

یادگیری ماشین
(۰۱-۸۰۵-۱۱-۱۳)
فصل هفتم



دانشگاه شهید بهشتی

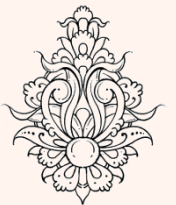
دانشکده مهندسی برق و کامپیوتر

پاییز ۱۳۹۴

احمد محمودی ازناوه

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 $(p(x | C_i))$).
- 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.

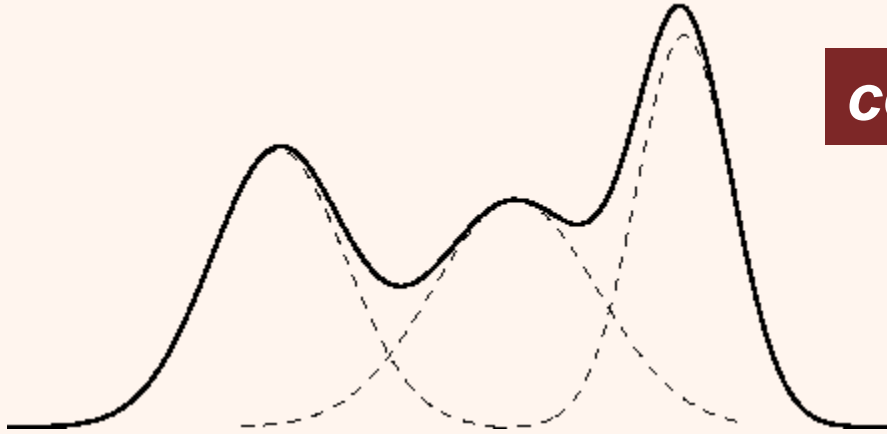


Mixture Densities

component densities

mixture proportions

$$p(\mathbf{x}) = \sum_{i=1}^k \overbrace{p(\mathbf{x} | G_i)}^{\text{component densities}} \underbrace{P(G_i)}_{\substack{\text{mixture proportions} \\ \downarrow \\ \text{components/groups/clusters}}}$$



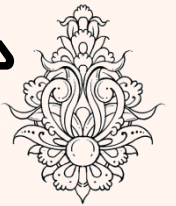
components/groups/clusters

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:

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



مفهوم «دسته» در مقایسه با «فوشه»

Classes vs. Clusters

Classification

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

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

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

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

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

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

- **Unsupervised:** $X = \{\mathbf{x}^t\}_t$
- Clusters $G_i, i=1, \dots, k$

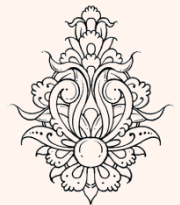
$$p(\mathbf{x}) = \sum_{i=1}^k p(\mathbf{x} | G_i) P(G_i)$$

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

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

Labels r^t ?

Clustering



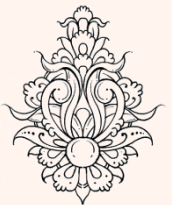
k-Means Clustering

- 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, $m_j, j = 1, \dots, k$

- بعد از مشخص شدن بردارهای مرجع، نمونه‌ها در خوشه‌ی نزدیک‌ترین بردار مرجع قرار می‌گیرند:
$$\|x^t - m_i\| = \min_j \|x^t - m_j\|$$
- بدین‌تربیب می‌توان به جای داده‌های از بردار مرجع متناظر آن استفاده کرد.



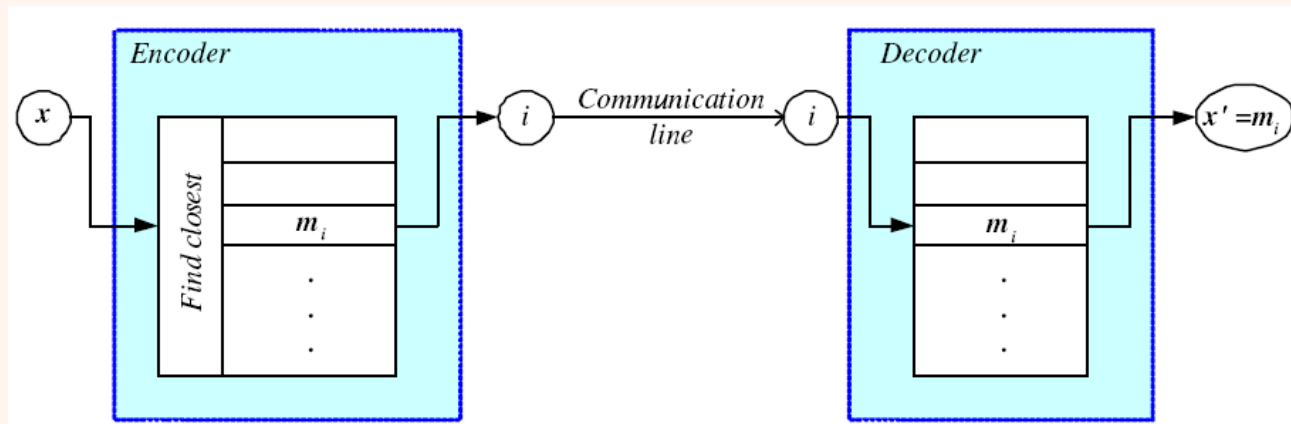
k-Means Clustering

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

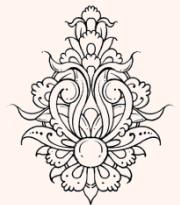
$$E(\{\mathbf{m}_i\}_{i=1}^k | \mathcal{X}) = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|$$

Reconstruction error

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



Encoding/Decoding



k-Means Clustering

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

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

- این رابطه افزون بر \mathbf{m}_i به برچسب‌ها b_i^t هم بستگی دارد، از این رو نمی‌توان برای آن راه حل تحلیلی یافت.

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

Repeat

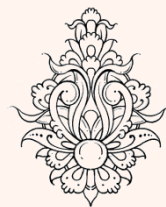
For all $\mathbf{x}^t \in \mathcal{X}$

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

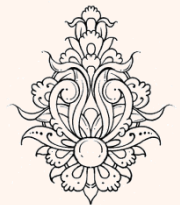
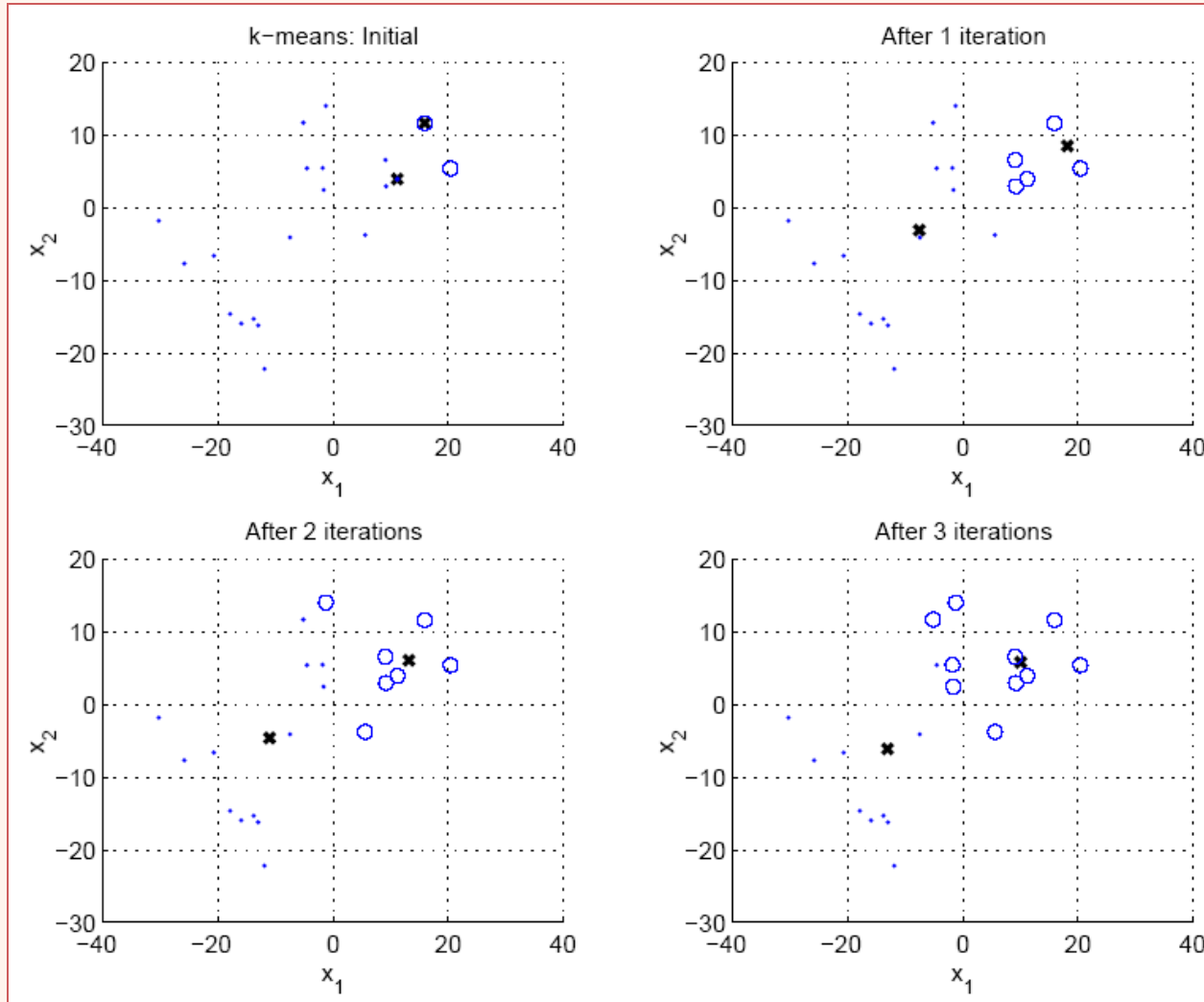
For all $\mathbf{m}_i, i = 1, \dots, k$

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

Until \mathbf{m}_i converge

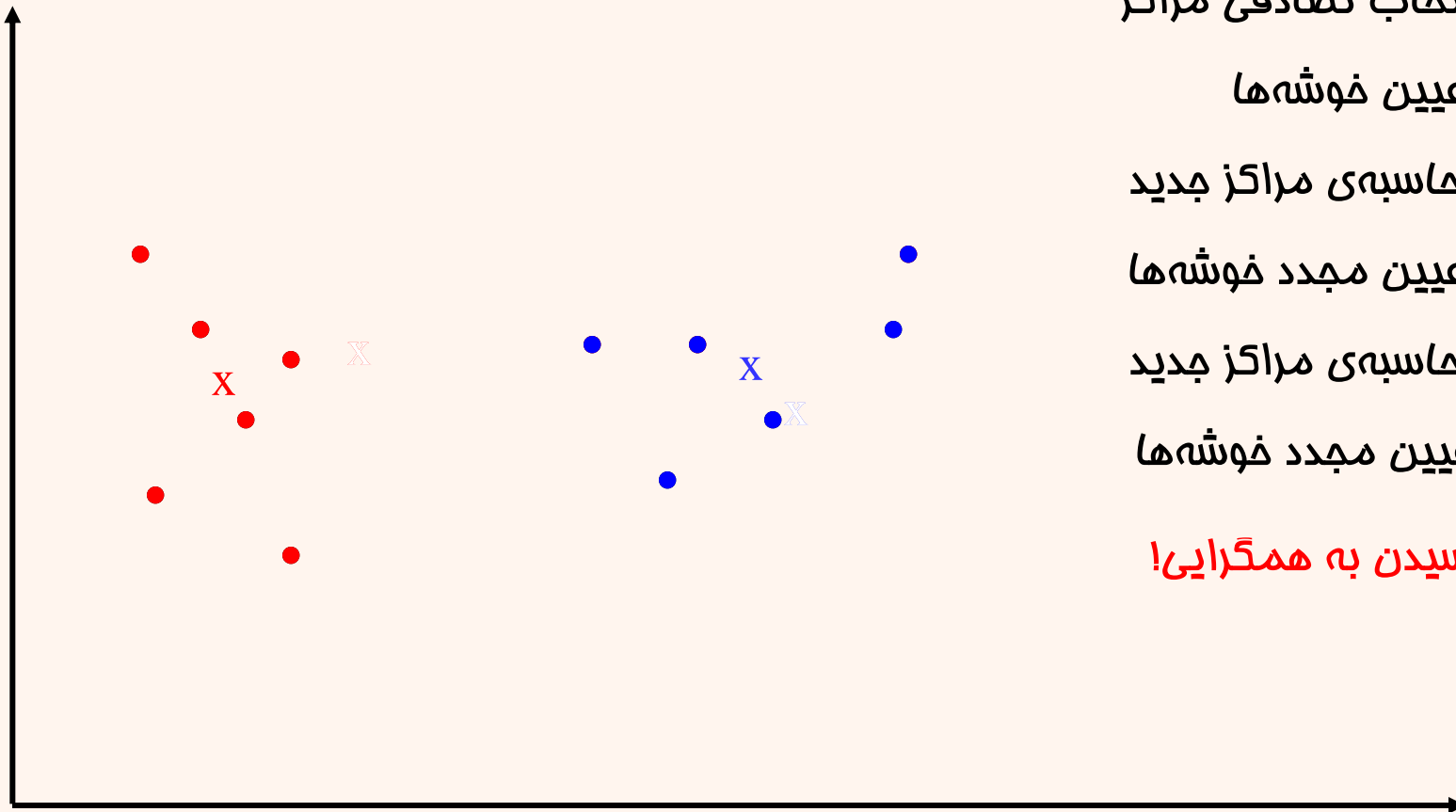


در صورتی که بخش‌بندی داده‌ها تخریب نکند، الگوریتم به پایان رسیده است. ری ماشین



K=2

مثال ۲



انتخاب تصادفی مراکز

تعیین خوشه‌ها

محاسبه‌ی مراکز جدید

تعیین مجدد خوشه‌ها

محاسبه‌ی مراکز جدید

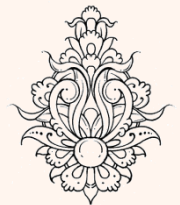
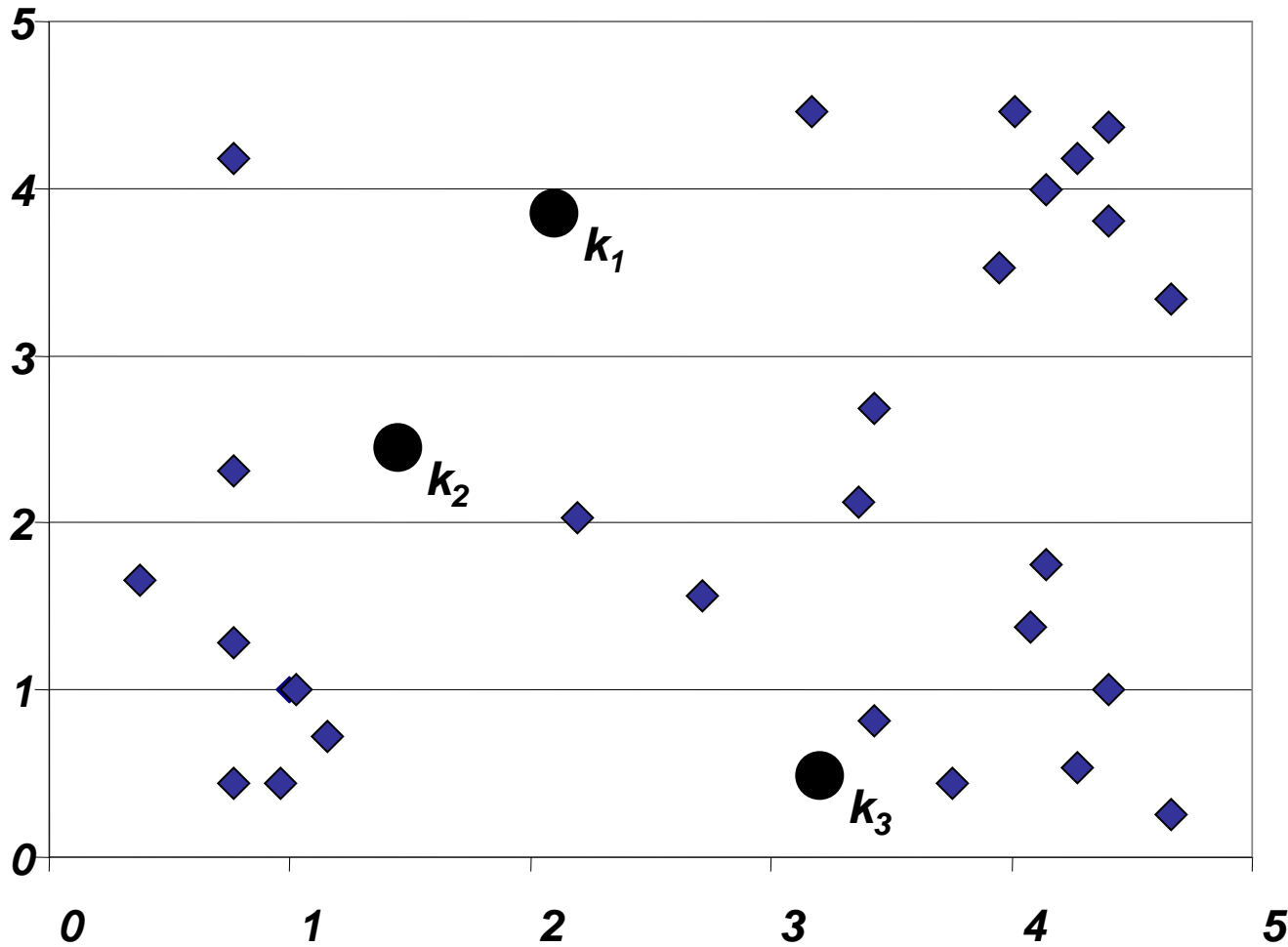
تعیین مجدد خوشه‌ها

رسیدن به همگرایی!



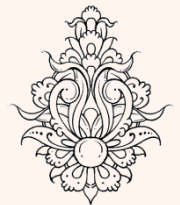
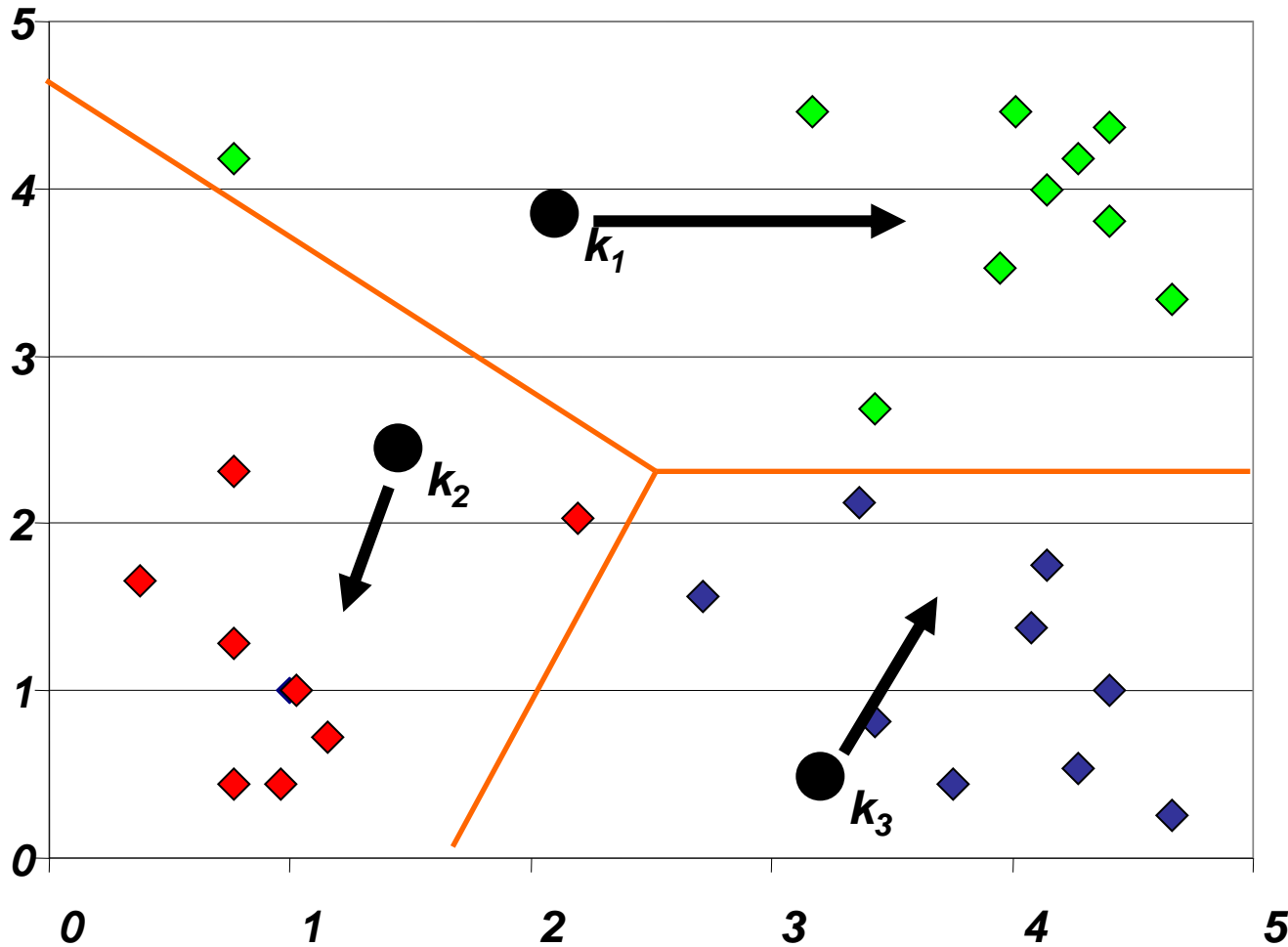
K=3

مثال ۳



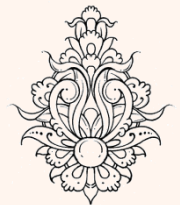
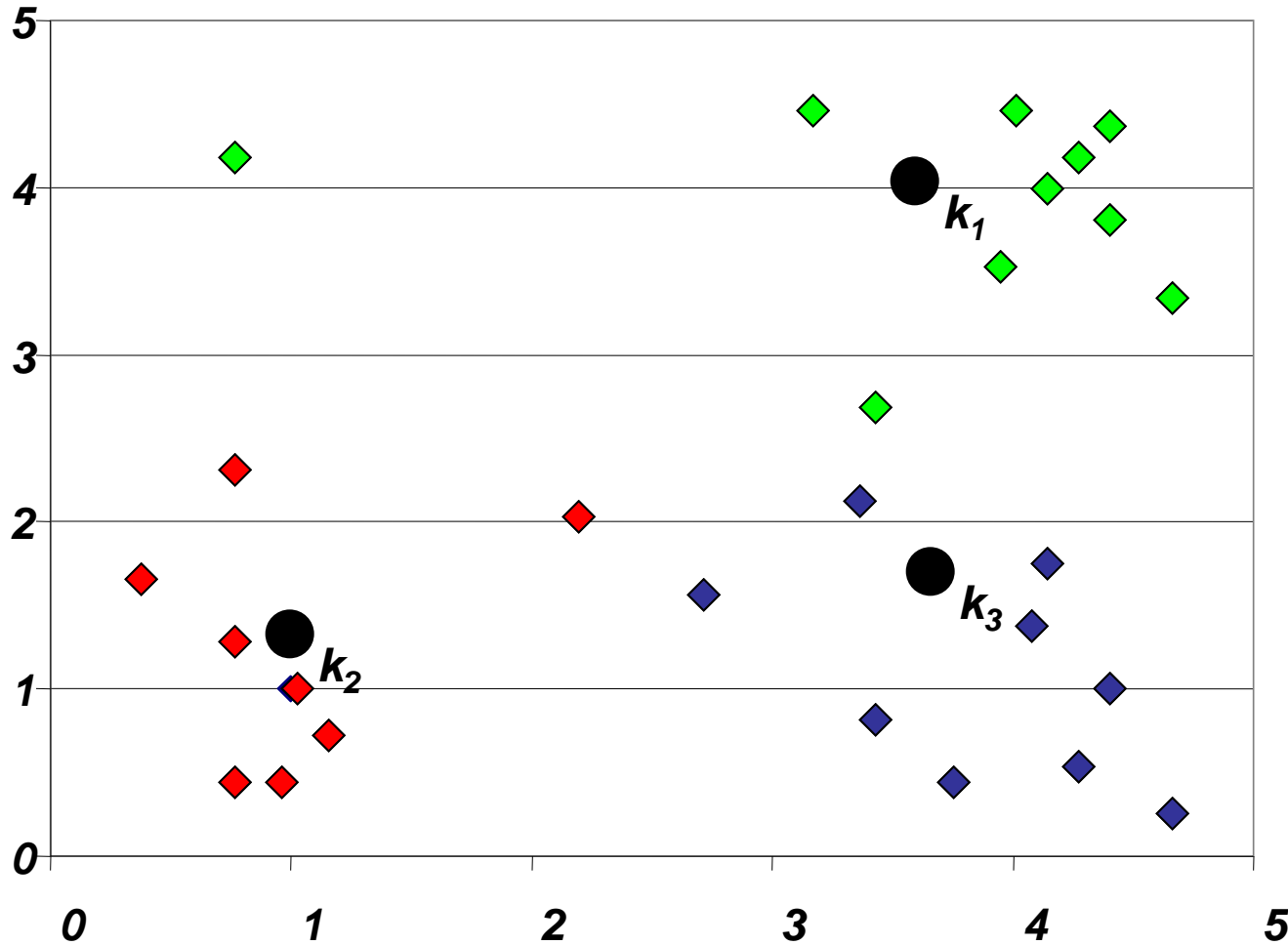
K=3

مثال ۳ (ادامه...)



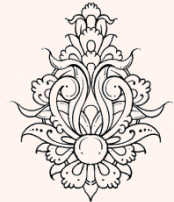
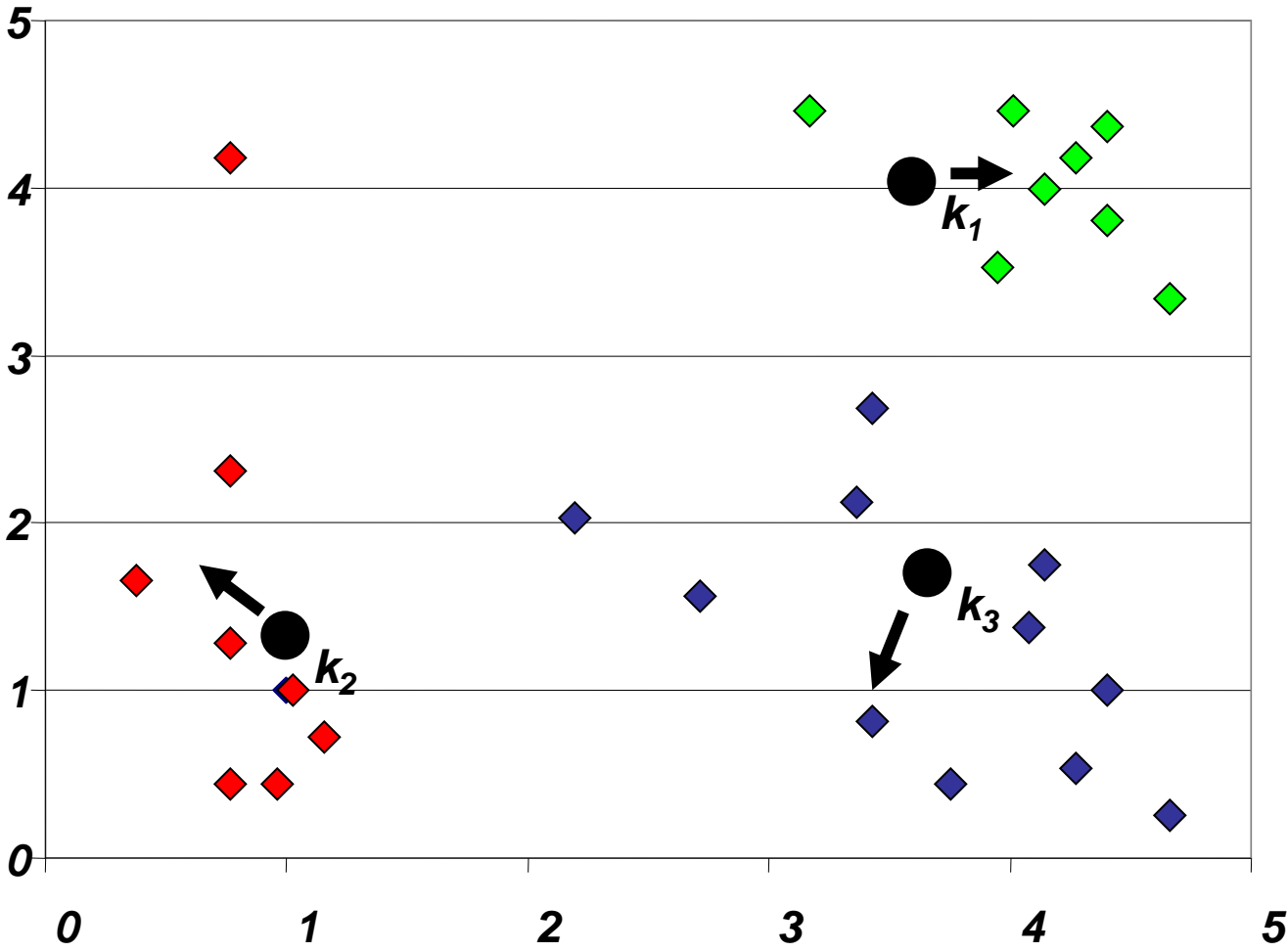
$K=3$

مثال ۳ (ادامه...)



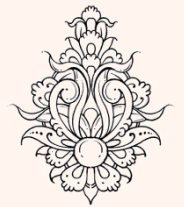
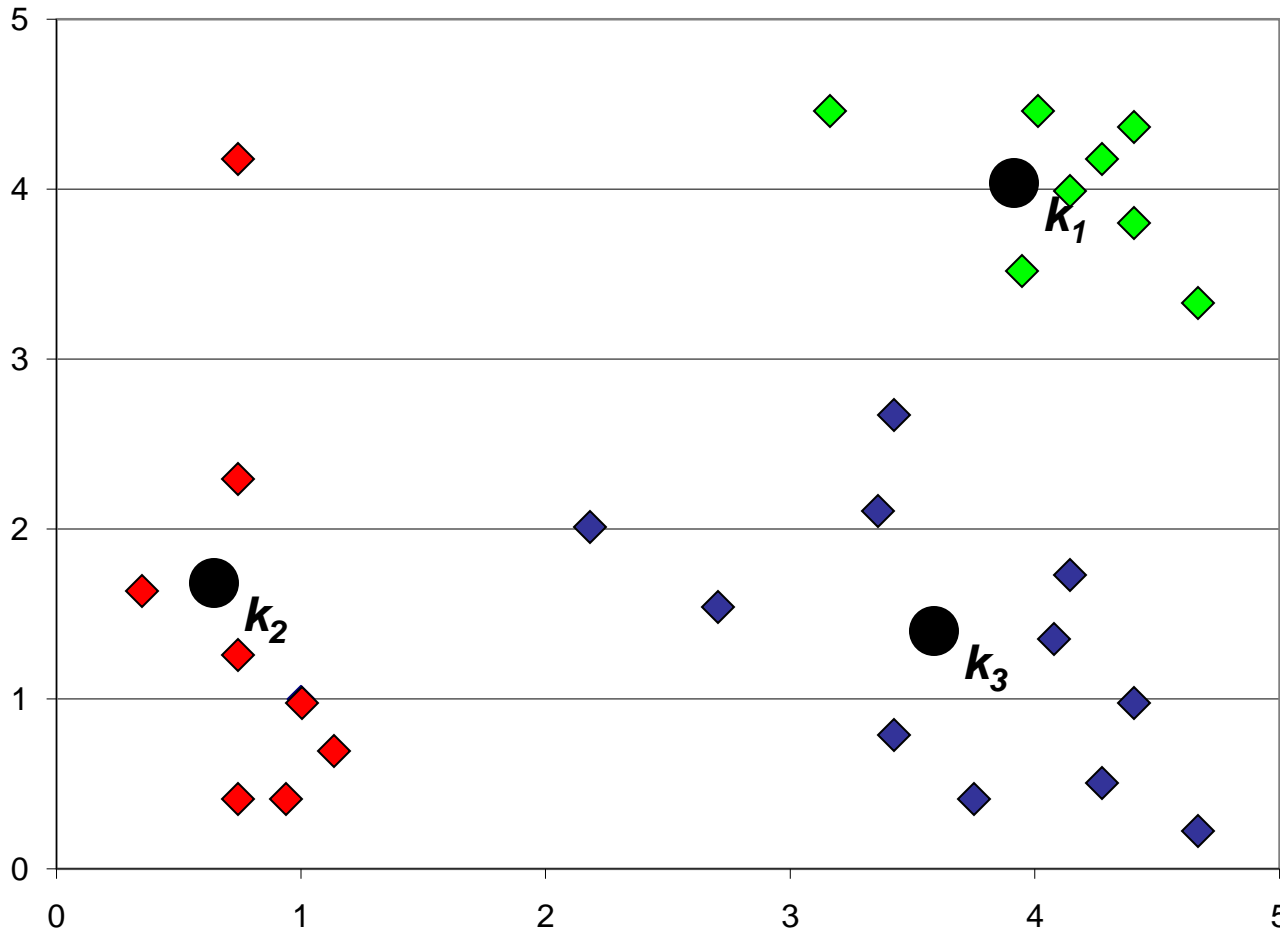
K=3

مثال ۳ (ادامه...)



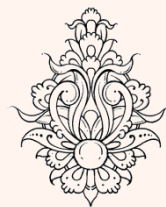
$K=3$

مثال ۳ (ادامه...)



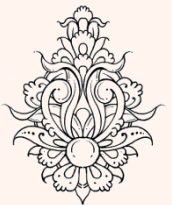
مشکلات k-means

- 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

$K = 2$



$K = 3$



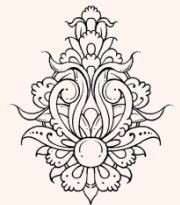
$K = 10$



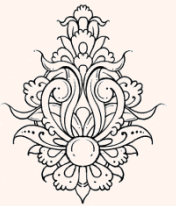
Original image



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



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



Expectation-Maximization (EM)

• در صورتی که بخواهیم با استفاده از MLE پارامترهای یک

مدل ترکیبی را تخمین بزنیم:

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

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

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

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

• 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.

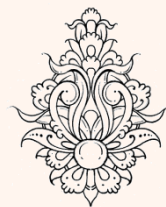
- هدف این الگوریتم یافتن پارامترهایی (Φ) است که احتمال رخداد متغیرهای قابل مشاهده ($L(\Phi | X)$) را بیشینه کند.

Incomplete likelihood

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

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

Complete likelihood



دو گام این الگوریتم

E-step

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

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

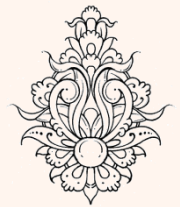
M-step

$$\text{E-step: } Q(\Phi | \Phi^l) = E[\mathcal{L}_c(\Phi | \mathcal{X}, Z) | \mathcal{X}, \Phi^l]$$

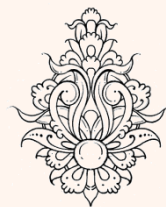
$$\text{M-step: } \Phi^{l+1} = \arg \max_{\Phi} Q(\Phi | \Phi^l)$$

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

$$\mathcal{L}(\Phi^{l+1} | \mathcal{X}) \geq \mathcal{L}(\Phi^l | \mathcal{X})$$



- در مثال ترکیب توزیع‌ها، «متغیرهای پنهان» مشخص می‌کنند کدام نمونه به کدام خوشه تعلق دارد.
- در صورتی که تعلق هر نمونه‌ی 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 \mathbf{z}^t متغیر پنهان در این مسأله است.

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

• $z_i^t = 1$ اگر \mathbf{x}^t به خوشه i -ام تعلق داشته باشد.

– این متغیر دارای توزیع برنولی تصمیم یافته است.

– در واقع شبیه به r_i^t در حالت بانظر است.

اگر \mathbf{z} ها مشخص باشند، مانند حالت «بانظارت» است:

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

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



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

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

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

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

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

$$\mathcal{L}_c(\Phi | \mathcal{X}, \mathcal{Z}) = \log \prod_t p(\mathbf{x}^t, \mathbf{z}^t | \Phi)$$

$$= \sum_t \log p(\mathbf{x}^t, \mathbf{z}^t | \Phi)$$

$$P(G_i) = \pi_i$$

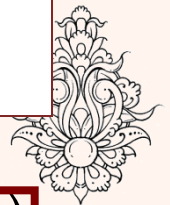
$$= \sum_t \log P(\mathbf{z}^t | \Phi) + \log p(\mathbf{x}^t | \mathbf{z}^t, \Phi)$$

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

$$= \sum_t \sum_i z_i^t [\log \pi_i + \log p_i(\mathbf{x}^t | \Phi)]$$

$$p_i(\mathbf{x}^t) = p(\mathbf{x}^t | G_i)$$

$$Q(\Phi | \Phi^l) = E[\mathcal{L}_c(\Phi | \mathcal{X}, \mathcal{Z}) | \mathcal{X}, \Phi^l]$$



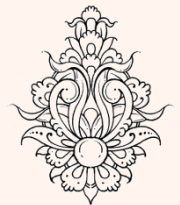
EM in Gaussian Mixtures (cnt'd...)

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

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

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

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



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

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

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

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

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

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

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

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

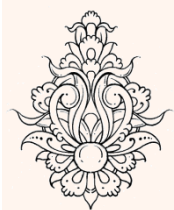
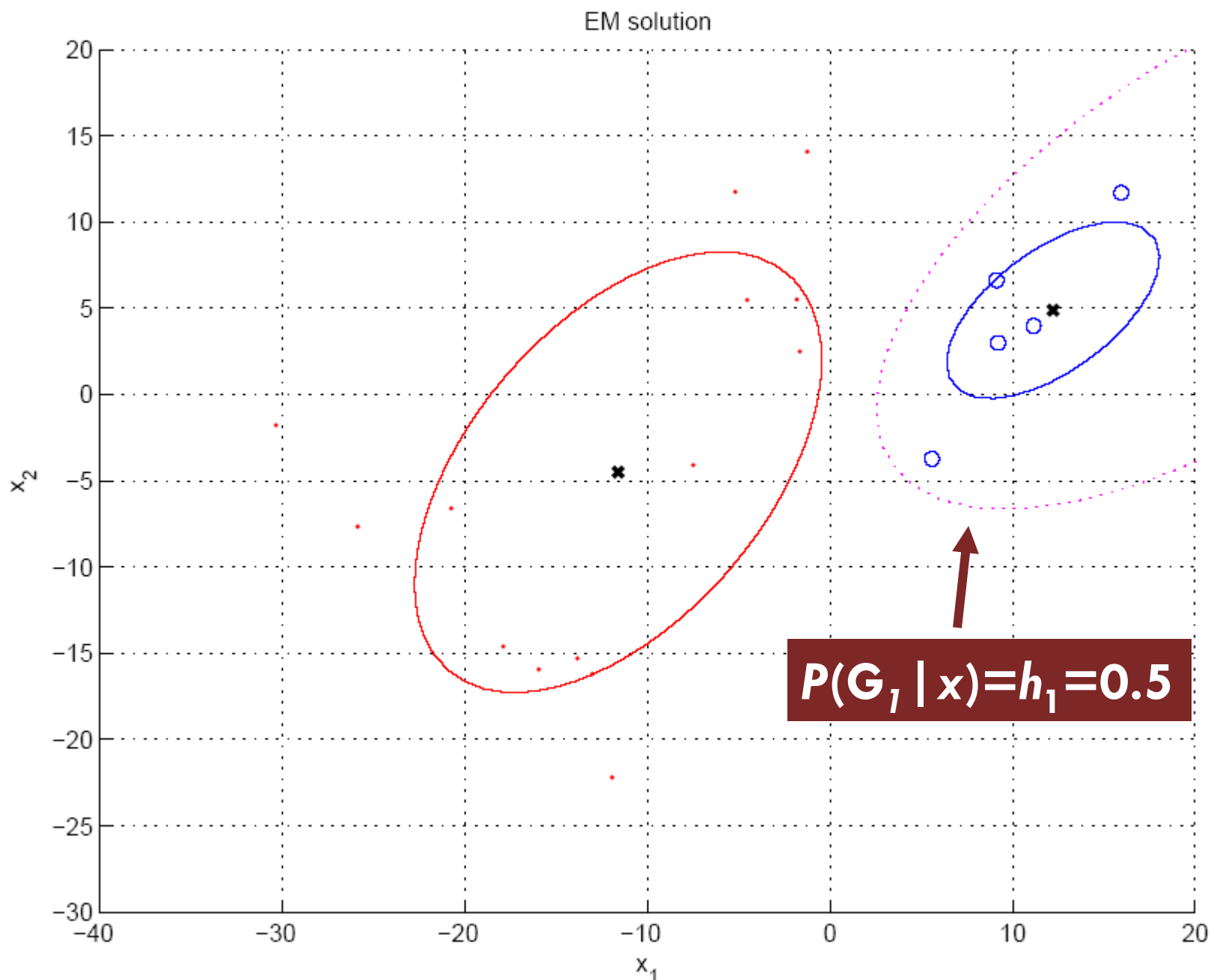
$$P(C_i) = \frac{\sum_t r_i^t}{N} \quad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

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

Soft label

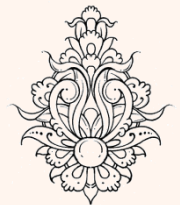
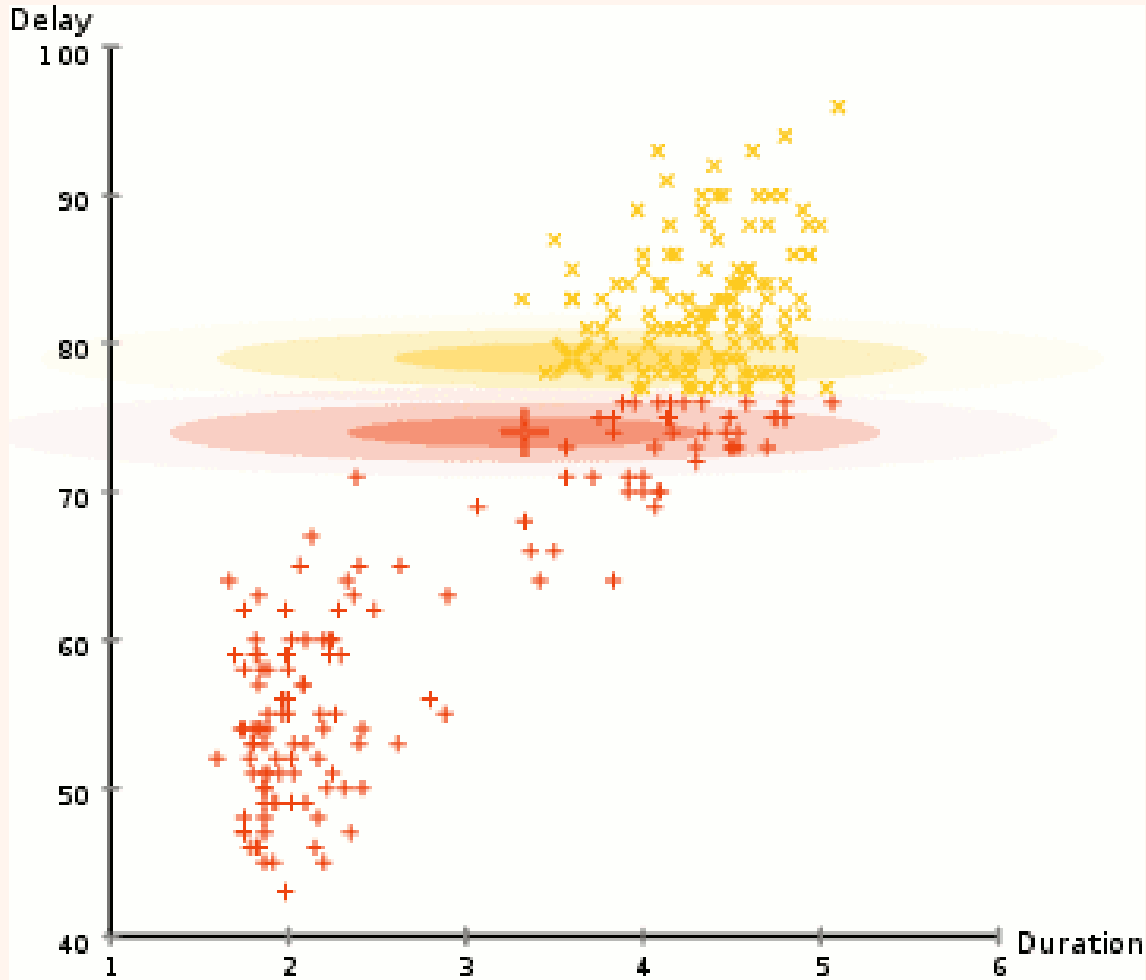


مثال



تراشگاه
سپهر
بهشتی

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

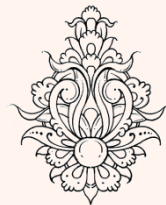
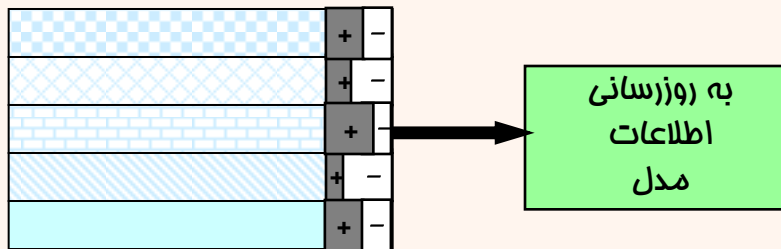
□	+	-
◇	+	-
■	+	-
▨	+	-
■	+	-



Expectation-Maximization (EM)

M Step:

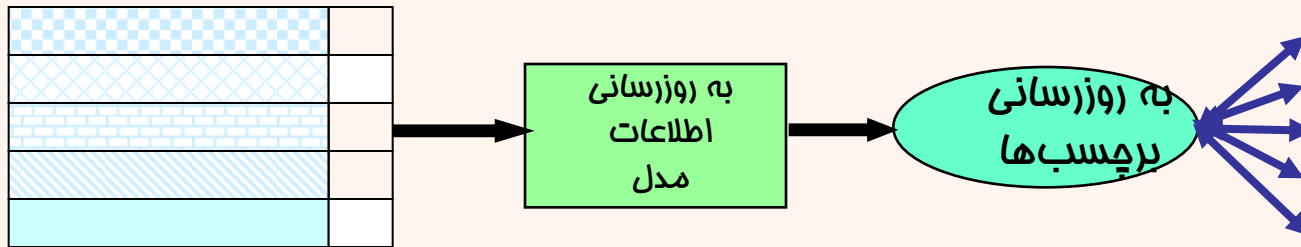
- براساس برچسب‌های نسبت داده شده، اطلاعات توزیع‌ها به روز می‌شود:



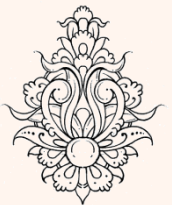
Expectation-Maximization (EM)

E Step:

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

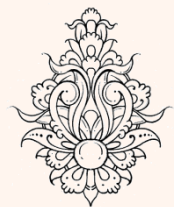
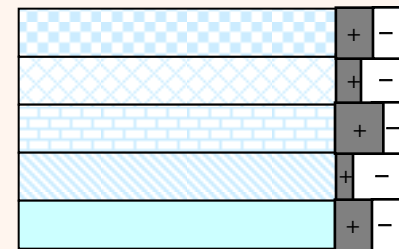
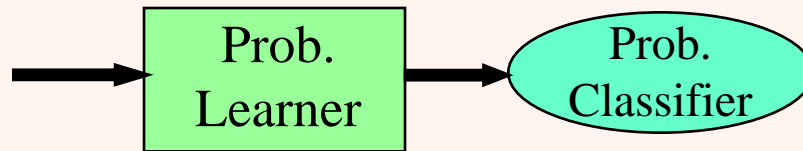


+	-
+	-
+	-
+	-
+	-



Expectation-Maximization (EM)

M Step:



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

جمع‌بندی

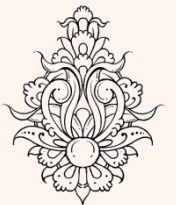
- پارامترهای اولیه مقداردهی می‌شود: Φ^0 initialize

- تا زمانی رسیدن به همگرایی تکرار کن:

- گام E: $P(Z | X, \Phi^l)$ را تقریب بزن

- گام M: $\Phi^{l+1} = \arg \max_{\Phi} Q(\Phi | \Phi^l)$

- 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 $P(G_i)$ تخمین زده شده و الگوریتم EM آغاز می‌شود.



Expectation-Maximization (EM)

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

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

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

- 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

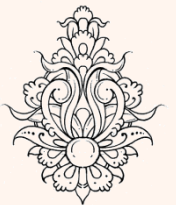
- در صورتی که برای همه‌ی خوشه‌ها کواریانس یکسانی در نظر بگیریم، با رابطه‌ی ساده‌تری مواجه خواهیم شد:

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

- در صورتی که توزیع‌های هر خوشه، ناهمبسته بوده و واریانس یکسانی داشته باشند:
- بسیار شبیه به k-means است، با این تفاوت که که برچسب‌ها در این جا بین صفر و یک هستند.

$$\min_{\mathbf{m}, S} \sum_t \sum_i h_i^t \frac{\|x^t - m_i\|^2}{s^2}$$

$$h_i^t = \frac{\exp[-(1/2s^2)\|x^t - m_i\|^2]}{\sum_j \exp[-(1/2s^2)\|x^t - m_j\|^2]}$$



انتخاب مدل (ادامه...)

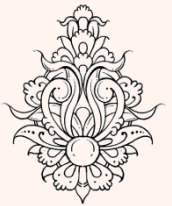
- در نظر گرفتن ماتریس کواریانس یکسان، موجب نادیده گرفتن شکل واقعی خوشه‌ها می‌شود.
 - در نظر گرفتن ماتریس کواریانس قطری با توجه به نادیده گرفتن همبستگی‌ها، به طریق اولی ساختار واقعی را نادیده می‌گیرد.
- به عنوان راه حل، پیش از خوشه‌بندی می‌توان از روش‌های کاهش ابعاد (PCA/FA) بهره برد.

$$p(\mathbf{x}_t | G_i) = \mathcal{N}(\mathbf{m}_i, \mathbf{V}_i \mathbf{V}_i^T + \psi_i)$$



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.



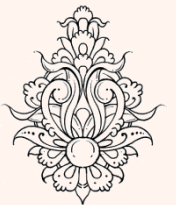
خوشه‌بندی به عنوان پیش‌پردازش

- 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)، بدین ترتیب ممکن با افزایش ابعاد هم مواجه شویم.
- در کاهش ابعاد هم‌ی 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 بدون برچسب در فرآیند آموزش بهره برد.
- در نهایت در دسته‌بندی با ترکیبی از ترکیب‌ها مواجه خواهیم بود:

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

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

Mixture of Mixtures



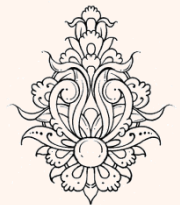
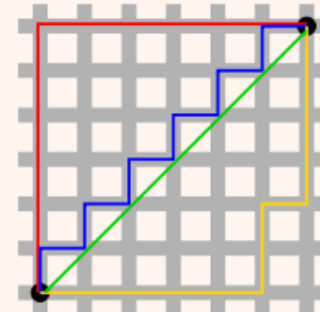
- در k-means هدف مینیمم کردن خطای بازسازی است.
- در «خوشه‌بندی سلسله‌مراتبی»، تنها شباهت بین نمونه‌ها در نظر گرفته می‌شود.
- افزون بر فاصله‌ی اقلیدسی معیارهای دیگری نیز در نظر گرفته می‌شوند:

Minkowski

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

City-block distance

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



Hierarchical Agglomerative Clustering (HAC)

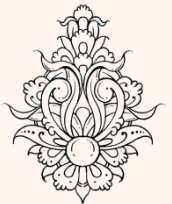
- با N خوشه کار آغاز می‌شود؛ هر خوشه شامل یک نمونه می‌باشد.
- خوشه‌های نزدیک به هم در هر تکرار با هم ادغام می‌شوند.
- برای انتخاب گروه‌های نزدیک، سه معیار مورد استفاده قرار می‌گیرد:

$$d(G_i, G_j) = \min_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s) \quad \text{Single-link}$$

$$d(G_i, G_j) = \max_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s) \quad \text{Complete-link}$$

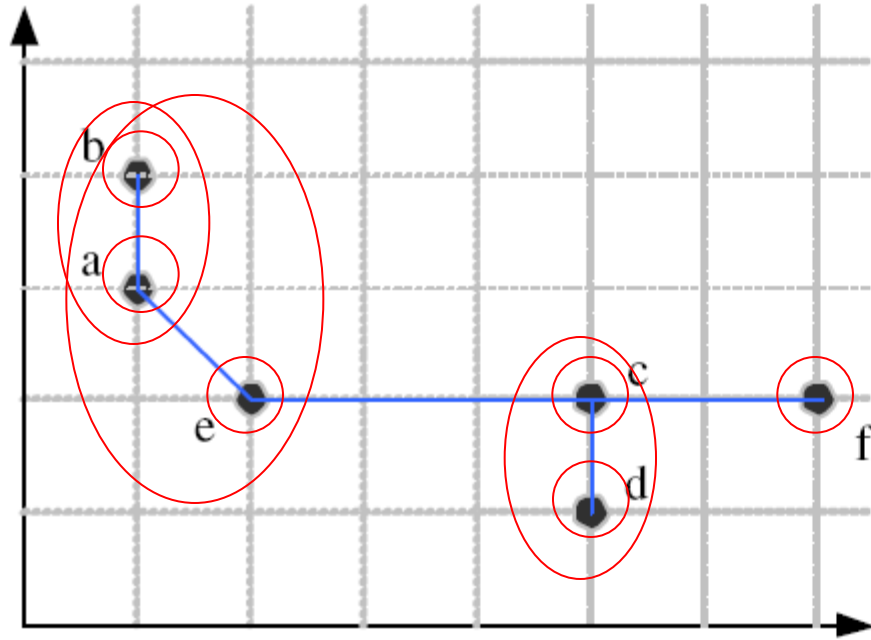
$$d(G_i, G_j) = \text{ave}_{\mathbf{x}^r \in G_i, \mathbf{x}^s \in G_j} d(\mathbf{x}^r, \mathbf{x}^s) \quad \text{Average-link, centroid}$$

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



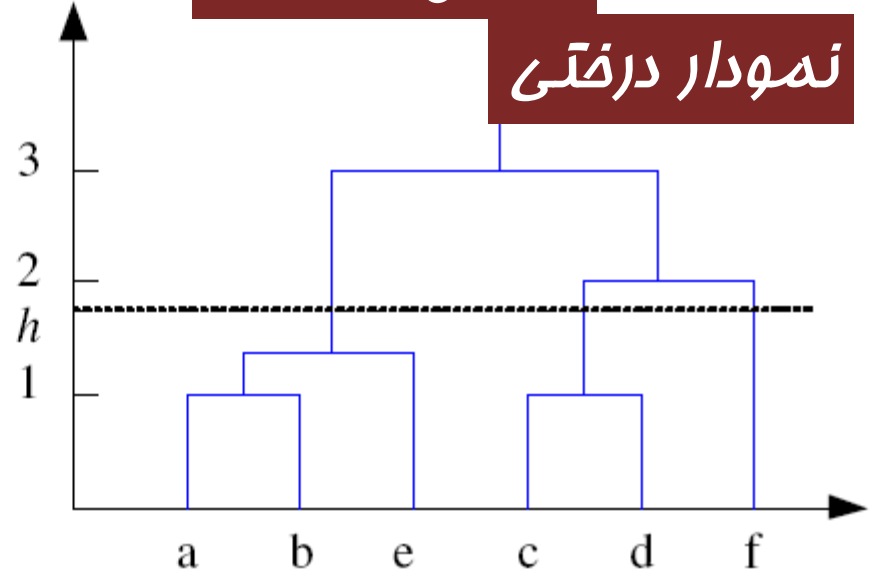
Agglomerative Clustering

فوشه بندی سلسله مراتبی (ادامه...)



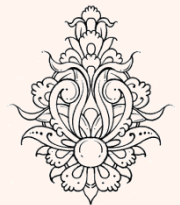
Dendrogram

نمودار درختی



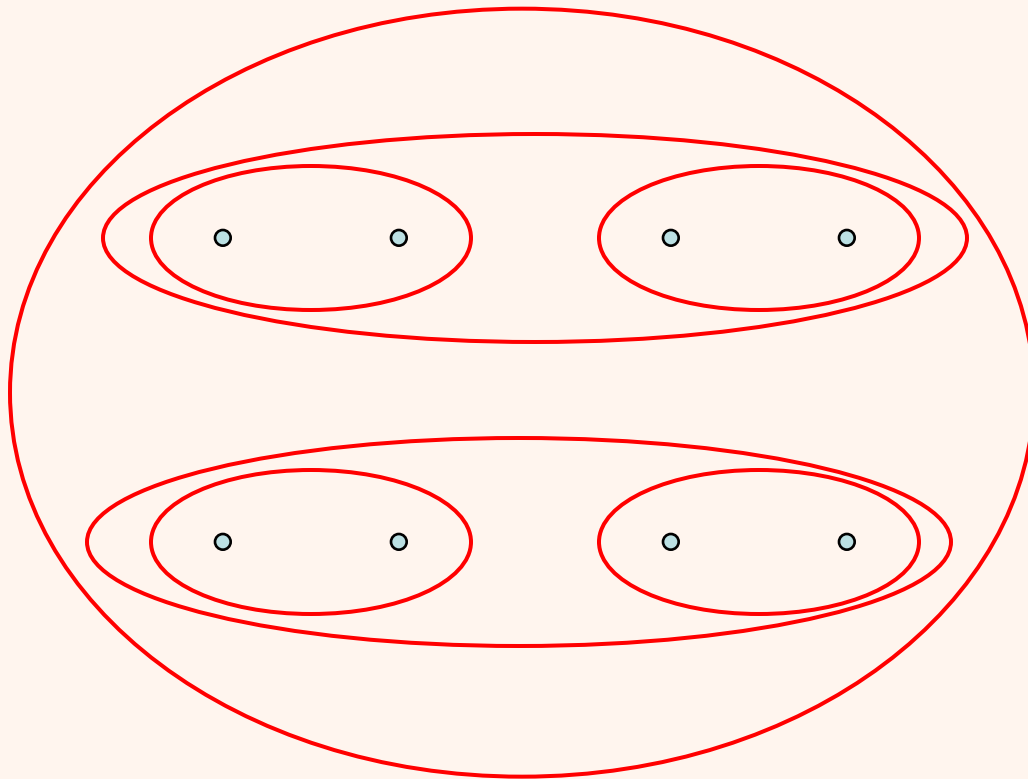
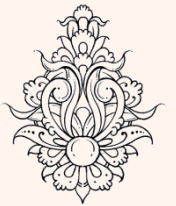
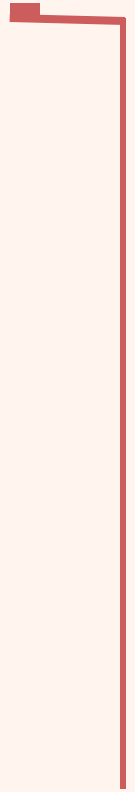
Divisive Clustering

در این شیوه به صورت عکس عمل می شود؛ از یک فوشه کار آغاز شده و فوشه ها در هر تکرار به فوشه های کوچک تر تقسیم می شوند تا زمانی که هر فوشه شامل یک نمونه باشد.



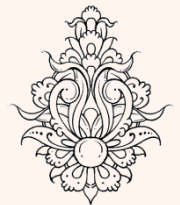
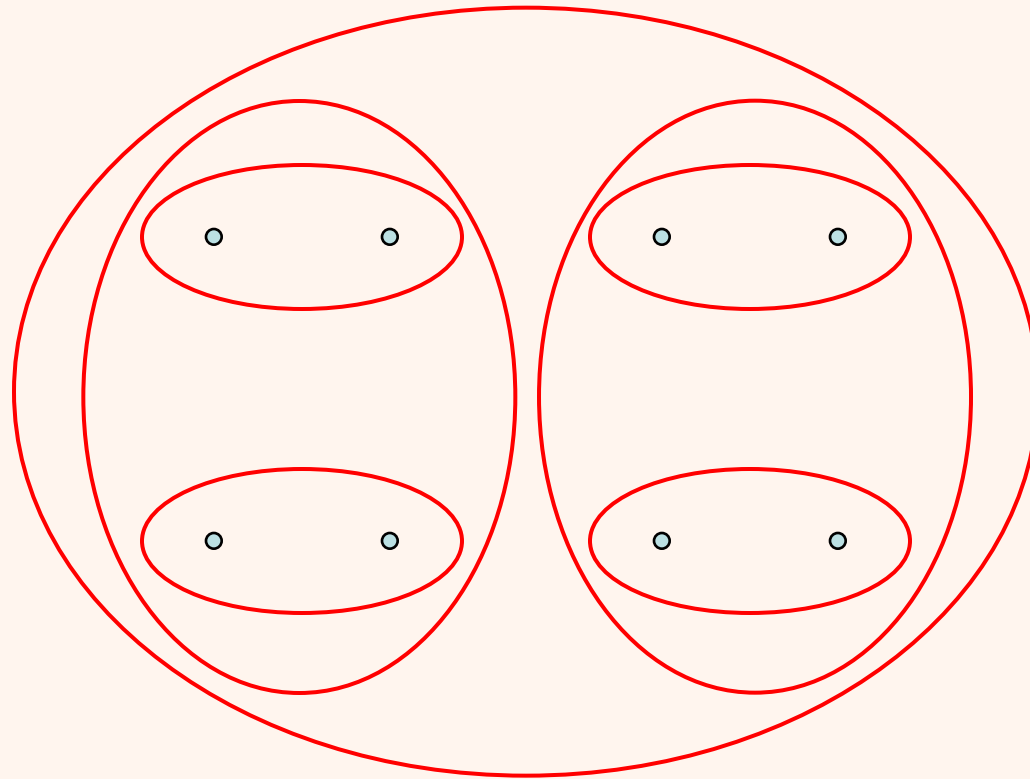
single-link clustering

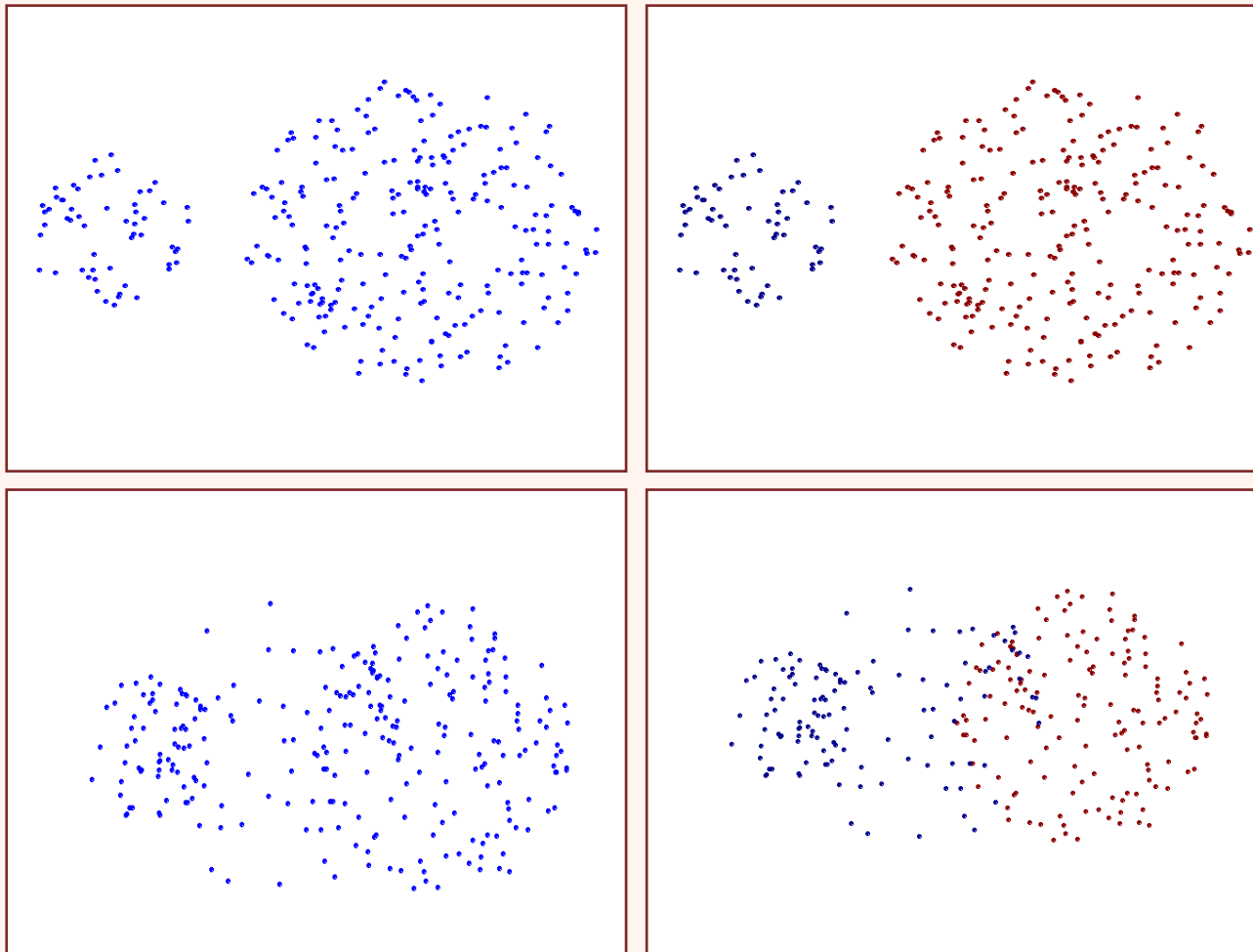
مثال



complete-link clustering

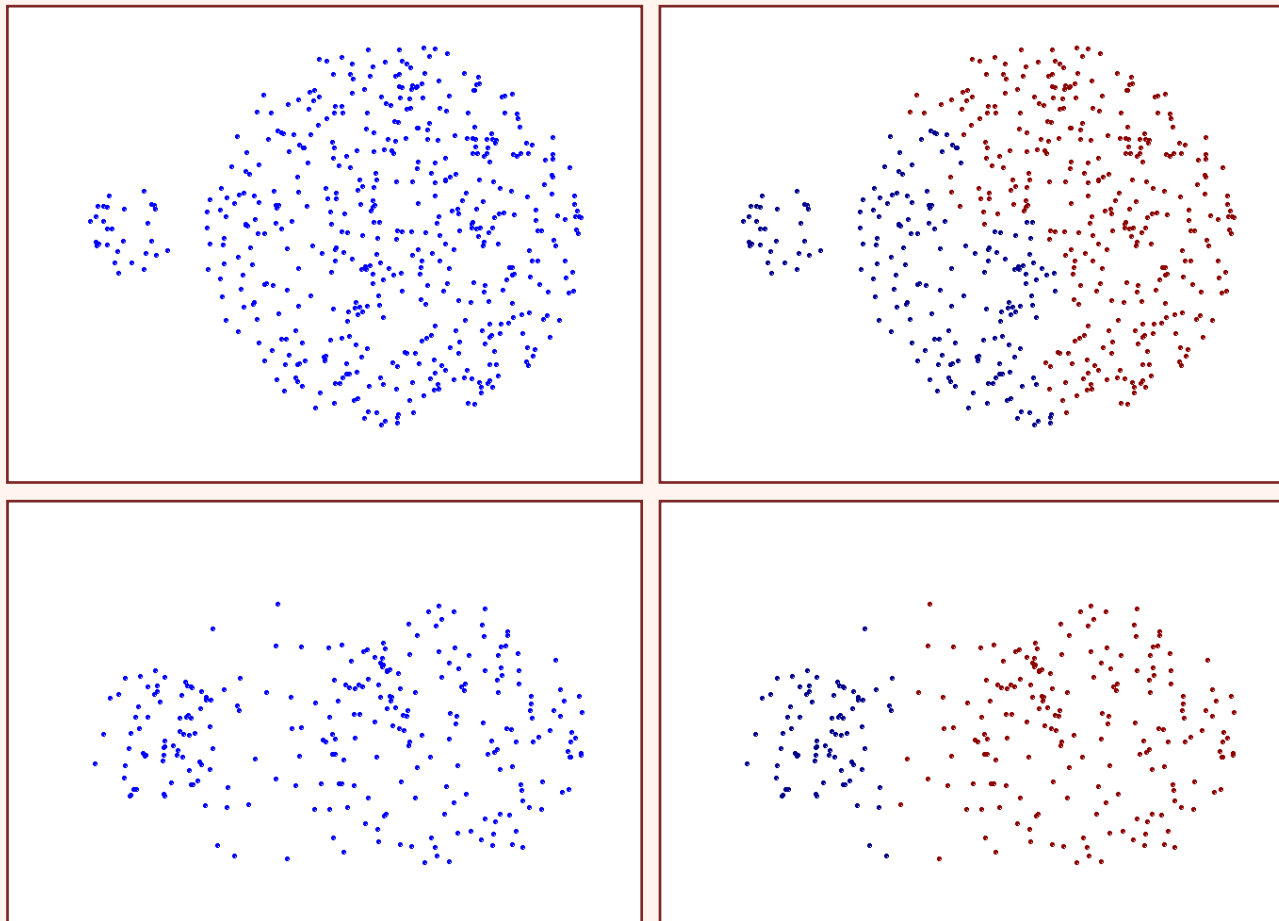
مثال





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



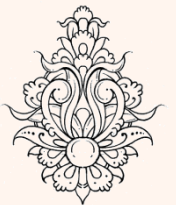


فوشه‌های بزرگ را می‌شکند، فوشه‌ها با قطر یکسان تولید می‌کند.
فوشه‌های کوچک را با فوشه‌های بزرگ ادغام می‌کند.

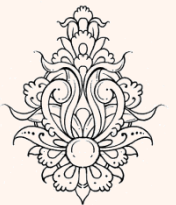


انتخاب تعداد خوشه‌ها

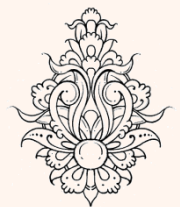
- در برخی کاربردها، با توجه به نیاز k مشخص می‌شود، مانند color quantization
- استفاده از PCA و رسم داده‌ها در دو بعد می‌تواند ساختار داده‌ها را تا حدی مشخص کرده و انتخاب مناسب k کمک کند.
- استفاده از روش‌های افزایشی (leader-cluster)
- در برخی کاربردها بعد از انجام خوشه‌بندی، به صورت دستی می‌توان مناسب 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.

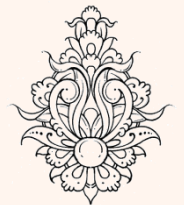
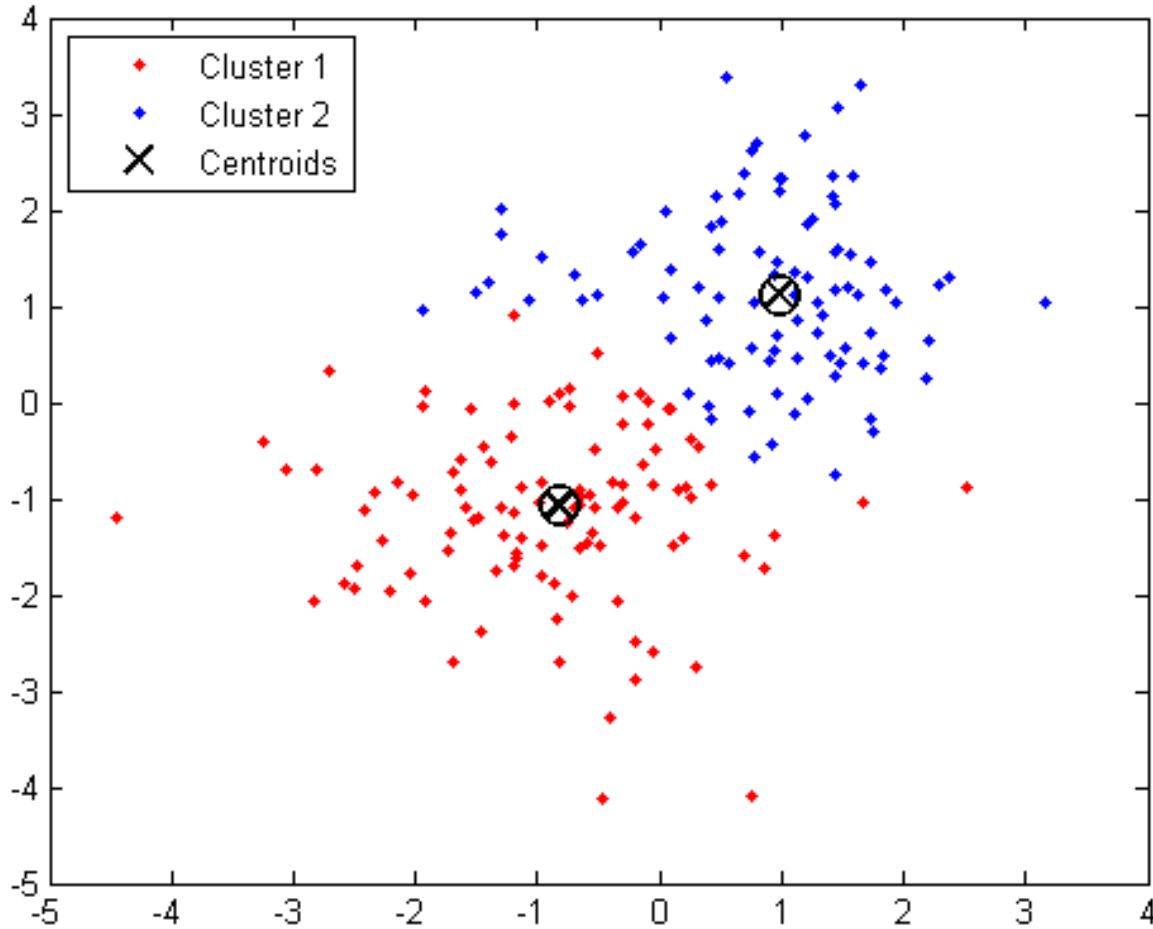


- در دسته‌بندی بانظارت، تهیه‌ی داده‌ی برچسب‌خورده معمولاً هزینه‌ی بالای تحمیل می‌کند.
 - در یادگیری نیمه‌نظارتی، از داده‌های برچسب‌نخورده برای افزایش کارایی کمک گرفته می‌شود.
- به عنوان مثال شیوه‌ی نیمه‌نظارتی EM به این صورت است که داده‌های برچسب‌خورده توسط ناظر با برچسب‌های اصلی خود در تخمین شرکت می‌کنند.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




```
clear all, clc;
X = [randn(100,2)+ones(100,2);...
     randn(100,2)-ones(100,2)];
opts = statset('Display','iter');
[idx,ctr] = 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(ctr(:,1),ctr(:,2),'kx',...
     'MarkerSize',12,'LineWidth',2)
plot(ctr(:,1),ctr(:,2),'ko',...
     'MarkerSize',12,'LineWidth',2)
legend('Cluster 1','Cluster 2','Centroids',...
      'Location','NW')
```





```
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)
figure;
dendrogram(Z)
T = cluster(Z,'maxclust',2)
```

