# باز شناسی آماری الگو

نشست دهم، طبقه بندی با یادگیری
بدون مربی

# فهرست مطالب

- مقدمه
- جدایی پذیری و توزیع های مختلط
- تخمین بیشترین احتمال(شباهت) ، کاربرد در حالت توزیع گوسی
- الگوریتم k-means
- یادگیری بدون مربی Bayesian
- توصیف داده ها و خوشه بندی
- معیار ها برای خوشه بندی
- خوشه بندی سلسله مراتبی
- مسئله تعداد خوشه ها
- قابل قبول بودن نتایج

# مقدمه

- در یادگیری با مربی علاوه بر داشتن ورودی ها، نشانه label هر نمونه نیز وجود داشت. Supervised learning
- در یادگیری بدون مربی دیگر اطلاعات مربوط به نشانه یا label نمونه ها در دسترس نیستند. unsupervised learning
- در اینکه اطلاعات با مربی برای استفاده آسانتر هستند شکی نیست اما صرفنظر کردن از اطلاعات بدون مربی، در همه موارد جایز نیست. مانند:
  - یافتن خط زمینه
  - جدا سازی تصویر مورد نظر از زمینه

# چرایادگیری بدون مربی؟

- ممکن است اساسا طبقه بندی اولیه وجود نداشته باشد (مانند طبقه بندی اعضای طبیعی و غیر طبیعی)
- ممکن است خصوصیات ( و نیز طبقه) نمونه ها در طول زمان تغییر کند
- ممکن است میزان اعتماد به انتساب های داده شده به نمونه های یادگیری پیش از یادگیری با مربی، کافی نباشد (عدم قطعیت برچسب ها)
- ممکن است هدف بازنمایی نمونه ها به روش دیگر باشد.
- ممکن است بخواهیم دانش بیشتری از آنچه داریم در مورد مسئله استخراج کنیم.

# رویکرد های اصلی:

- وقتی که تعدادی نمونه در دست داریم ولی برای آنها برچسبی وجود ندارد، این نمونه ها چه اطلاعات دیگری ممکن است داشته باشند که به حل مسئله طبقه بندی کمک کند؟
  - دسته بندی بر اساس توزیع کلاس ها:
    - اگر توزیع کلاس ها را به نحوی از روی نمونه های یادگیری بیابیم، (شبیه مرحله تست در کلاسه بندی) می توان بر اساس پیش فرض هائی مانند قانون Bayesبه برچسب گذاری مبادرت کرد.
  - دسته بندی مستقیم ( بر اساس نزدیکی و دوری نمونه ها)
    - اگر بتوان این پیش فرض را پذیرفت که اعضای هر کلاس،(یا زیر گروه هایی از هر کلاس) در فضا مشخصه ها، به هم نزدیک هستند، می توان از روی موقعیت نمونه ها در فضا و توزیع آن، به نحوی به کلاس بندی مبادرت کرد.

# Mixture Densities & Identifiably

- فرض اصلی:
  - شکل چگالی های توزیع برای کلاس ها معلوم است ولی پارامتر های آن شکل مجهول است
- پیش فرض ها:
  - نمونه ها مربوط به تعداد معلومی از کلاس ها هستند.
  - برای هر کلاس احتمال پیشین معلوم است: $P(\omega_j)$ for $(j = 1, \ldots, c)$
  - برای هر کلاس شکل تابع توزیع معلوم ولی پارامتر های تابع توزیع مجهول است $P(x \mid \omega_j, \theta_j)$ $(j = 1, \ldots, c)$

# Mixture Densities & Identifiably

- توابع احتمال پیشین، مانند وزن درتابع  توزیع مرکب عمل کرده اند:

$$P(x \mid \theta) = \sum_{j=1}^{c} \overbrace{P(x \mid \omega_j, \theta_j)}^{component\ densities} . \underbrace{P(\omega_j)}_{mixing\ parameters}$$

$$where \quad \theta = (\theta_1, \theta_2, ..., \theta_c)^t$$

- تابع توزیع فوق را تابع توزیع مختلط (مخلوط) می گویند.
- **هدف : یافتن $\theta$ با استفاده از داده ها**
- با پیدا شدن بردار مزبور می توان توزیع مستقل هر کلاس را پیدا نمود. سپس به مراحل دیگر رفت.

# توابع قابل شناسایی

- تابع چگالی $P(x \mid \theta)$ را قابل شناسایی identifiable گویند اگر $\theta \neq \theta'$ برای هر نمونه، منجر به تابع توزیع دیگری شود: $P(x \mid \theta) \neq P(x \mid \theta')$

مثال: برای x باینری : $P(x \mid \theta)$ زیر :

$$P(x \mid \theta) = \frac{1}{2}\theta_1^x (1-\theta_1)^{1-x} + \frac{1}{2}\theta_2^x (1-\theta_2)^{1-x}$$

$$= \begin{cases} \frac{1}{2}(\theta_1 + \theta_2) \text{ if } x = 1 \\ 1 - \frac{1}{2}(\theta_1 + \theta_2) \text{ if } x = 0 \end{cases}$$

برای مقادیر زیر

$$P(x = 1 \mid \theta) = 0.6 \Rightarrow P(x = 0 \mid \theta) = 0.4$$

$$\theta_1 + \theta_2 = 1.2$$

این توزیع مختلط قابل جدا سازی نیست

# توابع قابل شناسایی...

- در توزیع های گسسته، اگر تعداد مولفه ها ی زیادی در توزیع مخلوط داشته باشیم، ممکن است تعداد مجهول ها از معادلات بیشتر شود و در اینصورت جدایی پذیری identifiably به یک مسئله اصلی تبدیل خواهد شد.
- اگر چه می توان نشان داد که در توزیع های نرمال، عموما جدایی سازی ممکن است، حتی در حالتی که احتمال های پیشین مساویند، ممکن است بطور منحصر بفرد نتوان توزیع کلاس ها را از هم جدا نمود

$$P(x|\theta) = \frac{P(\omega_1)}{\sqrt{2\pi}} exp\left[-\frac{1}{2}(x-\theta_1)^2\right] + \frac{P(\omega_2)}{\sqrt{2\pi}} exp\left[-\frac{1}{2}(x-\theta_2)^2\right]$$

مانند حالتیکه $\theta = (\theta_1, \theta_2)$ and $\theta = (\theta_2, \theta_1)$

- گرچه این مسئله یک چالش است، ما همواره ممکن بودن جدا سازی را بعنوان پیش فرض پذیرفته ایم.

# تخمین ML

$D = \{x_1, \ldots, x_n\}$ و داده ها از هم مستقل هستند. ($\theta$ is fixed but unknown)

$$p(x \mid \theta) = \sum_{j=1}^{c} p(x \mid \omega_j, \theta_j) P(\omega_j)$$

$$\hat{\theta} = \arg\max_{\theta} p(D \mid \theta) \text{ with } p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta)$$

$$(l = \sum_{k=1}^{n} \ln p(x_k \mid \theta))$$

برای یافتن مقدار فوق، یک روش متداول استفاده از محاسبه مشتق (گرادیان) و نقطه تغییر علامت (گذر از صفر) آن است.

$$\nabla_{\theta_i} l = \sum_{k=1}^{n} P(\omega_i \mid x_k, \theta) \nabla_{\theta_i} \ln p(x_k \mid \omega_i, \theta_i)$$

$\hat{\theta}_i$

# ریشه های گرادیان و جواب مسئله

$$\sum_{k=1}^{n} P(\omega_i \mid x_k, \hat{\theta}) \nabla_{\theta_i} \ln p(x_k \mid \omega_i, \hat{\theta}_i) = 0 \quad (\mathrm{i} = 1, \ldots, \mathrm{c}) \qquad (a)$$

$$\hat{P}(\omega_i) = \frac{1}{n} \sum \hat{P}(\omega_i \mid x_k, \hat{\theta})$$

$$\text{and} \sum_{k=1}^{n} \hat{P}(\omega_i \mid x_k, \hat{\theta}) \nabla_{\theta_i} \ln p(x_k \mid \omega_i, \hat{\theta}_i) = 0$$

$$\text{where :} \quad \hat{P}(\omega_i \mid x_k, \hat{\theta}) = \frac{p(x_k \mid \omega_i, \hat{\theta}_i) \hat{P}(\omega_i)}{\sum_{j=1}^{c} p(x_k \mid \omega_j, \hat{\theta}_j) \hat{P}(\omega_j)}$$

## کاربرد در حالتی که توزیع ها نرمال هستند:

اگر برای هر یک از کلاس ها : $p(x \mid \omega_i, \theta_i) \sim N(\mu_i, \Sigma_i)$

- برای این مسئله حالت های مختلفی را می توان در نظر گرفت:

| Case | $\mu_i$ | $\Sigma_i$ | $P(\omega_i)$ | c |
|---|---|---|---|---|
| 1 | ? | X | x | x |
| 2 | ? | ? | ? | x |
| 3 | ? | ? | ? | ? |

## حالت اول: یافتن مقادیر متوسط

$$\mu_i = \theta_i \quad \forall\, i = 1, \ldots, c$$

- با تخمین ML $\mu = (\mu_i)$:

$$\ln p(x|\omega_i, \mu_i) = -\ln\left[(2\pi)^{d/2}|\Sigma_i|^{1/2}\right] - \frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)$$

$$\hat{\mu}_i = \frac{\sum_{k=1}^{n} P(\omega_i \mid x_k, \hat{\mu})\, x_k}{\sum_{k=1}^{n} P(\omega_i \mid x_k, \hat{\mu})} \qquad (1)$$

- چنانکه دیده می شود در این تخمین رابطه ای بازگشتی وجود دارد نه صریح

# حالت اول: یافتن مقادیر متوسط

- با داشتن حدس اولیه ای از آن $\hat{\mu}_i(0)$ و روش های بازگشتی تکراری، می توان به مقادیر دقیق تر رسید.

$$\hat{\mu}_i(j+1) = \frac{\sum_{k=1}^{n} P(\omega_i \mid x_k, \hat{\mu}(j))x_k}{\sum_{k=1}^{n} P(\omega_i \mid x_k, \hat{\mu}(j))}$$

## حالت دو کلاسه

$$p(x \mid \mu_1, \mu_2) = \frac{1}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu_1)^2\right] + \frac{2}{3\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-\mu_2)^2\right]$$

میدانیم دو خوشه از داده ها داریم:

$$\mu_1 = -2,\ \mu_2 = 2$$

اگر مثلا از هر کلاس ۲۵ نمونه را تولید کنیم و تابع لگاریتم شباهت را بر اساس فرمول زیر بدست آوریم، بیشترین احتمال در مقادیر زیر بدست می آید:

$$l(\mu_1, \mu_2) = \sum_{k=1}^{n} \ln p(x_k \mid \mu_1, \mu_2)$$

$$\hat{\mu}_1 = -2.130 \text{ and } \hat{\mu}_2 = 1.668$$

# تجزیه و تحلیل نتیجه:

*مقادیر بدست آمده نزدیک جواب های اصلی است:*

$$\mu_1 = -2 \text{ and } \mu_2 = +2$$

می توان پیک های دیگری نیز دید:

$$\hat{\mu}_1 = 2.085 \text{ and } \hat{\mu}_2 = -1.257$$

در حالتیکه تابع قابل شناسایی نباشد جواب یکتا وجود ندارد

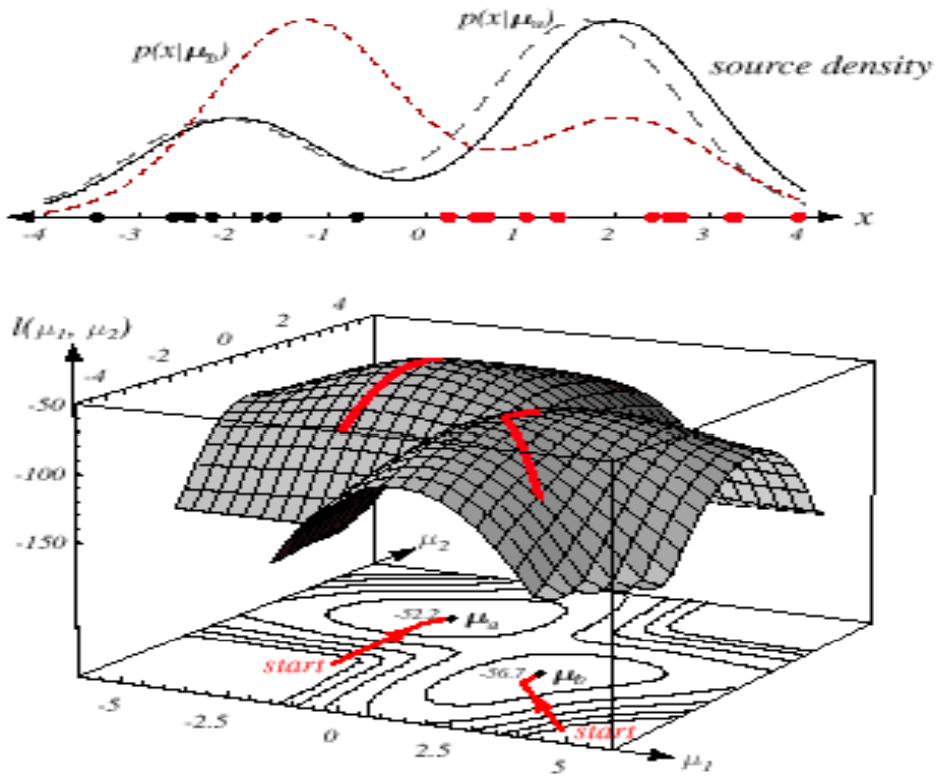



**FIGURE 10.1.** (Above) The source mixture density used to generate sample data, and two maximum-likelihood estimates based on the data in the table. (Bottom) Log-likelihood of a mixture model consisting of two univariate Gaussians as a function of their means, for the data in the table. Trajectories for the iterative maximum-likelihood estimation of the means of a two-Gaussian mixture model based on the data are shown as red lines. Two local optima (with log-likelihoods −52.2 and −56.7) correspond to the two density estimates shown above. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

## حالت دوم: همه پارامتر ها مجهول هستند:

- محدودیتی نیز برای ماتریس کوواریانس وجود ندارد
- $p(x \mid \mu, \sigma^2)$

$$p(x \mid \mu, \sigma^2) = \frac{1}{2\sqrt{2\pi}.\sigma} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] + \frac{1}{2\sqrt{2\pi}} exp\left[-\frac{1}{2}x^2\right]$$

$$p(x_1 \mid \mu, \sigma^2) = \frac{1}{2\sqrt{2\pi}\ \sigma} + \frac{1}{2\sqrt{2\pi}} \exp\left[-\frac{1}{2}x_1^2\right]$$

# ادامه..

با فرض $\mu = x_1$ :

$$p(x_k \mid \mu, \sigma^2) \geq \frac{1}{2\sqrt{2\pi}} exp\left[-\frac{1}{2}x_k^2\right]$$

برای بقیه نمونه ها:

$$p(x_1,...,x_n \mid \mu, \sigma^2) \geq \underbrace{\left\{\frac{1}{\sigma} + exp\left[-\frac{1}{2}x_1^2\right]\right\}}_{\left(\begin{matrix}this\ term \to \infty \\ \sigma \to 0\end{matrix}\right)} \frac{1}{(2\sqrt{2\pi})^n} exp\left[-\frac{1}{2}\sum_{k=2}^{n} x_k^2\right]$$

در این حالت تابع شباهت بزرگ بوده و حل ML به سوی تکین شدن پیش می رود.

# الگوریتم تکرار

Iterative scheme

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} \hat{P}(\omega_i \mid x_k, \hat{\theta})$$

$$\hat{\mu}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i \mid x_k, \hat{\theta})\, x_k}{\sum_{k=1}^{n} \hat{P}(\omega_i \mid x_k, \hat{\theta})}$$

$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{n} \hat{P}(\omega_i \mid x_k, \hat{\theta})(x_k - \hat{\mu}_i)(x_k - \hat{\mu}_i)^t}{\sum_{k=1}^{n} \hat{P}(\omega_i \mid x_k, \hat{\theta})}$$

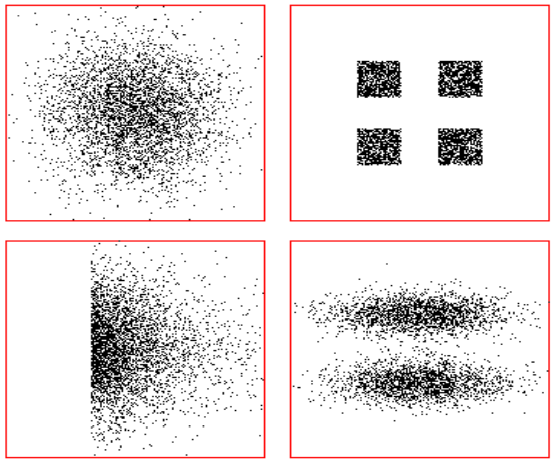
$$\hat{P}(\omega_i \mid x_k, \hat{\theta}) = \frac{|\Sigma_i|^{-1/2} \exp\left[-\frac{1}{2}(x_k - \hat{\mu}_i)^t \hat{\Sigma}_i^{-1}(x_k - \hat{\mu}_i)\right]\hat{P}(\omega_i)}{\sum_{j=1}^{c} |\hat{\Sigma}_j|^{-1/2} \exp\left[-\frac{1}{2}(x_k - \hat{\mu}_j)^t \hat{\Sigma}_j^{-1}(x_k - \hat{\mu}_j)\right]\hat{P}(\omega_j)}$$

## بحث

- ایده استفاده از مخلوط توزیع های نرمال می تواند برای گروه بزرگی از مسائل تقریب خوبی ارائه کند.
- در این حالت ها، می توان از ایده روشهای پارامتری برای تقریب تابع توزیع مرکب استفاده نمود.
- اما اگر اطلاعات ما کافی نباشد، پیش فرض های پارامتری بی معنی بوده و در واقع ما به جای اینکه ساختار داده ها را بیابیم، مبادرت به این کرده این که داده ها را به جبر به ساختار مورد نظر خود بخورانیم.
- بدیهی است در این شرایط، ممکن است روش های غیر پارامتری به جواب های بهتری منجر شوند.

# Data Clustering

- چگونگی ساختار داده ها در مورد نمونه های چند بعدی در فرآیند خوشه بندی موثر است.
- اگر ما به نحوی بدانیم که نمونه ها از فرآیند توزیع خاصی تولید شده اند، این امر در ارائه مدل فشرده ای ازداده ها موثر است و البته این شرط برای ارائه مدل کافی است.
  - در صورتیکه پیش فرض اتخاذ شده ( در مورد فرآیند توزیع داده ها) درست نباشد، بدیهی است که این امر به برخی اشتباهات منجر خواهد شد.
  - خوشه بندی بدون این پیش فرض ها چگون خواهد بود؟



# خوشه بندی

- همچون مسئله  طراحی کلاسیفایر در یادگیری با مربی، ممکن است انتساب نمونه به یک کلاس یا زیر کلاس ضرورتا نیازی به یافتن تابع توزیع نداشته باشد.
- اگر هدف پیدا کردن زیر-کلاس ها باشد، استفاده از ایده خوشه بندی، به معنی یافتن **تجمع هایی از داده ها یا نمونه هایی که  در فضای مشخصه ها، به هم نزدیک و از گروه ها یا تجمع های دیگر دور هستند**، می تواند مفید باشد.
- به عنوان یک مثال ساده، معروفترین الگوریتم خوشه بندی موسوم به k مقدار متوسط را بررسی میکنیم

# خوشه بندی C-Means

- هدف: یافتن C بردار مقدار متوسط: $\mu_1, \mu_2, \ldots, \mu_c$
- با استفاده از مفهوم فاصله Mahalanobis

$$(x_k - \hat{\mu}_i)^t \hat{\Sigma}_i^{-1} (x_k - \hat{\mu}_i) \text{ by the squared Euclidean distance } \|x_k - \hat{\mu}_i\|^2$$

- Find the mean $\hat{\mu}_m$ earest to $x_k$ and approximate

$$\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_c$$

$$\hat{P}(\omega_i \mid x_k, \hat{\theta}) \qquad \hat{P}(\omega_i \mid x_k, \theta) \cong \begin{cases} 1 & \text{if } i = m \\ 0 & \text{otherwise} \end{cases}$$

# شبه کد الگوریتم C-means

- اگر تعداد کلاس ها معلوم باشد الگوریتم C-means، عبارت خواهد بود از:

```
Begin
initialize n, c, μ1, μ2, …, μc(randomly
do classify n samples according
    nearest μi                          to
          re compute μi
        until no change in μi
              return μ1, μ2, …, μc
                                    End
```

# نمایش الگوریتم c-means بر اساس ایده تپه نوردی

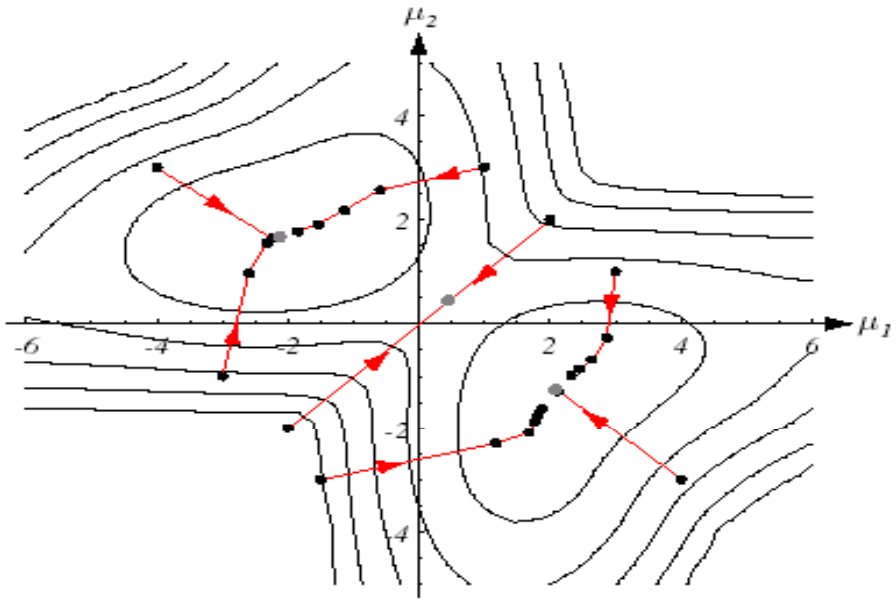



**FIGURE 10.2.** The *k*-means clustering procedure is a form of stochastic hill climbing in the log-likelihood function. The contours represent equal log-likelihood values for the one-dimensional data in Fig. 10.1. The dots indicate parameter values after different iterations of the *k*-means algorithm. Six of the starting points shown lead to local maxima, whereas two (i.e., $\mu_1(0) = \mu_2(0)$) lead to a saddle point near $\boldsymbol{\mu} = \mathbf{0}$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

# Similarity measures

- یک سوال اولیه در این ایده آن است که شباهت نمونه های موجود در یک خوشه را چگونه می توان بطور کمی ارزیابی کرد؟
- در شق معمول از مسئله:
  - چگونه شباهت بین نمونه ها قابل ارزیابی کمی است؟
  - چگونه کیفیت یک خوشه ( یا خوشه بندی) قابل ارزیابی کمی است؟
- بدیهی برای پاسخ گوئی به هر یک از این سوال ها، به یک شاخص یا متریک نیاز است.
- با داشتن چنان شاخصی، می توان فرض کرد که فاصله بین نمونه ها در داخل یک خوشه (بر اساس آن متریک) باید به حد کافی کوچکتر از فاصله نمونه های داخل و خارج خوشه باشد.
- به بیان دیگر می توان خوشه را با یک مرکزیت و یک شعاع مشخص نمود.

# فاصله اقلیدسی

- دیدیم که نمونه ها را با بردار نشان می دهیم.
- فاصله اقلیدسی معمولترین متریک است برای سنجش فاصله بردار هاست.
- اندازه معیار معرف اعضای یک خوشه آن خواهد بود که (مثلا) فاصله آنها از هم ( و یا از نماینده خوشه) از فاصله معین $d_0$ کوچکتر باشد.
- خوشه ای که با این معیار تعریف می شود، نسبت به چرخش نامتغیر است اما به برخی تبدیلات دیگرکه روی فواصل تاثیر می گذارند، خیر

# تاثیر معیار فاصله در خوشه ها

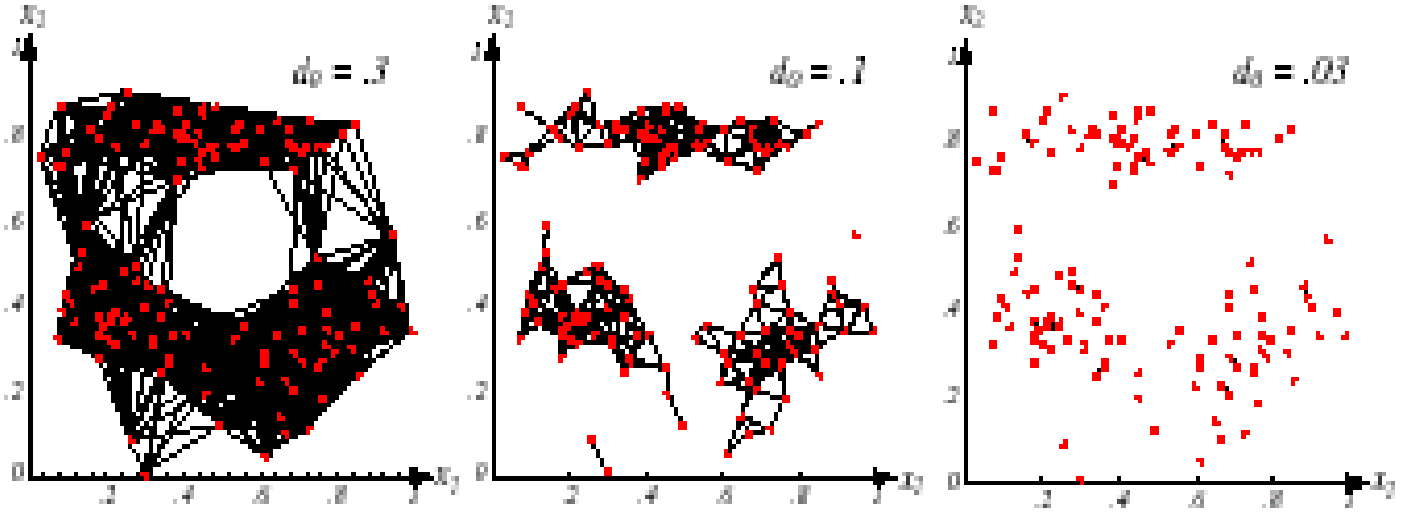



FIGURE 10.7. The distance threshold affects the number and size of clusters in similarity based clustering methods. For three different values of distance $d_0$, lines are drawn between points closer than $d_0$—the smaller the value of $d_0$, the smaller and more numerous the clusters. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

# تاثیر معیار فاصله

- برای اینکه به نامتغیر بودن خوشه دست یابی پیدا شود، ممکن است از انواع نرمالیزه کردن ها (بشکلی که همگی دارای مقدار متوسط صفر و واریانس یک باشند) و یا موافه های اساسی ( برای مقابله با تغییرات ناشی از چرخش) استفاده نمود.
- متریک مینکووسکی Minkowsky یک متریک معمول دیگر است که به وفور مورد استفاده قرار می گیرد.

$$d(\mathbf{x},\mathbf{x}') = \left( \sum_{k=1}^{d} |x_k - x_k'|^q \right)^{1/q}$$

با تغییر پارامتر q متریک های مختلفی به دست می آید:

$q = 1 \Rightarrow$ Manhattan or city block metric

$q = 2 \Rightarrow$ Euclidean metric

# تاثیر معیار فاصله

- ممکن است از توابع یا عملگر های  غیر متریک برای تعریف شباهت بین دو برداراستفاده شود. این توابع معمولا خاصیت تعویض پذیری داشته و مقدار بزرگتر در خروجی آنها نشان دهنده شباهت بیشتر است ( بر عکس معیار فاصله) مثلا ضرب داخلی در بردار

$$s(\mathbf{x},\mathbf{x}') = \frac{\mathbf{x}^t\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}$$

- در حالتی که مشخصه ها باینری باشند:

$$s(\mathbf{x},\mathbf{x}') = \frac{\mathbf{x}^t\mathbf{x}'}{d}$$

$$s(\mathbf{x},\mathbf{x}') = \frac{\mathbf{x}^t\mathbf{x}'}{\mathbf{x}^t\mathbf{x} + \mathbf{x}'^t\,\mathbf{x}' + \mathbf{x}^t\mathbf{x}'}$$

Tanimoto distance

# Clustering as optimization

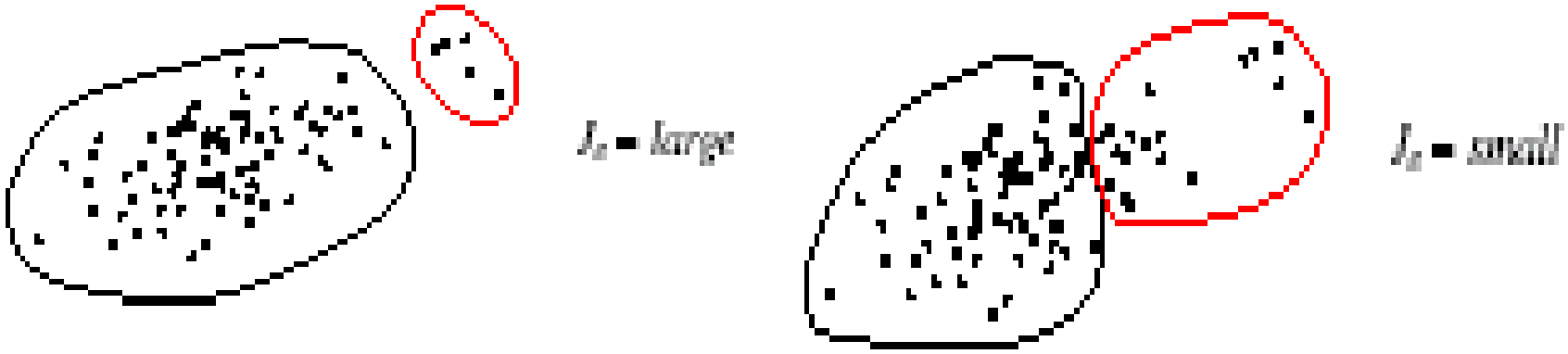
- حال به سوال دوم بپردازیم، چگونه با داشتن معیار شباهت، کیفیت خوشه بندی را بالا ببریم؟
- می توان آن خوشه بندی را انتخاب کرد که دارای کیفیت بهتری است. به تعبیر دیگر فرآیند خوشه بندی به یک مسئله بهینه سازی تبدیل می شود.
- هر مسئله بهینه سازی خود به یک معیار بهینه سازی و یک روش جستجو برای یافتن بهینه نیاز دارد.
  - معیار های معمول بهینه سازی
    - مجموع مربعات خطا
    - معیار پراکندگی
  - برای مجموع مربعات خطا

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|$$

# معیار مجموع مربعات

- در این معیار خوشه ها با میانگین اعضایشان معرفی می شوند $\mathbf{m}_i$ و *فاصله هر نمونه* $\mathbf{x} - \mathbf{m}_i$ *یا این نماینده مورد توجه خواهد بود.*
- خوشه بندی بهینه آن است که در آن مجموع مربعات این فواصل $J_e$ کمترین باشد.
- وقتی که خوشه ها در عمل بخوبی از هم جدا باشند این روش کاراست خوب عمل می کند ولی اگر همپوشانی خوشه ها زیاد باشد به مشکل برخورد می کند.



# معیار پراکندگی
# Scatter criteria

- استفاده از ماتریس پراکندگی برای تعریف خوشه ها، روش معمول دیگر است

within-scatter matrix $\mathbf{S}_W$

between-scatter matrix $\mathbf{S}_B$

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$$

- معیار trace (sum of diagonal elements) معمولترین معیار بهینه سازی برای پروسه خوشه بندی بر مینای پراکنرگی است

$$tr[S_W] = \sum_{i=1}^{c} tr[S_i] = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \|\mathbf{x} - \mathbf{m}_i\|^2 = J_e$$

# Trace و پراکندگی

- می توان نشان داد که این دو روش تقریبا هم ارز هستند
- چون این معیار مستقل از چرخش است
- $tr[\mathbf{S}_T] = tr[\mathbf{S}_W] + tr[\mathbf{S}_B]$ and $tr[\mathbf{S}_T]$
- minimizing $tr[\mathbf{S}_B]$
- اگر m متوسط عمومی باشد

$$tr[S_B] = \sum_{i=1}^{c} n_i \|\mathbf{m}_i - \mathbf{m}\|^2$$

$$\mathbf{m} = \frac{1}{n}\sum_{D} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i$$

# Iterative optimization

- روش های جستجو:
  - جستجوی کامل که با توجه به محدود بودن تعداد خوشه بندی های ممکن و برخی موارد عملی است
  - جستجوی گام به گام  فرموله ( روش های درجه دوم و مبتنی بر گرادیان)
  - جستجو های رندوم ( هدایت شده) مانند روش جستجوی ژنتیک
  - روش های مختلط: که در آنها در هر بخشی از کار از یکی از روش های یاد شده استفاده می شود.

# جستجو بر اساس تغییر عضویت نمونه ها در خوشه ها

- اگرفرض کنیم مقدار معیار در یک شکل خوشه بندی عبارت باشد از:

$$J_e = \sum_{i=1}^{c} J_i \quad \text{where} \quad J_i = \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

- حال اگر یک نمونه از یکی از خوشه **i** به خوشه **j** منتقل شود، مقدار معیار عوض شده و در صورت کاهش، خوشه بندی بهتر میشود:

$$J_j^* = J_j + \frac{n_j}{n_j + 1} \left\| \hat{\mathbf{x}} - \mathbf{m}_j \right\|^2 \qquad J_i^* = J_i - \frac{n_i}{n_i - 1} \left\| \hat{\mathbf{x}} - \mathbf{m}_i \right\|^2$$

$$\frac{n_i}{n_i - 1} \left\| \hat{\mathbf{x}} - \mathbf{m}_i \right\|^2 > \frac{n_j}{n_j + 1} \left\| \hat{\mathbf{x}} - \mathbf{m}_j \right\|^2$$

# الگوریتم تغییرات پیاپی

■ **Algorithm 3. (Basic Iterative Minimum-Squared-Error Clustering)**

1 **begin initialize** $n, c, \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_c$
2 **do** randomly select a sample $\hat{\mathbf{x}}$
3 $i \leftarrow \arg\min_{i'} \|\mathbf{m}_{i'} - \hat{\mathbf{x}}\|$ (classify $\hat{\mathbf{x}}$)
4 **if** $n_i \neq 1$ **then** compute
5 $$\rho_j = \begin{cases} \frac{n_j}{n_j+1}\|\hat{\mathbf{x}} - \mathbf{m}_j\|^2 & j \neq i \\ \frac{n_j}{n_j-1}\|\hat{\mathbf{x}} - \mathbf{m}_i\|^2 & j = i \end{cases}$$
6 **if** $\rho_k \leq \rho_j$ for all $j$ **then** transfer $\hat{\mathbf{x}}$ to $\mathcal{D}_k$
7 recompute $J_e, \mathbf{m}_i, \mathbf{m}_k$
8 **until** no change in $J_e$ in $n$ attempts
9 **return** $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_c$
10 **end**

# خوشه بندی سلسله مراتبی
# Hierarchical Clustering

- میتوان یک خوشه را شامل تعدادی زیر خوشه در نظر گرفت حال با این تعداد از چه نقطه ای می توان کار خوشه بندی را شروع کرد؟
- برای این کار معمولابرای کاهش بار عملیات و سادگی الگوریتم، از روش های چند مرحله ای یا سلسله مراتبی استفاده می شود. در این مسیردو رویکرد قابل تصور است:
  - رویکرد تراکم agglomerative: *طی آن از تعداد زیادی خوشه شروع کرده و با چسباندن خوشه ها به کم کردن تعداد آنها مبادرت می کنیم.*
  - رویکرد تقسیم divisive : *در آن از تعداد کم خوشه ها شروع کرده و با تقسیم خوشه ها بر تعداد آنها می افزائیم.*

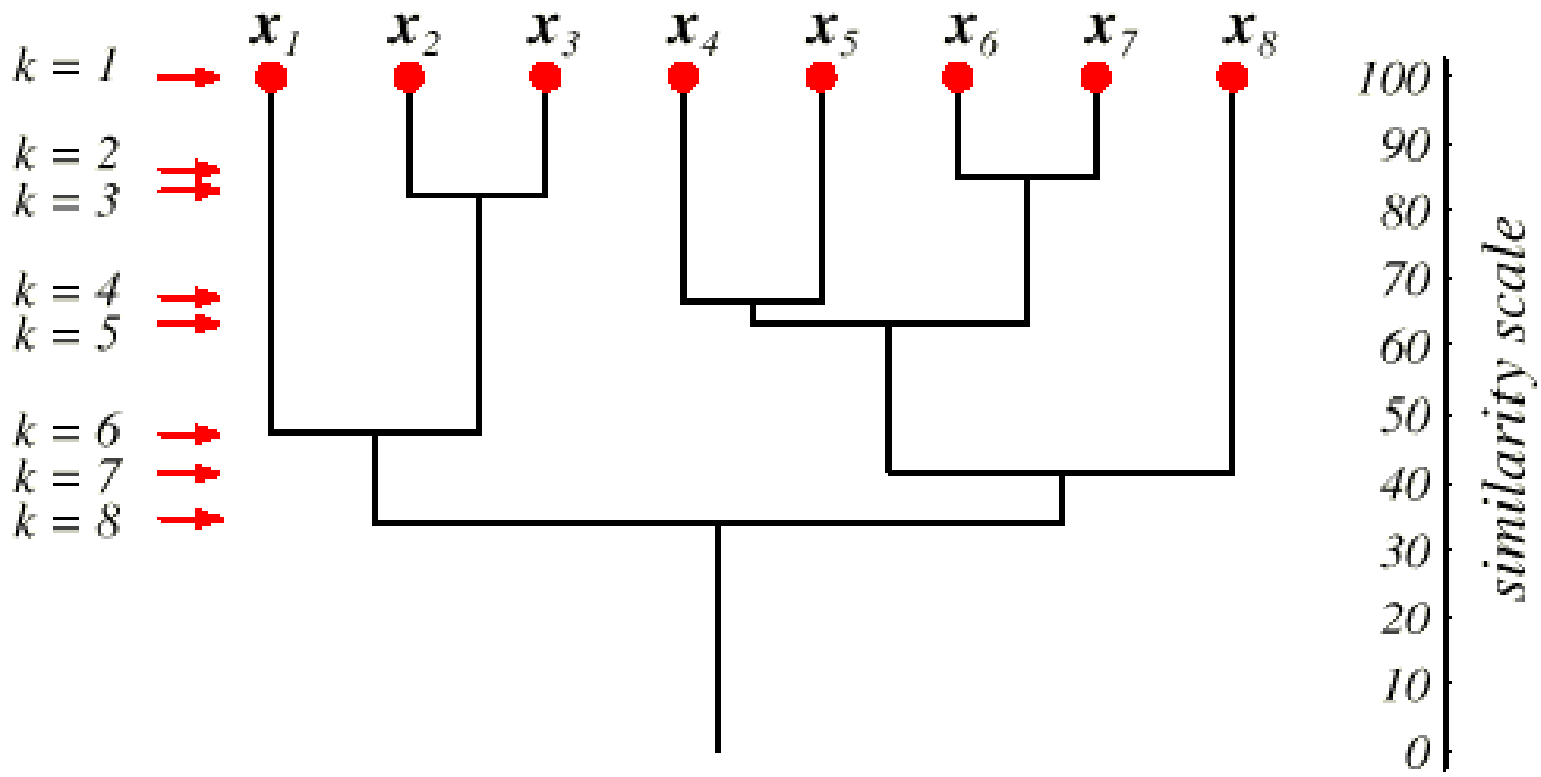
# رویکرد تراکم *agglomerative* (bottom up, clumping)

- فرض کنیید می خواهیم n نمونه را به c خوشه تقسیم کنیم .
  - درنخستین گام می توان هر یک نمونه را یک خوشه دانست
  - در گام بعدی، با به هم چسباندن دو خوشه که نزدیکترین فاصله به هم را دارند، تعداد خوشه ها به n-1 کاهش می یابد.
  - با تکرار این امر، در هر مرحله فاصله حداقل فاصله خوشه ها کمتر میشود تا جائیکه همه نمونه ها در تنها یک خوشه جای بگیرد.
  - با داشتن ایده در مورد **حداکثر حداقل فاصله یا تعداد خوشه ها** می توان چسباندن ها را متوقف کرد.

# رویکرد تقسیم
# (top down, splitting)

- ابتداهمه نمونه هاعضویک خوشه شمرده می شود.
- با مشخص کردن دورترین نمونه ها، آنها از این خوشه جدا شده وخوشه جدیدی را تشکیل می دهند.
- این کار آنقدر ادامه می یابد که یابه تعداد مشخصی از خوشه ها برسیم و یا اینکه فواصل اعضای  تمام نمونه های هر خوشه از مقدار معینی کوچکتر شوند.

# نمودار درختی در تقلیل تعداد خوشه ها



# دیاگرام ون در تشریح تراکم

- استفاده از دیاگرام هائی مانند نمودار ون نیزبخصوص زمانی که ابعاد نمونه ها زیاد نباشد در تشریح نمودار درختی معمول است

VM16

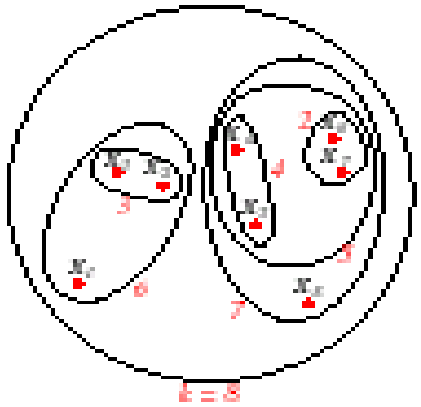


FIGURE 10.12. A set or Venn diagram representation of two-dimensional data (which was used in the dendrogram of Fig. 10.11) reveals the hierarchical structure but not the quantitative distances between clusters. The levels are numbered by *k*, in red. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

## Slide 43

**VM16** this rep reveals hierarchical structure, but does not represent similarities quantitatively!
swan, 5/31/2004

# خوشه بندی چسباننده(Agglomerative)

- **Algorithm 4. (Agglomerative Hierarchical Clustering)**

```
1 begin initialize c, ĉ ← n, D_i ← {x_i}, i = 1, ..., n
2         do ĉ ← ĉ − 1
3             find nearest clusters, say, D_i and D_j
4             merge D_i and D_j
5         until c = ĉ
6   return c clusters
7 end
```

# متریک های مورد نیاز در خوشه بندی سلسه مراتبی

$$d_{\min}(D_i, D_j) = \min_{\mathbf{x}\in D_i, \mathbf{x}'\in D_j} \|\mathbf{x} - \mathbf{x}'\|$$

$$d_{\max}(D_i, D_j) = \max_{\mathbf{x}\in D_i, \mathbf{x}'\in D_j} \|\mathbf{x} - \mathbf{x}'\|$$

$$d_{avg}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{\mathbf{x}\in D_i} \sum_{\mathbf{x}'\in D_j} \|\mathbf{x} - \mathbf{x}'\|$$

$$d_{mean}(D_i, D_j) = \|\mathbf{m}_i - \mathbf{m}_j\|$$

VM17

**VM17** d dim space of the feature x, n number of samples, c clusters

swan, 5/31/2004

# Nearest-neighbor algorithm

- اگر ازمتریک $d_{min}$ استفاده شود الگوریتم نزدیکترین همسایگی حاصل می شود.
- اگر خاتمه این الگوریتم منوط شود به اینکه حداقل فاصله هر دو زیرخوشه، از یک مقدار معین بزرگتر باشد، این الگوریتم پیوند منفرد *single* linkage خوانده می شود.
- می توان نقاط (نمونه ها) را گره های یک گراف فرض کرد که لبه ها در آن مسیریا شاخه ای می سازند که در بر دارنده همه اعضای زیرخوشه $D_i$, است. وصل شدن این زیر خوشه به زیر خوشه $D_i$ای که به آن نزدیک است، می تواند زیر خوشه بزرگتری بسازد
- نتیجه حاصل به درخت گسترده *spanning tree* منجر می شود.
- استفاده از $d_{min}$ ورویکرد تراکم موجب تولید درخت با کمینه گسترش *minimal spanning tree* *خواهد شد.*
- *تاثیر زنجیره ای، این متریک را مورد تاثیر قرار می دهد.*

# مثال: جدا سازی دو دسته بردار گوسی با الگوریتم نزدیکترین همسایگی و حساسیت آن نسبت به داده ها

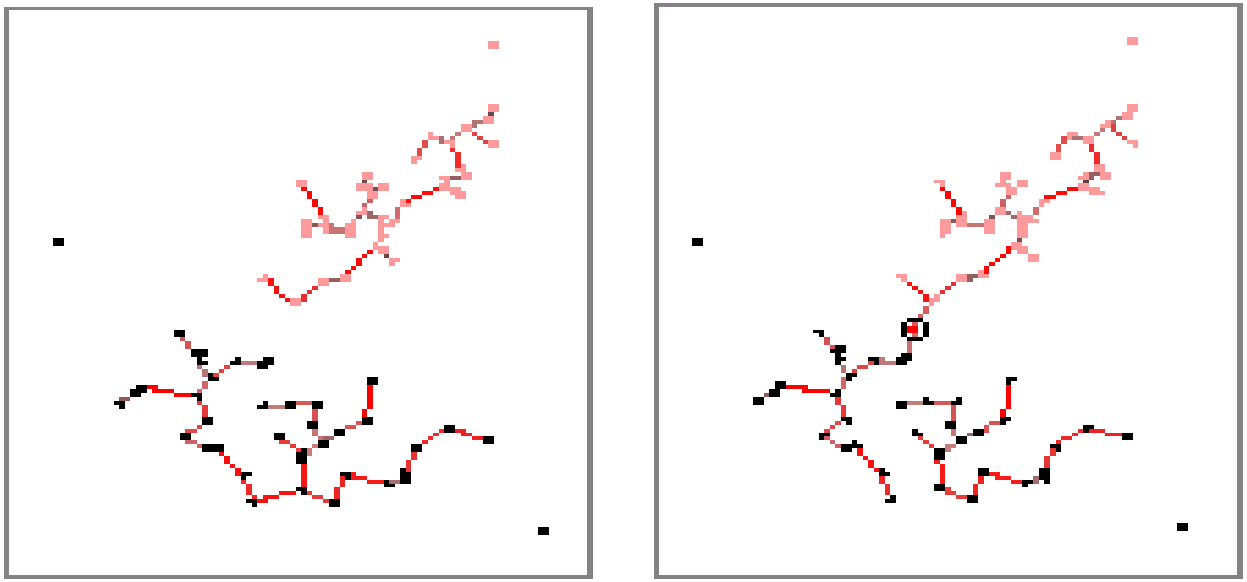


FIGURE 10.13. Two Gaussians were used to generate two-dimensional samples, shown in pink and black. The nearest-neighbor clustering algorithm gives two clusters that well approximate the generating Gaussians (left). If, however, another particular sample is generated (circled red point at the right) and the procedure is restarted, the clusters do not well approximate the Gaussians. This illustrates how the algorithm is sensitive to the details of the samples. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

# الگوریتم دورترین همسایگی
# The farthest neighbor algorithm

- در این الگوریتم از $d_{max}$ استفاده می شود.
- اگر معیار خاتمه آن باشد که فاصله بین دو خوشه نزدیک از حد معینی بیشتر باشد، این الگوریتم به پیوند کامل complete-linkage موسوم است.
- این الگوریتم باعث تضعیف رشد کشیدگی خوشه می شود.
- در تئوری گراف، این الگوریتم باعث می شودهر خوشه شامل یک زیر گراف کامل بوده و فاصله دو خوشه با بیشترین فاصله هر دو گره در دو خوشه تعیین می گردد.

VM9

**VM9** complete graph=a graph in which edges connect all of the nodes in a cluster

swan, 5/31/2004

# مسئله تعداد خوشه ها

- نوعا تعداد خوشه ها از قبل دانسته فرض می شود.
- در غیر اینصورت، به یکی از رویکرد های زیر عمل می شود:
- نخست اینکه بر اساس یک معیارمشخص (مثلا مجموع فواصل داخی اعضای خوشه ها)، خوشه بندی به ازای تعداد مختلف خوشه ها صورت گرفته و هر کدام که بهترین نتیجه را داد پذیرفته می شود.
- رویکرد دیگر بر اساس رسیدن به یک آستانه در فاصله داخی یا خارجی خوشه هاست.
- فی الواقع این رویکرد ها بسیار شبیه روش های موجود در مدل سازی سیستم هاست

## ارزیابی عملکرد در طبقه بندی بدون مربی و خوشه بندی

- بطور کلی در ارزیابی عملرد مسائلی که با یادگیری بدون مربی حل شده اند، با توجه به معلوم نبودن جواب، مشکل اساس وجود دارد.
- رویکرد های معمول:
  - استفاده از نظر خبرگان
  - استفاده از داده های با برچسب و اطلس ها

# خلاصه بحث

- یادگیری بدون مربی، موارد استفاده و پیش فرض ها
- روش های معمول:
  - مبتنی بر یافتن توابع توزیع
  - مبتنی بر فاصله
- خوشه بندی
  - متریک ها
  - معیار ها
  - روش های جستجو
- خوشه بندی سلسله مراتبی