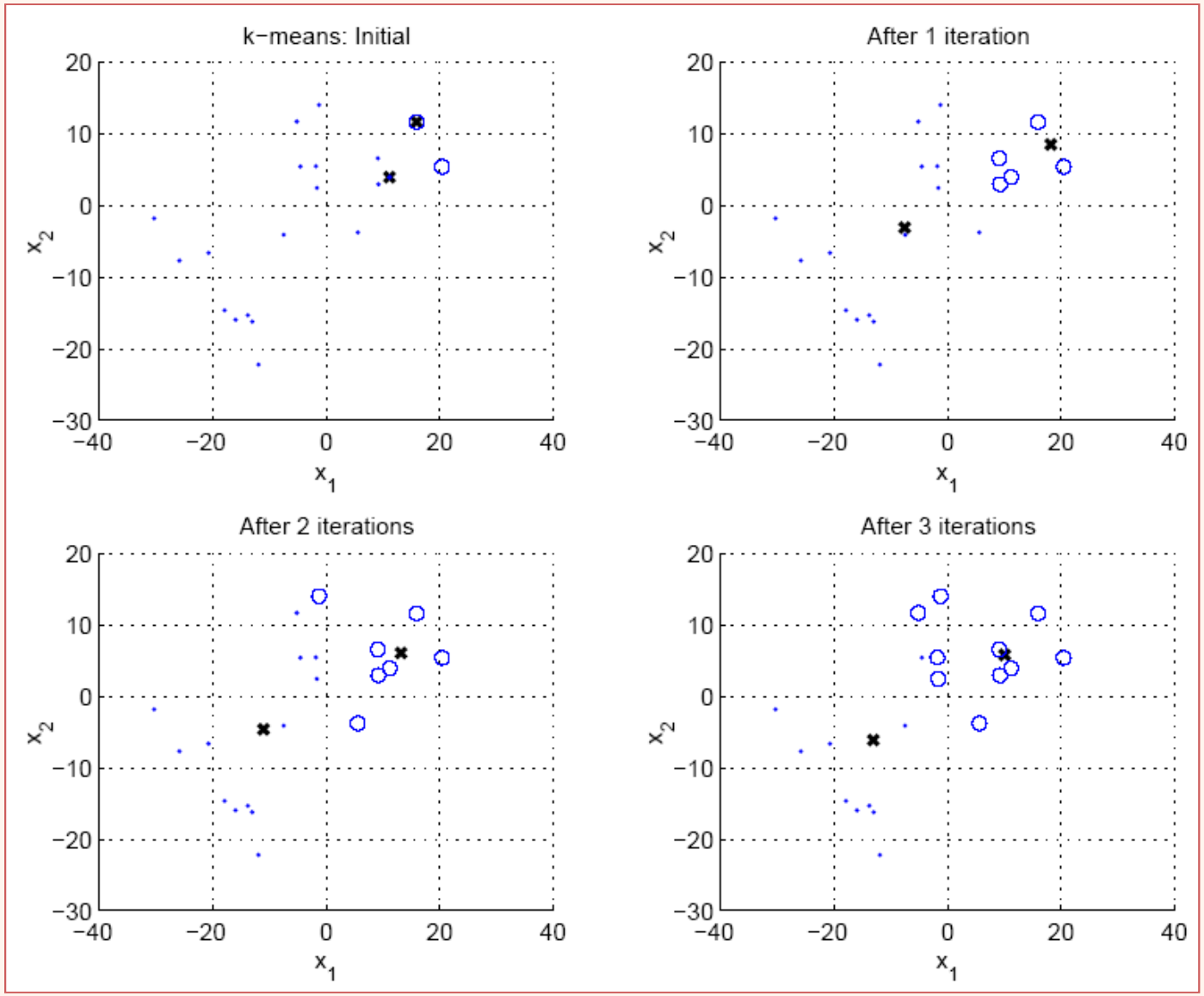
$$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$

Until $\boldsymbol{m}_i$ converge

در صورتی که بخش‌بندی داده‌ها تغییر نکند، الگوریتم به پایان رسیده است.

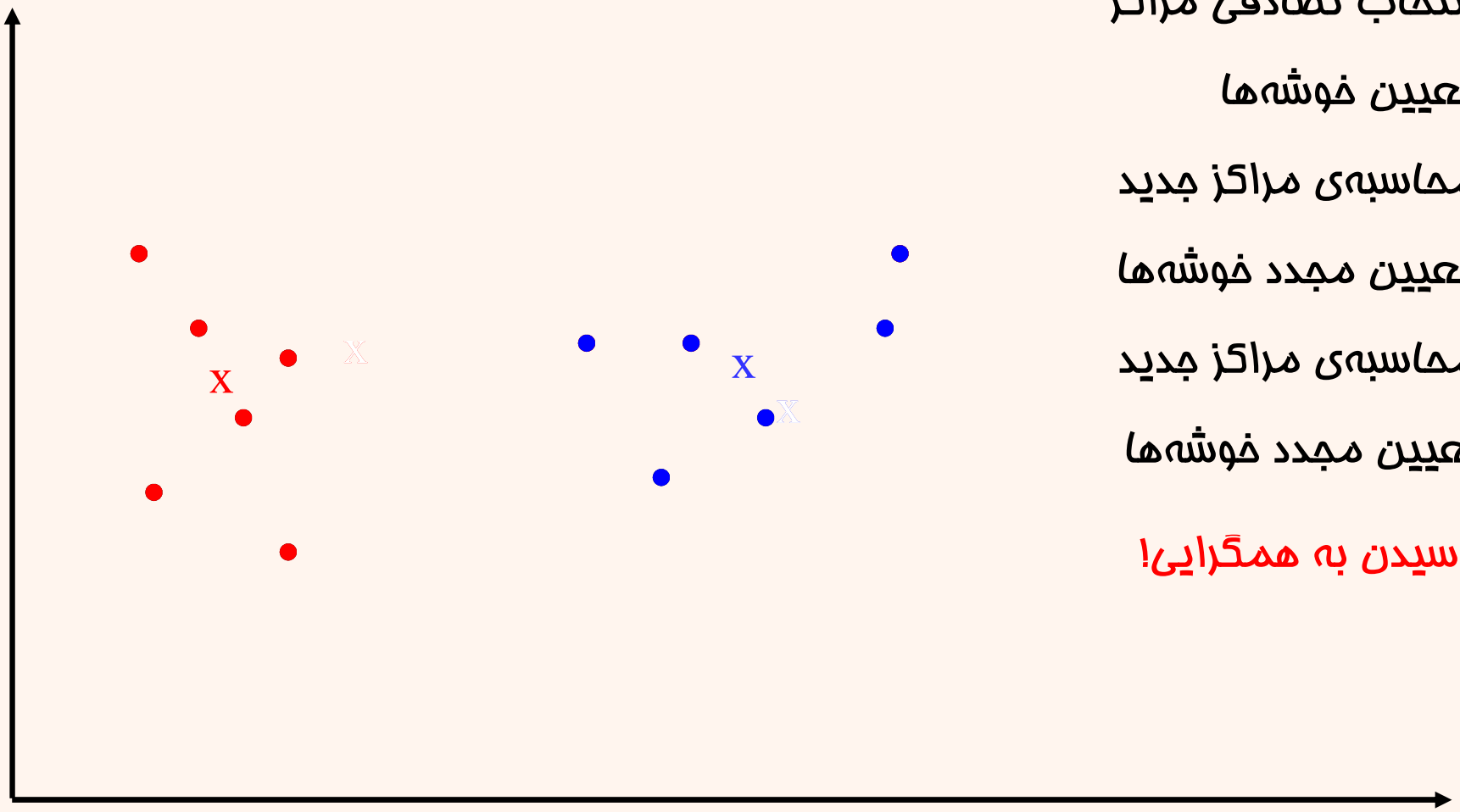
k-Means Clustering

مثال ۱



مثال ۲

K=2



انتخاب تصادفی مراکز

تعیین خوشه‌ها

محاسبه‌ی مراکز جدید

تعیین مجدد خوشه‌ها

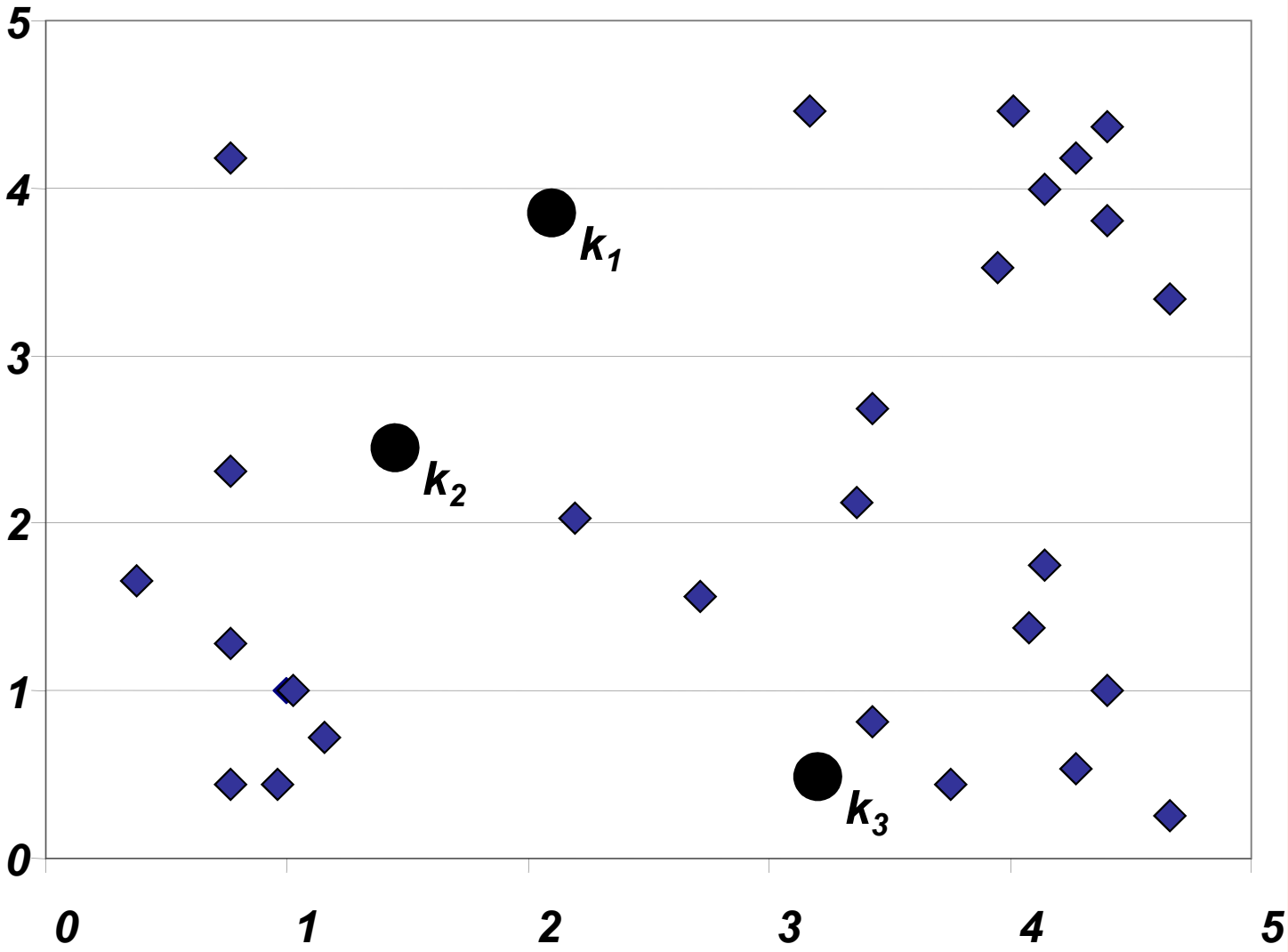
محاسبه‌ی مراکز جدید

تعیین مجدد خوشه‌ها

رسیدن به همگرایی!

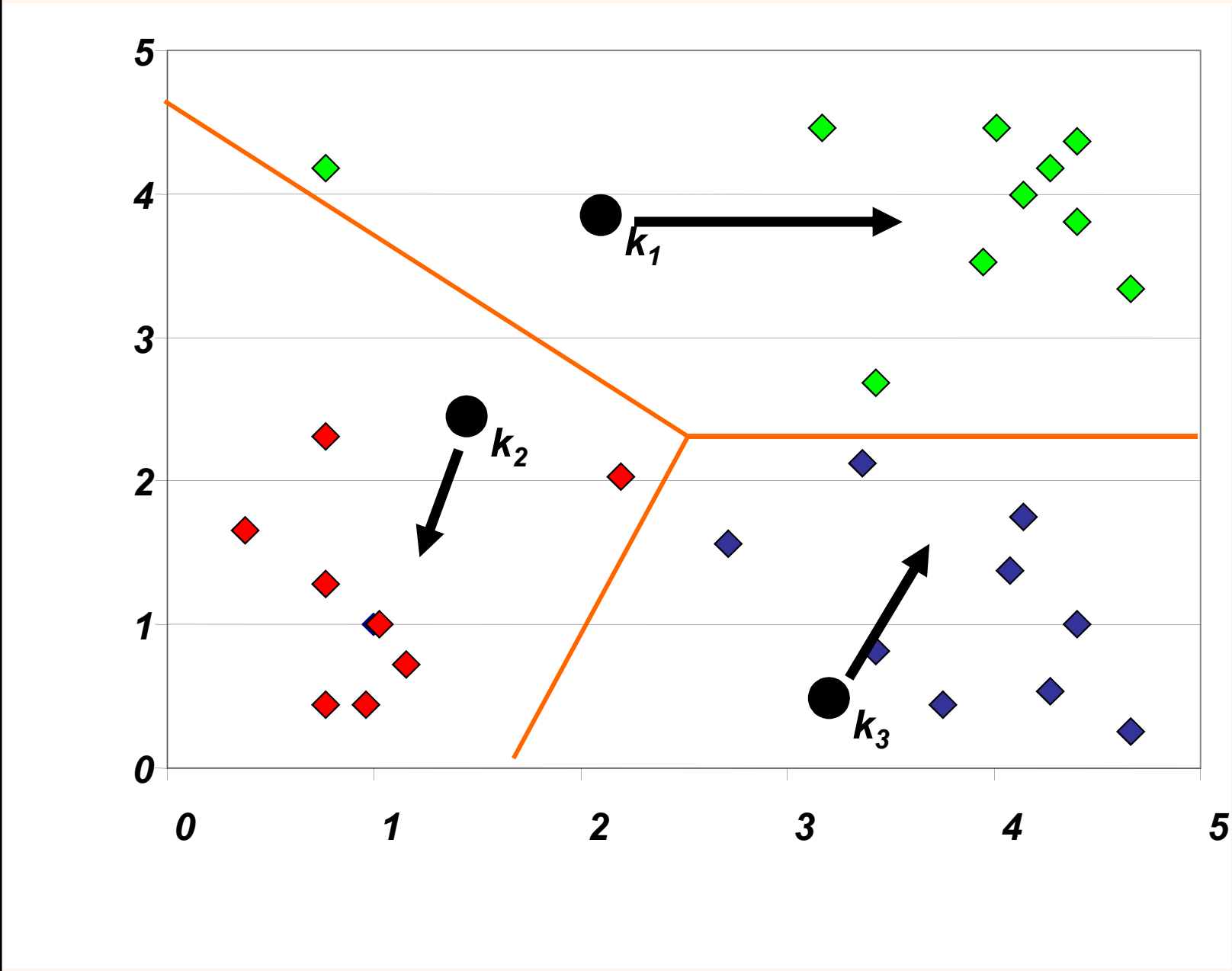
K=3

مثال ۳



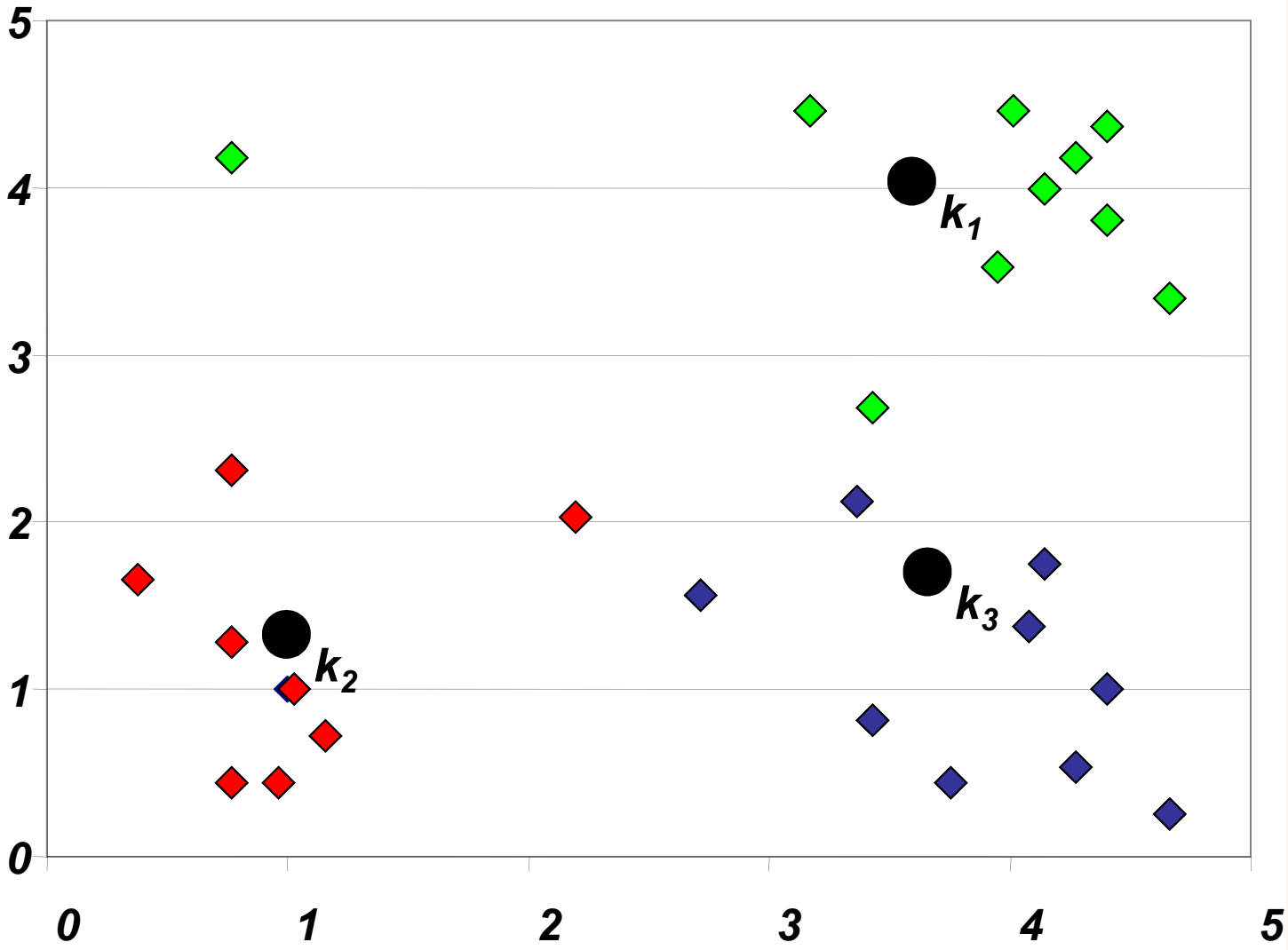
مثال ۳ (ادامه...)

K=3



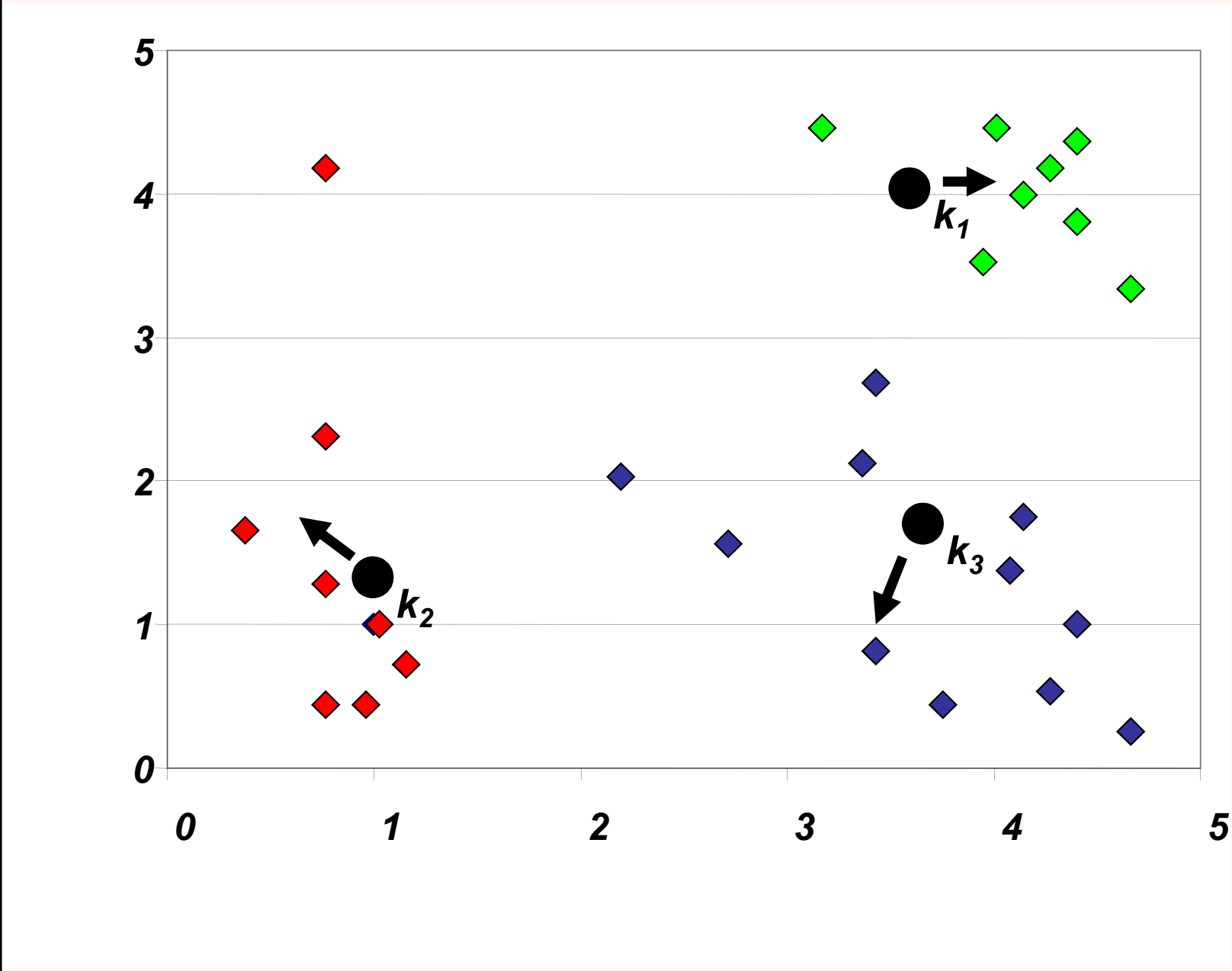
K=3

مثال ۳ (ادامه...)



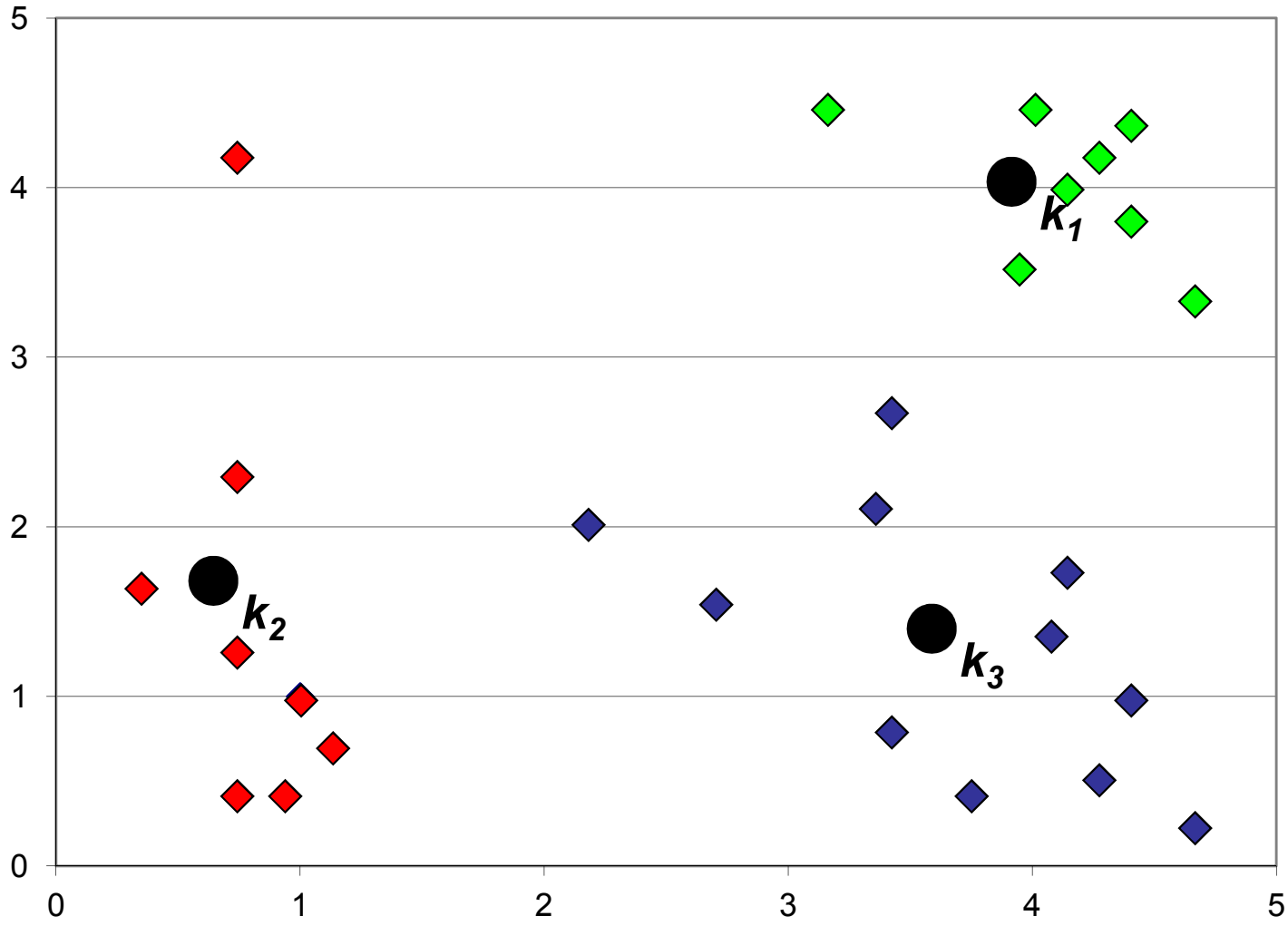
K=3

مثال ۳ (ادامه...)



K=3

مثال ۳ (ادامه...)



مشکلات k-means

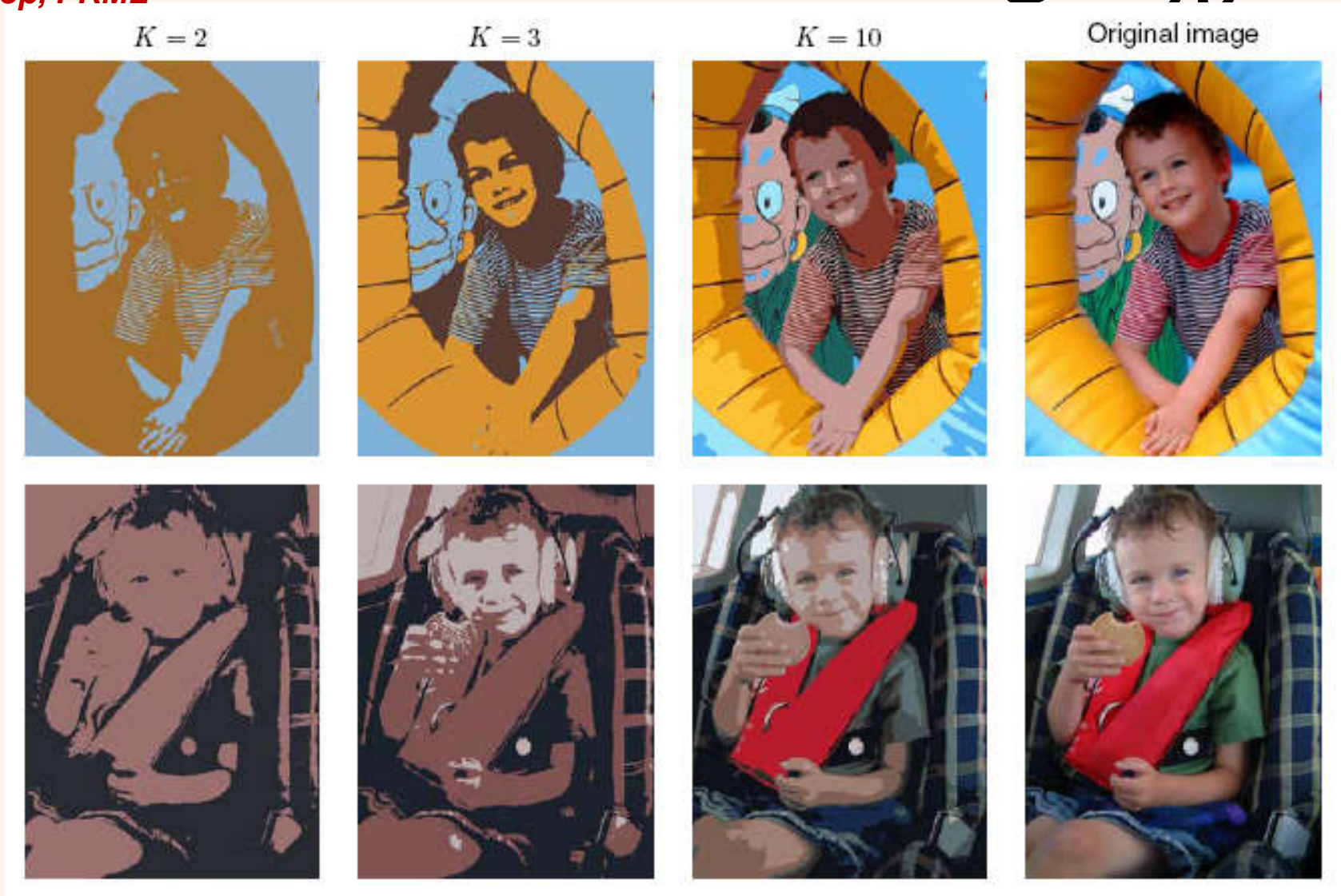
- این فرآیند، جستجوی محلی است و پاسخ نهایی وابسته به مقدار اولیه‌ی بردارهای مرجع است.
- نسبت به داده‌ها پرت مقاوم نیست.
- مقدار k باید از قبل مشخص شود.

مقدار اولیه‌ی بردارهای مرجع

- انتخاب تصادفی همه‌ی بردارهای مرجع
- محاسبه‌ی مقدار میانگین همه نمونه‌ها و انتساب آن به بردارهای مرجع پس از افزودن مقداری تصادفی
- محاسبه‌ی اولین مؤلفه‌ی اساسی و تقسیم آن به k قسمت مساوی و انتساب مقدار میانگین هر قسمت به هر بردار مرجع

کاربردهای k-means

Bishop, PRML



- از کاربردهای دیگر، دسته‌بندی مشتریان، کشف داده‌های پرت، کشف نمونه‌های غیرعادی را می‌توان نام برد.

Leader cluster algorithm

- در این شیوه در صورتی که یک نمونه از بردارهای مرجع از یک حدآستانه دورتر باشد، یک بردار مرجع برابر با نمونه‌ی مذکور ایجاد می‌شود.
- در صورتی که ناحیه‌ی مربوط به یک بردار مرجع شامل تعداد زیادی نمونه باشند، در آن ناحیه نمونه‌ی جدیدی ایجاد می‌شود.
- به طریق مشابه، در صورتی که ناحیه مربوط به یک بردار مرجع، شامل تعداد کمی نمونه باشد، آن ناحیه حذف می‌شود.

Expectation-Maximization (EM)

- در صورتی که بخواهیم با استفاده از MLE پارامترهای یک مدل ترکیبی را تخمین بزنیم:

$$\mathcal{L}(\Phi \mid \mathcal{X}) = \log \prod_t p\left(\mathbf{x}^t \mid \Phi\right)$$

$$= \sum_t \log \sum_{i=1}^{k} p\left(\mathbf{x}^t \mid G_i\right) P\left(G_i\right)$$

- به عنوان مثال در حالتی که توزیع هر خوشه، گاوسی باشد، پارامترهای مدل

$$p(\mathbf{x} \mid \mathcal{G}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i), \text{ and } \Phi = \{P(\mathcal{G}_i), \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^{k}$$

- در این حالت راه حل تحلیلی وجود ندارد، از این رو روش‌های تکرار شونده مورد استفاده قرار می‌گیرد.
- این روش برای زمانی مناسب است که برخی پارامترها «پنهان» هستند.

Expectation-Maximization (EM)

- فرض می‌شود متغیرهای پنهان(Z) وجود دارند که چنان‌چه مشخص باشند، مسأله‌ی بهینه‌سازی به سادگی حل می‌شود.
- هدف این الگوریتم یافتن پارامترهایی (Φ) است که احتمال رخداد متغیرهای قابل مشاهده ($L(\Phi \mid X)$) را بیشینه کند.

Incomplete likelihood

- در مواردی که یافتن پارامترها، امکان‌پذیر نیست، متغیرهای پنهان نیز مورد استفاده قرار می‌گیرند:

$L_c(\Phi \mid X,Z)$

Complete likelihood

دو گام این الگوریتم

E-step

- تخمین Z از روی داده‌های آموزشی و پارامترهای فعلی
  - در واقع $E(Z|X, \Phi^l)$ را محاسبه می‌کنیم.

M-step

با در اختیار داشتن تخمین متغیرهای پنهان و داده‌های آموزشی مقدار پارامترها به گونه‌ای انتخاب می‌شوند که تابع درست‌نمایی بیشینه شود.

$$\text{E-step}: \mathcal{Q}\left(\Phi \mid \Phi^l\right) = E\left[\mathcal{L}_C\left(\Phi \mid \mathcal{X}, \mathcal{Z}\right) \mid \mathcal{X}, \Phi^l\right]$$

$$\text{M-step}: \Phi^{l+1} = \arg\max_{\Phi} \mathcal{Q}\left(\Phi \mid \Phi^l\right)$$

- ثابت شده است با این شیوه در هر تکرار درست‌نمایی افزایش می‌یابد.

$$\mathcal{L}\left(\Phi^{l+1} \mid \mathcal{X}\right) \geq \mathcal{L}\left(\Phi^l \mid \mathcal{X}\right)$$

Dempster, Arthur P, Nan M. Laird, and Donald B. Rubin. 1976. Maximum Likelihood from Incomplete Data via the EM Algorithm. Paper presented at the Royal Statistical Society at a meeting organized by the Research Section, December 8, 1976.

EM in Gaussian Mixtures

- در مثال ترکیب توزیع‌ها، «متغیرهای پنهان» مشخص می‌کنند کدام نمونه به کدام خوشه تعلق دارد.
  - در صورتی که تعلق هر نمونه به خوشه‌ی متناظرش (برچسب) مشخص باشد(مانند حالت باناظر)، می‌توان به راحتی پارامترهای هر توزیع را به دست آورد.
- در گام E، بر اساس دانش فعلی، این برچسب‌ها تقریب زده می‌شوند.
- در گام M، بر اساس تخمین زده شده، اطلاعاتی که در مورد کلاس داریم، را به روز می‌کنیم.

این دو گام چه شباهتی با دو مرحله‌ی k-means دارند؟

EM in Gaussian Mixtures (cnt'd…)

- بردار z^t متغیر پنهان در این مسأله است.

$$\mathbf{z}^t = \{z_1^t, \ldots, z_k^t\}$$

- $z^t_i = 1$ اگر x^t به خوشه‌ی i-ام تعلق داشته باشد.
  - این متغیر دارای توزیع برنولی تعمیم‌یافته است.
  - در واقع شبیه به r^t_i در حالت باناظر است.

اگر zها مشخص باشند، مانند حالت «بانظارت» است:

$$P(C_i) = \frac{\sum_t r_i^t}{N} \qquad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S_i} = \frac{\sum_t r_i^t \left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)\left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)^T}{\sum_t r_i^t}$$

EM in Gaussian Mixtures (cnt'd...)

- در گام نخست، مقادیر متغیر پنهان را با توجه به دانش فعلی تقریب می‌زنیم:

$$h_i^t \equiv E\left(G_i|\mathbf{x}^t,\Phi^l\right)=P\left(z_i^t=1|\mathbf{x}^t,\Phi^l\right)=\frac{p\left(\mathbf{x}^t|G_i,\Phi^l\right)P\left(G_i\right)}{\sum_j p\left(\mathbf{x}^t|G_j,\Phi^l\right)P\left(G_j\right)}$$

$$P(\mathbf{z}^t)=\prod_{i=1}^{k}\pi_i^{z_i^t}$$

$$p(\mathbf{x}^t|\mathbf{z}^t)=\prod_{i=1}^{k}p_i(\mathbf{x}^t)^{z_i^t}$$

$$\begin{aligned}\mathcal{L}_c(\Phi|\mathcal{X},\mathcal{Z}) &= \log\prod_t p(\mathbf{x}^t,\mathbf{z}^t|\Phi)\\ &= \sum_t \log p(\mathbf{x}^t,\mathbf{z}^t|\Phi)\\ &= \sum_t \log P(\mathbf{z}^t|\Phi)+\log p(\mathbf{x}^t|\mathbf{z}^t,\Phi)\\ &= \sum_t\sum_i z_i^t[\log\pi_i+\log p_i(\mathbf{x}^t|\Phi)]\end{aligned}$$

$$P(G_i)=\pi_i$$

$$p(\mathbf{x}^t,\mathbf{z}^t)=P(\mathbf{z}^t)p(\mathbf{x}^t|\mathbf{z}^t)$$

$$p_i\left(\mathbf{x}^t\right)=p\left(\mathbf{x}^t\mid G_i\right)$$

$$\mathcal{Q}\left(\Phi\mid\Phi^l\right)=E\left[\mathcal{L}_C\left(\Phi\mid\mathcal{X},\mathcal{Z}\right)\mid\mathcal{X},\Phi^l\right]$$

EM in Gaussian Mixtures (cnt'd…)

$$\text{M-step}: \Phi^{l+1} = \arg\max_{\Phi} \mathcal{Q}\left(\Phi \mid \Phi^{l}\right)$$

$$\begin{aligned}
\mathcal{Q}(\Phi|\Phi^l) &\equiv E\left[\log P(X,Z)|\mathcal{X},\Phi^l\right] \\
&= E\left[\mathcal{L}_c(\Phi|\mathcal{X},\mathcal{Z})|\mathcal{X},\Phi^l)\right] \\
&= \sum_t \sum_i E[z_i^t|\mathcal{X},\Phi^l][\log \pi_i + \log p_i(\boldsymbol{x}^t|\Phi^l)]
\end{aligned}$$

$$E\left[z_i^t|\mathbf{x}^t,\Phi^l\right] = P\left(z_i^t = 1|\mathbf{x}^t,\Phi^l\right) \equiv h_i^t$$

$$\begin{aligned}
\mathcal{Q}(\Phi|\Phi^l) &= \sum_t \sum_i h_i^t[\log \pi_i + \log p_i(\boldsymbol{x}^t|\Phi^l)] \\
&= \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p_i(\boldsymbol{x}^t|\Phi^l)
\end{aligned}$$

EM in Gaussian Mixtures (cnt'd…)

$$\text{M-step}: \Phi^{l+1} = \arg\max_{\Phi} \mathcal{Q}\left(\Phi \mid \Phi^{l}\right)$$

- در گام M: در این مرحله بر اساس تخمین متغیرهای پنهان، پارامترهای به روز شده‌ی مدل به گونه‌ای انتخاب می‌شوند که Q ماکزیمم شود:

$$\nabla_{\pi_i} \sum_t \sum_i h_i^t \log \pi_i - \lambda \left(\sum_i \pi_i - 1 \right) = 0$$

$$P(G_i) = \pi_i, \quad \sum \pi_i = 1$$

$$P(G_i) = \frac{\sum_t h_i^t}{N}$$

$$P(C_i) = \frac{\sum_t r_i^t}{N} \qquad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

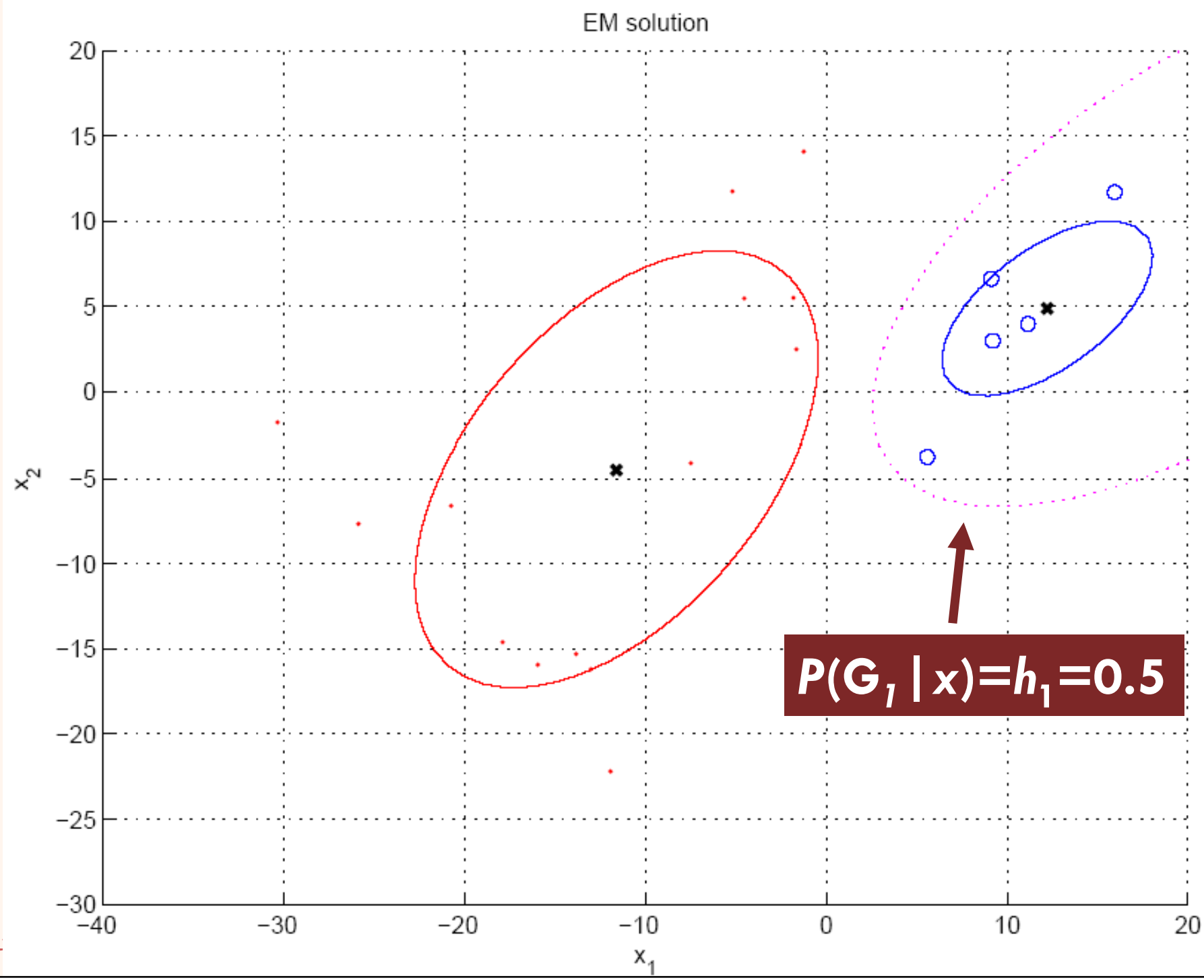
$$\mathbf{S_i} = \frac{\sum_t r_i^t \left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)\left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)^T}{\sum_t r_i^t}$$

$$\nabla_{\Phi} \sum_t \sum_i h_i^t \log p_i(\mathbf{x}^t \mid \Phi) = 0$$

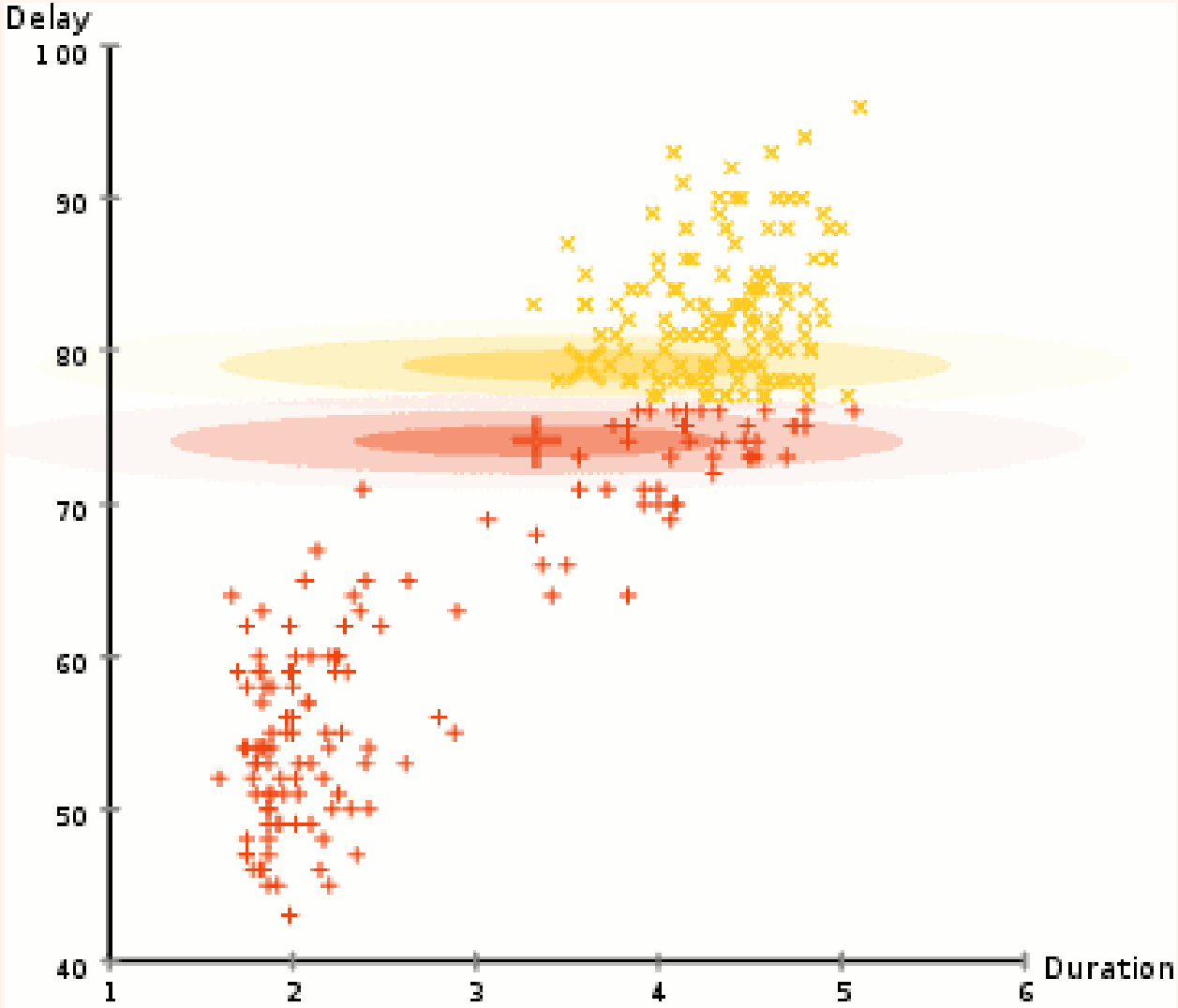
Soft label

$$\mathbf{m}_i^{l+1} = \frac{\sum_t h_i^t \mathbf{x}^t}{\sum_t h_i^t} \qquad \mathbf{S}_i^{l+1} = \frac{\sum_t h_i^t \left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)\left(\mathbf{x}^t - \mathbf{m}_i^{l+1}\right)^T}{\sum_t h_i^t}$$

مثال


EM solution
20
15
10
5
0
−5
−10
−15
−20
−25
−30
−40
−30
−20
−10
0
10
20
x_1
x_2
$P(G_1 \mid x)=h_1=0.5$

مثال

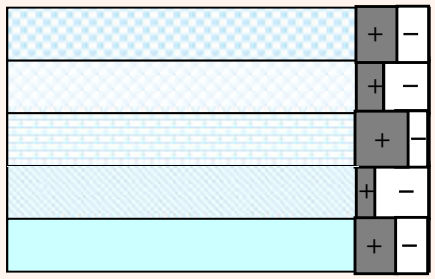


Expectation-Maximization (EM)

E Step:

- مقداردهی اولیه:
  - بر اساس پارامترهای کنونی به داده‌ها، برچسب‌هایی نرم (احتمال تعلق به خوشه) نسبت داده می‌شود.

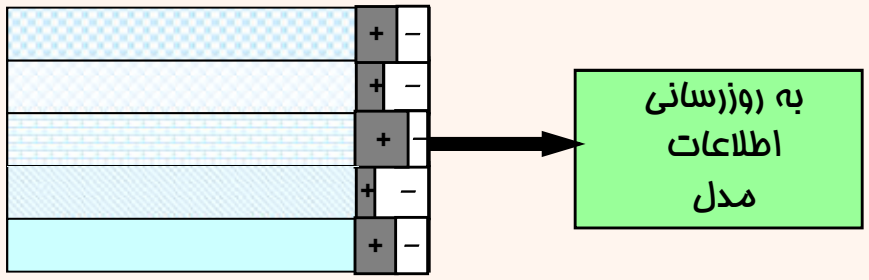
داده‌های بدون برچسب



Expectation-Maximization (EM)

M Step:

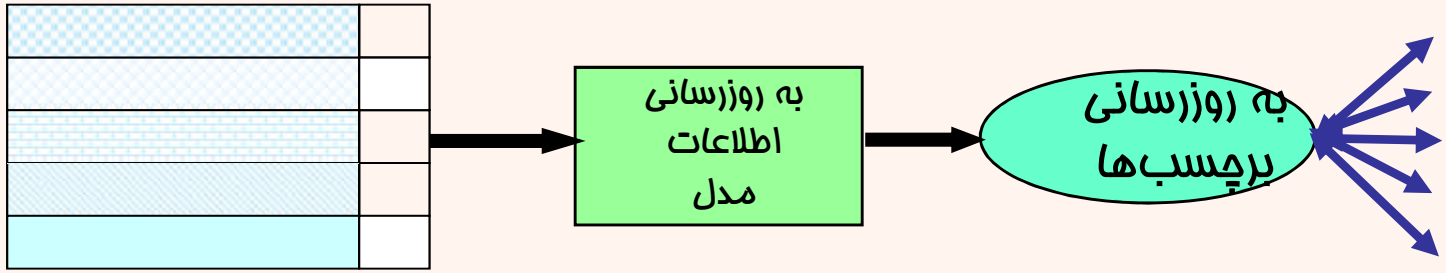
- براساس برچسب‌های نسبت داده شده، اطلاعات توزیع‌ها به روز می‌شود:

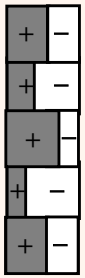


Expectation-Maximization (EM)

E Step:

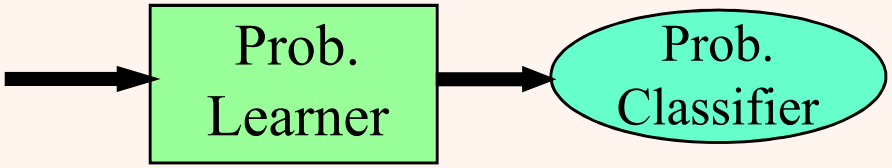
- اطلاعات برچسب‌ها براساس مدل به دست آمده به روز می‌شوند:

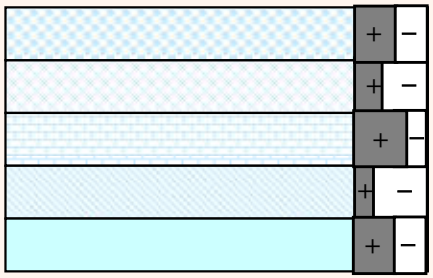




Expectation-Maximization (EM)

M Step:





این روند تا زمانی که برچسب‌های به دست آمده به همگرایی برسد ادامه خواهد یافت.

جمع‌بندی

- پارامترهای اولیه مقداردهی می‌شود: $initialize \quad \Phi^0$
- تا زمانی رسیدن به همگرایی تکرار کن:
  - گام E: $E\left(Z|X,\Phi^l\right)$ را تقریب بزن
  - گام M: $\Phi^{l+1} = \arg\max_{\Phi} \mathcal{Q}\left(\Phi \mid \Phi^l\right)$

- برای مقداردهی اولیه، از K-means استفاده می‌شود، بعد از چند تکرار، تخمین میانگین محاسبه شده و پس از مشخص شدن اعضای هر خوشه، ماتریس کواریانس تخمین زده شده و $E(G_i)$ تخمین زده شده و الگوریتم EM آغاز می‌شود.

Expectation-Maximization (EM)

- در صورتی که داده‌ها با توزیع گاوسی در نظر گرفته شوند:

$$h_i^t \equiv E\left(G_i | \mathbf{x}^t, \Phi^l\right)$$

$$h_i^t = \frac{\pi_i |\mathbf{S}_i|^{-1/2} \exp[-(1/2)(\boldsymbol{x}^t - \boldsymbol{m}_i)^T \mathbf{S}_i^{-1} (\boldsymbol{x}^t - \boldsymbol{m}_i)]}{\sum_j \pi_j |\mathbf{S}_j|^{-1/2} \exp[-(1/2)(\boldsymbol{x}^t - \boldsymbol{m}_j)^T \mathbf{S}_j^{-1} (\boldsymbol{x}^t - \boldsymbol{m}_j)]}$$

- مانند روش‌های پارامتری در این جا نیز در حالتی که داده‌های آموزشی کم‌تعداد است یا ابعاد ورودی زیاد است، می‌توان از مدل‌های ساده‌تری استفاده کرد تا مشکل overfitting رخ ندهد.

انتخاب مدل

- در صورتی که برای همه‌ی خوشه‌ها کواریانس یکسانی در نظر بگیریم، با رابطه‌ی ساده‌تری مواجه خواهیم شد:

$$\min_{\boldsymbol{m}_i, S} \sum_t \sum_i h_i^t (\boldsymbol{x}^t - \boldsymbol{m}_i)^T S^{-1} (\boldsymbol{x}^t - \boldsymbol{m}_i)$$

- در صورتی که توزیع‌های هر خوشه، ناهمبسته بوده و واریانس یکسانی داشته باشند:

$$\min_{\boldsymbol{m}_i, s} \sum_t \sum_i h_i^t \frac{\|\boldsymbol{x}^t - \boldsymbol{m}_i\|^2}{s^2}$$

- بسیار شبیه به k-means است ، با این تفاوت که برچسب‌ها در این جا بین صفر و یک هستند.

$$h_i^t = \frac{\exp\left[-(1/2s^2)\|\boldsymbol{x}^t - \boldsymbol{m}_i\|^2\right]}{\sum_j \exp\left[-(1/2s^2)\|\boldsymbol{x}^t - \boldsymbol{m}_j\|^2\right]}$$

انتخاب مدل (ادامه...)

- در نظر گرفتن ماتریس کواریانس یکسان، موجب نادیده گرفتن شکل واقعی خوشه‌ها می‌شود.
  - در نظر گرفتن ماتریس کواریانس قطری با توجه به نادیده گرفتن همبستگی‌ها، به طریق اولی ساختار واقعی را نادیده می‌گیرد.
- به عنوان راه حل، پیش از خوشه‌بندی می‌توان از روش‌های کاهش ابعاد(PCA/FA) بهره برد.

$$p(\mathbf{x}_t \mid G_i) = \mathcal{N}\left(\mathbf{m}_i, \mathbf{V}_i\mathbf{V}_i^T + \boldsymbol{\psi}_i\right)$$

خوشه‌بندی برای استخراج دانش

- مانند کاهش ابعاد برای خوشه‌بندی نیز می‌توان دو هدف متفاوت در نظر گرفت:
- «استخراج دانش»: برای فهم بهتر ساختار داده‌ها مورد استفاده قرار می‌گیرد.
  - کاهش ابعاد همبستگی بین خصیصه‌ها را می‌یابد.
  - خوشه‌بندی شباهت بین نمونه‌های داده را مشخص می‌کند.
- پس از خوشه‌بندی استخراج دانش توسط متخصص قابل انجام است، همچنین پارامترهای خوشه‌بندی نظیر میانگین خوشه‌ها و تعداد آن هم قابل استفاده می‌باشد.
  - از کاربردها می‌توان به CRM اشاره کرد.

خوشه‌بندی به عنوان پیش‌پردازش

- همان‌گونه که در کاهش ابعاد، فضای جدید برای فرآیندهای بعدی (دسته‌بندی، رگرسیون) مورد استفاده قرار می‌گیرد، خوشه‌بندی نیز داده‌ها را به یک فضای k-بعدی نگاشت می‌کند. ابعاد فضای جدید شامل برچسب‌های به دست آمده است (h یا b)، بدین‌ترتیب ممکن با افزایش ابعاد هم مواجه شویم.
- در کاهش ابعاد همه‌ی داده‌ها در فرآیند مشارکت دارند، در حالی که در خوشه‌بندی مشارکت به صورت محلی صورت می‌پذیرد.
- در صورت استفاده از چنین پیش‌پردازش‌هایی می‌توان از یک مجموعه داده‌های بدون‌برچسب در فرآیند آموزش بهره برد.
- در نهایت در دسته‌بندی با ترکیبی از ترکیب‌ها مواجه خواهیم بود:

$$p(\mathbf{x}\mid C_i)=\sum_{j=1}^{k_i} p(\mathbf{x}\mid G_{ij})P(G_{ij})$$

$$p(\mathbf{x})=\sum_{i=1}^{K} p(\mathbf{x}\mid C_i)P(C_i)$$

Mixture of Mixtures

خوشه‌بندی سلسله‌مراتبی

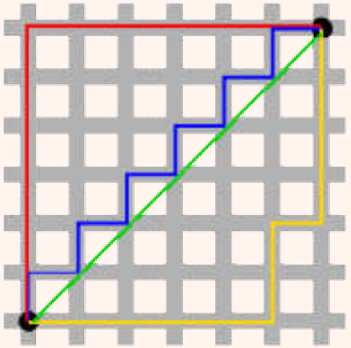
- در k-means هدف مینیمم کردن خطای بازسازی است.
- در «خوشه‌بندی سلسله مراتبی»، تنها شباهت بین نمونه‌ها در نظر گرفته می‌شود.
- افزون بر فاصله‌ی اقلیدسی معیارهای دیگری نیز در نظر گرفته می‌شوند:

Minkowski

$$d_m\left(\mathbf{x}^r,\mathbf{x}^s\right)=\left[\sum_{j=1}^{d}\left(x_j^r-x_j^s\right)^p\right]^{1/p}$$

City-block distance

$$d_{cb}\left(\mathbf{x}^r,\mathbf{x}^s\right)=\sum_{j=1}^{d}\left|x_j^r-x_j^s\right|$$



خوشه‌بندی سلسله‌مراتبی(ادامه...)

Hierarchical Agglomerative Clustering(HAC)

- با N خوشه کار آغاز می‌شود؛ هر خوشه شامل یک نمونه می‌باشد.
- خوشه‌های نزدیک به هم در هر تکرار با هم ادغام می‌شوند.
  - برای انتخاب گروه‌های نزدیک، سه معیار مورد استفاده قرار می‌گیرد:

$$d(G_i, G_j) = \min_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

Single-link

$$d(G_i, G_j) = \max_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

Complete-link

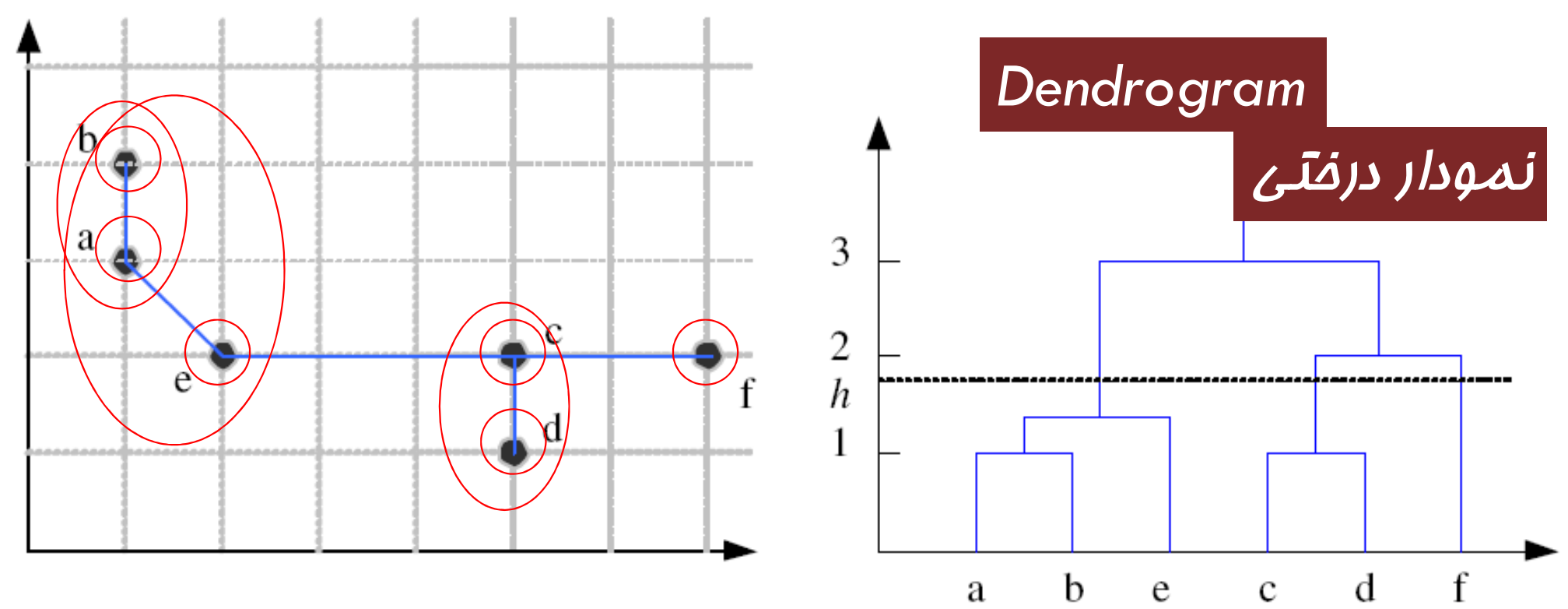
$$d(G_i, G_j) = \operatorname{ave}_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s)$$

Average-link, centroid

- این روند تا زمانی که تنها یک خوشه وجود داشته باشد، ادامه می‌یابد.

Agglomerative Clustering

خوشه‌بندی سلسله‌مراتبی(ادامه...)

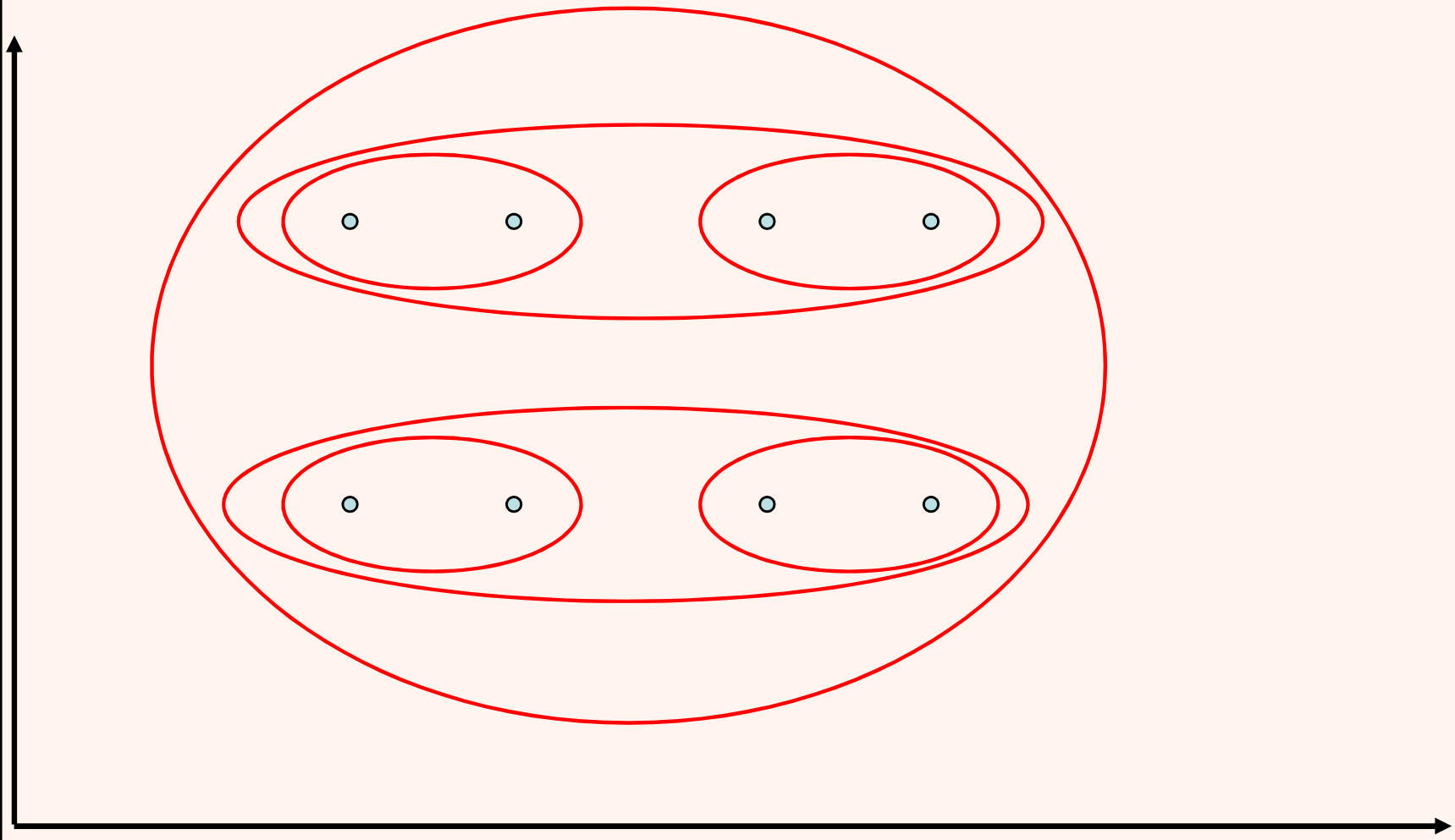


Divisive Clustering

در این شیوه به صورت عکس عمل می‌شود؛ از یک خوشه کار آغاز شده و خوشه‌ها در هر تکرار به خوشه‌های کوچک‌تر تقسیم می‌شوند تا زمانی که هر خوشه شامل یک نمونه باشد.

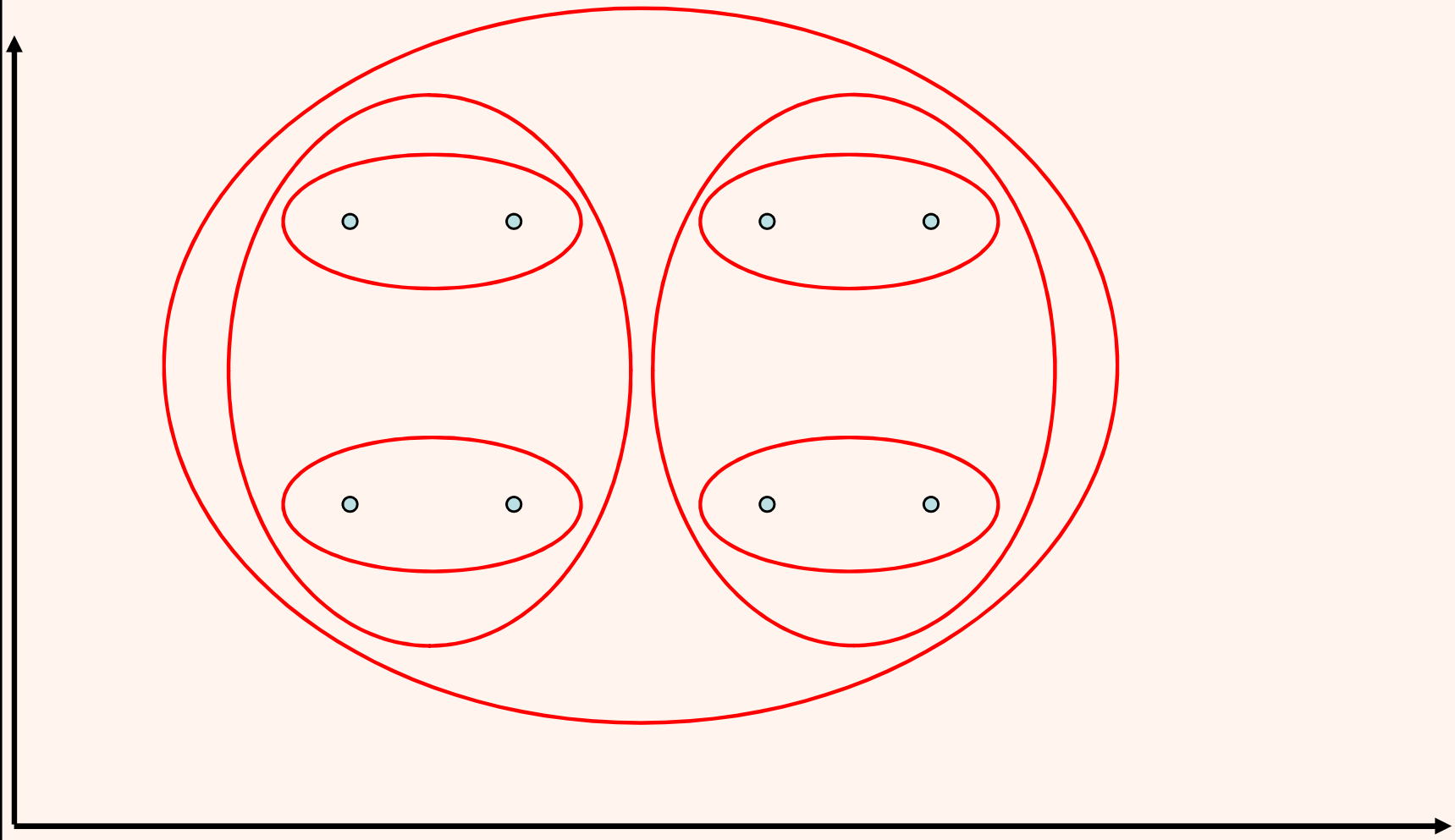
single-link clustering

مثال



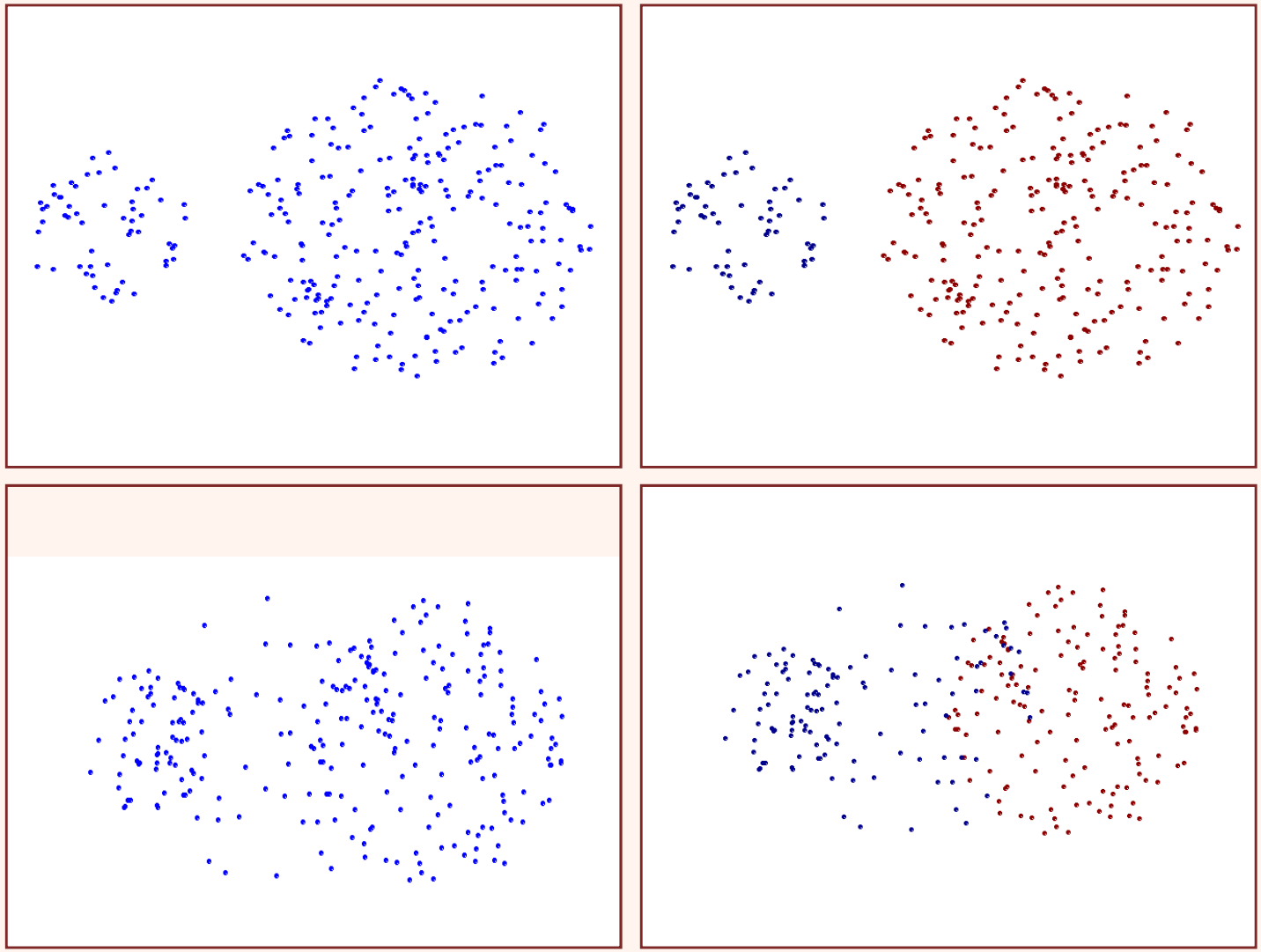
complete-link clustering

مثال



single-link clustering

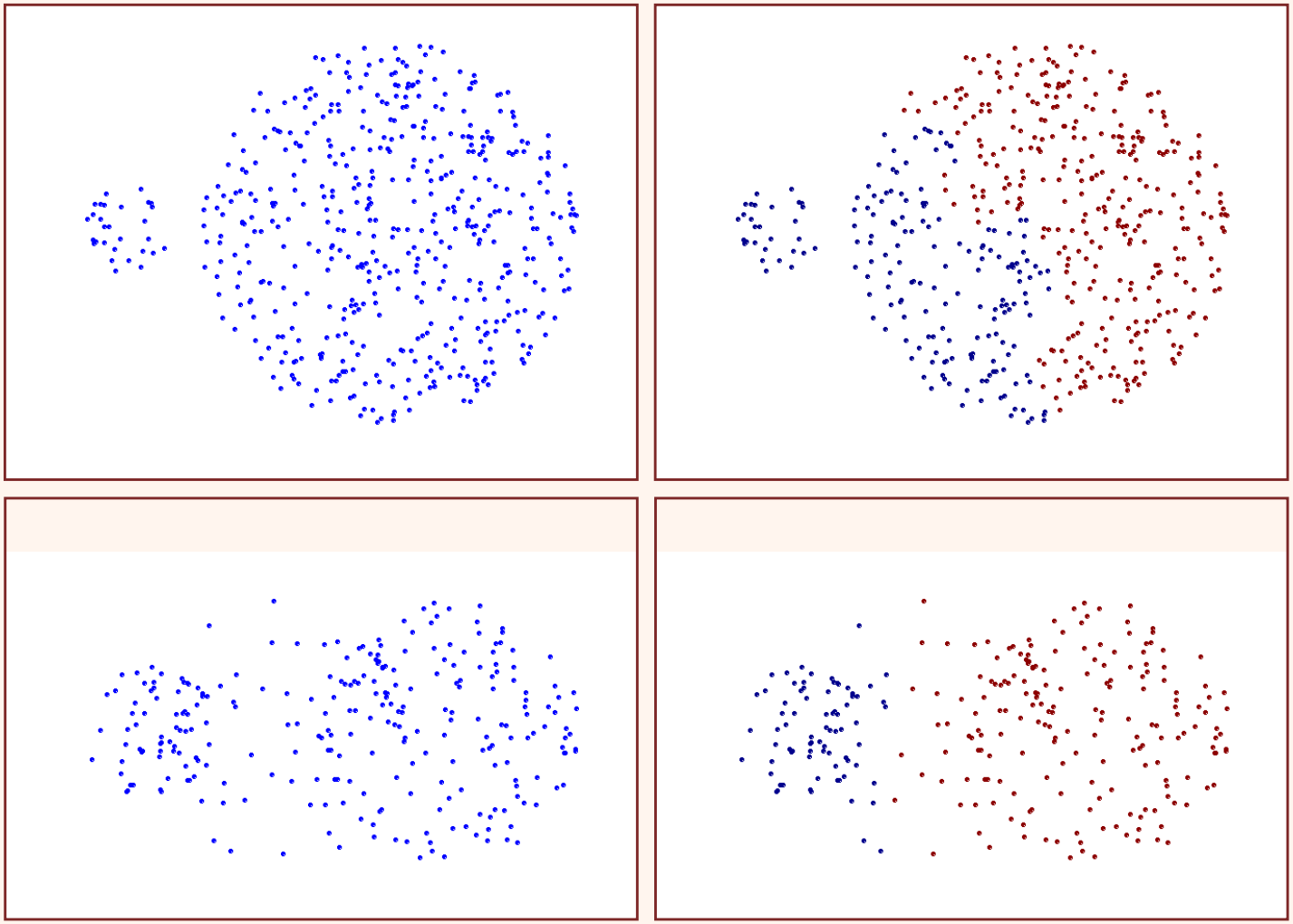
مقایسه‌ی معیارهای فاصله



این معیار به نویز و داده‌های پرت حساس است و خوشه‌های «کشیده» ایجاد می‌کند.

complete-link clustering

مقایسه‌ی معیارهای فاصله



خوشه‌های بزرگ را می‌شکند، خوشه‌ها با قطر یکسان تولید می‌کند.
خوشه‌های کوچک را با خوشه‌های بزرگ ادغام می‌کند.

انتخاب تعداد خوشه‌ها

- در برخی کاربردها، با توجه به نیاز k مشخص می‌شود، مانند color quantization
- استفاده از PCA و رسم داده‌ها در دو بعد می‌تواند ساختار داده‌ها را تا حدی مشخص کرده و انتخاب مناسب k کمک کند.
- استفاده از روش‌های افزایشی (leader-cluster)
- در برخی کاربردها بعد از انجام خوشه‌بندی، به صورت دستی می‌توان مناسب بودن خوشه‌ها را بررسی کرد؛ به عنوان مثال در برخی کاربردهای داده‌کاوی
- بسته به نوع الگوریتم خوشه‌بندی مورد استفاده می‌توان نمودار خطای بازسازی بر حسب k را رسم کرده و بر این اساس مقدار مناسب تعداد خوشه‌ها را یافت.

یادگیری نیمه‌نظارتی

- در دسته‌بندی بانظارت، تهیه‌ی داده‌ی برچسب‌خورده معمولا هزینه‌ی بالای تحمیل می‌کند.
- در یادگیری نیمه‌نظارتی، از داده‌های برچسب‌نخورده برای افزایش کارایی کمک گرفته می‌شود.
  - به عنوان مثال شیوه‌ی نیمه‌نظارتی EM به این صورت است که داده‌های برچسب‌خورده توسط ناظر با برچسب‌های اصلی خود در تخمین شرکت می‌کنند. این برچسب‌ها برخلاف برچسب‌های تخمین زده شده در طی فرآیند EM تغییر نمی‌کنند.

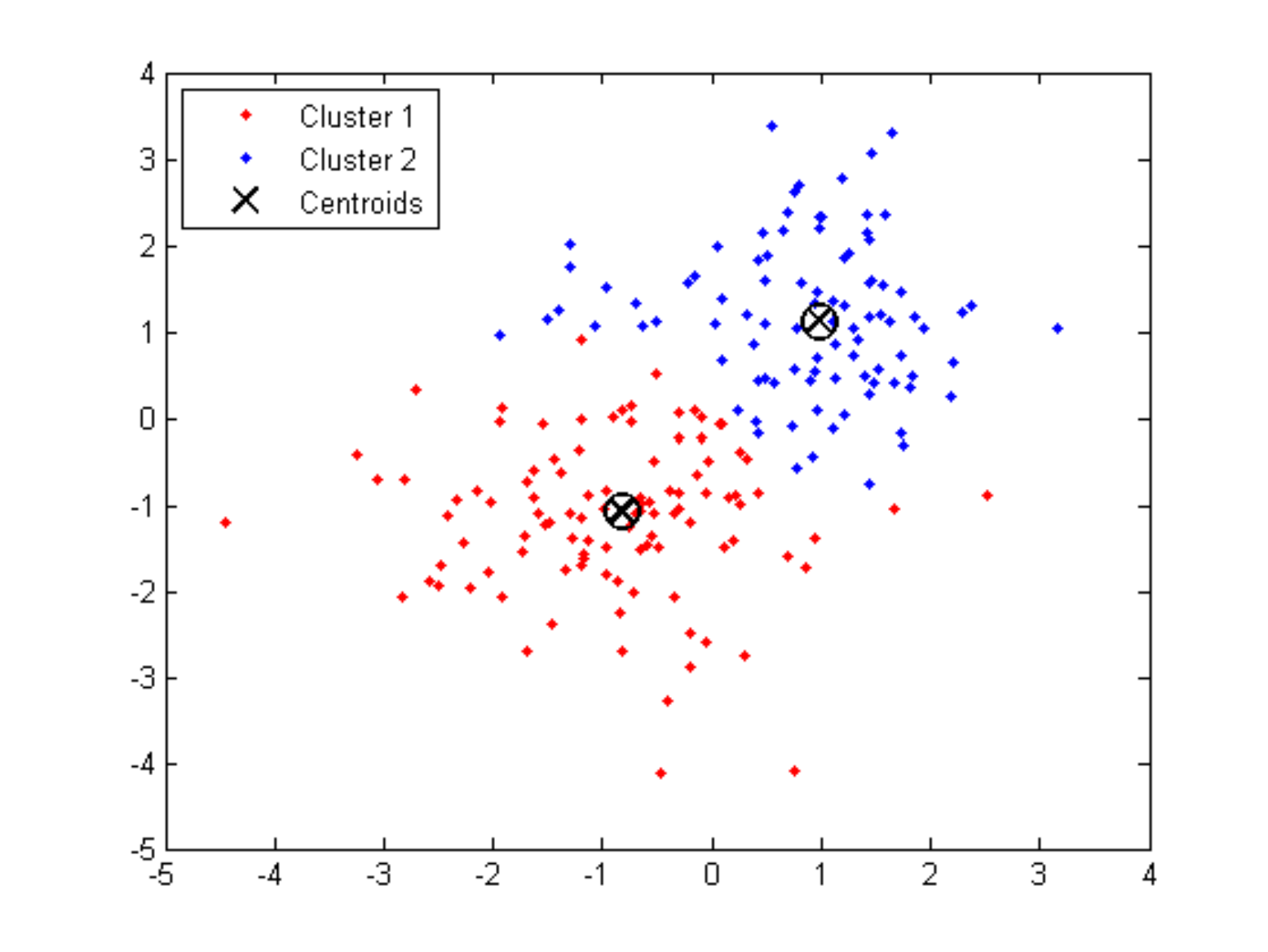
Matlab

مثال ۱

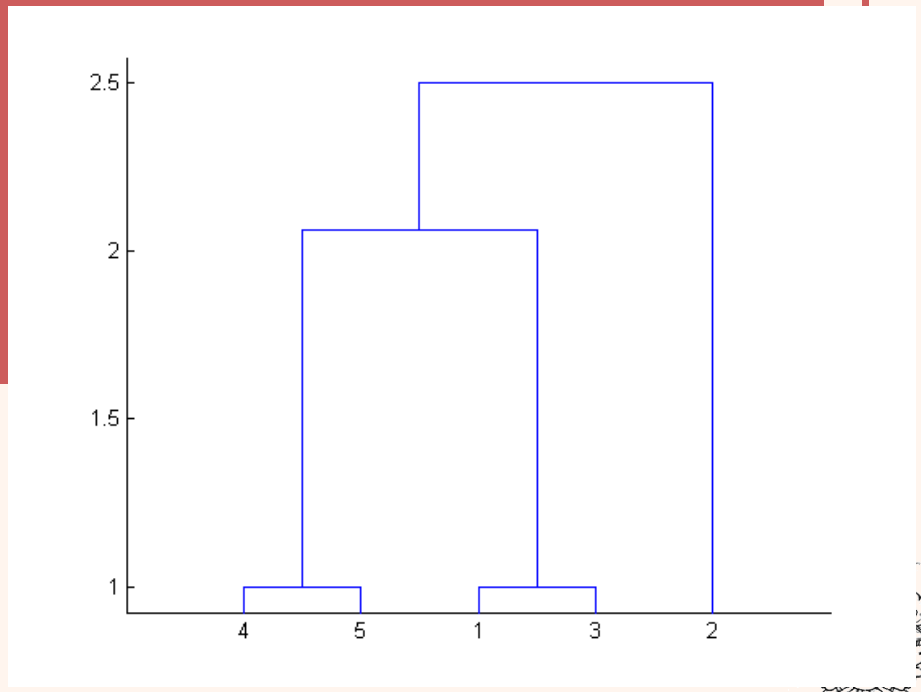
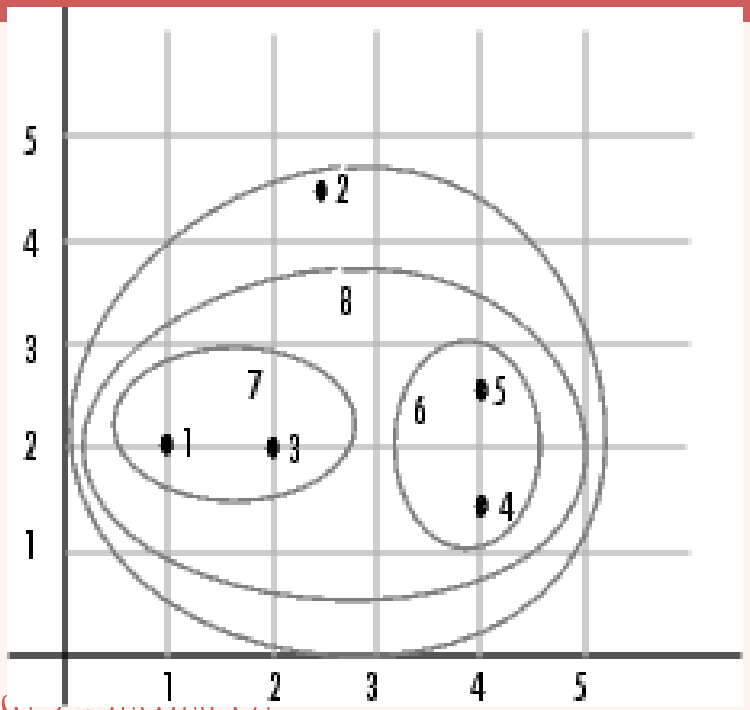
```
clear all, clc;
X = [randn(100,2)+ones(100,2);...
     randn(100,2)-ones(100,2)];
opts = statset('Display','iter');
[idx,ctrs] = kmeans(X,2,...
                    'Distance','city',...
                    'Replicates',5,...
                    'Options',opts);
plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',12)
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',12)
plot(ctrs(:,1),ctrs(:,2),'kx',...
     'MarkerSize',12,'LineWidth',2)
plot(ctrs(:,1),ctrs(:,2),'ko',...
     'MarkerSize',12,'LineWidth',2)
legend('Cluster 1','Cluster 2','Centroids',...
       'Location','NW')
```

Matlab

مثال ۱



Matlab

مثال ۲

```
clear all, close all, clc;
X=[1 2; 2.5 4.5; 2 2; 4 1.5; 4 2.5];
plot(X(:,1),X(:,2),'*');
axis([0 5 0 5]);
Y = pdist(X);
squareform(Y)
Z = linkage(Y, ,'single')
figure;
dendrogram(Z)
T = cluster(Z,'maxclust',2)
```