中山大學

Chapter 10

无监督学习和聚类

方艳梅
中山大学数据科学与计算机学院

1

1

中山大學

本章目录

2

2

中山大學

10.1 引言

有监督学习和无监督学习：

- 有监督训练过程
  ——训练样本集中每个样本的类别已经被标记
- 无监督训练过程
  ——使用未被标记的训练样本

3

3

中山大學

“无监督”方法非常有用

- 收集并标记大型样本集非常费时费力
  ——例如：语音信息的记录
- 逆向解决问题：用大量未标记样本集训练，再人工标记数据分组
  ——例如：数据挖掘的应用
- 对于待分类模式性质会随时间变化的情况，使用无监督方法可以大幅提升分类器性能
  ——例：自动食品分类器中食品随季节而改变

4

4

中山大學

- 用无监督方法提取一些对进一步分类很有用的基本特征
  ——独立于数据的“灵巧预处理”，“灵巧特征提取”

- 揭示观测数据的一些内部结构和规律
  ——就能更有效设计有针对性的分类器

5

5

中山大學

10.2 混合密度和可辨识性

基本假设

1. 所有样本来自c种类别，c已知。
2. 每种类别的先验概率 $P(\omega_j)$ 已知， $j=1,\cdots,c$
3. 样本的类条件概率密度具有确定的数学形式 $p(x|\omega_j,\theta_j)$ $j=1,\cdots,c$
4. 参数向量 $\theta_1,\cdots,\theta_c$ 未知
5. 样本类别未标记

6

6

中山大學

混合密度

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \sum_{j=1}^{c} p(\mathbf{x}\,|\,\omega_j, \boldsymbol{\theta}_j) P(\omega_j)$$

参数向量： $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_c)^t$

分量密度： $p(\mathbf{x}\,|\,\omega_j, \boldsymbol{\theta}_j)$

混合参数： $P(\omega_j)$

7

7

中山大學

目标和方法

- 目标：使用从混合密度中取出的样本去估计未知的参数向量θ。
- 一旦θ已知时，将样本的混合密度分解为基本分量，据此设计最大后验（MAP）分类器。

8

8

中山大學

解的存在性

- 假设样本数量无穷；用非参数技术可获得任意样本x上的概率 $p(x|\theta)$
- 如果仅仅存在一个q满足 $p(x|\theta)$，那么理论上存在解。
- 如果几个不同的q取值都产生相同的 $p(x|\theta)$，那么不可能得到唯一的解。

9

9

中山大學

可辨识密度

$p(\mathbf{x}|\theta)$ 可辨识的是指：

如果$\theta \neq \theta'$ $\Rightarrow$ 存在某个x使得$p(\mathbf{x}|\theta) \neq p(\mathbf{x}|\theta')$

$p(\mathbf{x}|\theta)$ 不可辨识：

无论样本数量多少，都不存在唯一的解θ

$p(\mathbf{x}|\theta)$ 完全不可辨识：

参数向量θ的任何部分都无法求出

10

10

例子：不可辨识的离散分布混合密度

x : binary

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

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

$$P(x=1\,|\,\boldsymbol{\theta}) = 0.6,\ P(x=0\,|\,\boldsymbol{\theta}) = 0.4 \Rightarrow$$

$$\theta_1+\theta_2 = 1.2$$

11

11

例子：不可辨识的离散分布混合密度

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

when $P(\omega_1) = P(\omega_2)$

由于 θ_1与 θ_2是可交换的，不影响 $p(x|\theta)$

12

12

中山大學

10.3 最大似然估计

$D=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$

n个样本集合，样本未标记，从混合密度中独立抽取

混合密度 $p(\mathbf{x}\mid\boldsymbol{\theta})=\sum_{j=1}^{c}p(\mathbf{x}\mid\omega_j,\boldsymbol{\theta}_j)P(\omega_j)$

参数向量 $\boldsymbol{\theta}$具有确定但未知的值

样本集的似然函数 $p(D\mid\boldsymbol{\theta})=\prod_{k=1}^{n}p(\mathbf{x}_k\mid\boldsymbol{\theta})$

最大似然估计参数值 $\hat{\boldsymbol{\theta}}$

13

13

中山大學

最大似然估计

样本集似然函数的对数 $l=\sum_{k=1}^{n}\ln p\left(\mathbf{x}_k\mid\boldsymbol{\theta}\right)$

$$\nabla_{\boldsymbol{\theta}_i}l=\sum_{k=1}^{n}\frac{1}{p\left(\mathbf{x}_k\mid\boldsymbol{\theta}\right)}\nabla_{\boldsymbol{\theta}_i}\left[\sum_{j=1}^{c}p(\mathbf{x}_k\mid\omega_j,\boldsymbol{\theta}_j)P(\omega_j)\right]$$

假设参数向量 $\boldsymbol{\theta}_i$ 和 $\boldsymbol{\theta}_j$ 互相独立 （$i\neq j$）

后验概率 $P(\omega_i\mid\mathbf{x}_k,\boldsymbol{\theta})=\dfrac{p(\mathbf{x}_k\mid\omega_i,\boldsymbol{\theta}_i)P(\omega_i)}{p\left(\mathbf{x}_k\mid\boldsymbol{\theta}\right)}$

14

14

中山大學

最大似然估计

$$\nabla_{\boldsymbol{\theta}_i} l = \sum_{k=1}^{n} P(\omega_i \mid \mathbf{x}_k, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}_i} \ln p(\mathbf{x}_k \mid \omega_k, \boldsymbol{\theta}_i)$$

当l最大时，$\hat{\boldsymbol{\theta}}_i$必须满足

$$\sum_{k=1}^{n} P\left(\omega_i \middle| x_k, \hat{\theta}\right) \nabla_{\theta_i} \ln p\left(x_k \middle| \omega_k, \hat{\theta}_i\right) = 0, i = 1, \ldots, c$$

对这个方程求解$\hat{\theta}_i$，就可以得到最大似然估计

15

15

中山大學

先验概率未知时的最大似然估计

$$l = \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} \ln \sum_{j=1}^{c} p(\mathbf{x}_k \mid \omega_j, \boldsymbol{\theta}_j) P(\omega_j)$$

寻找$\boldsymbol{\theta}$ 和 $P(\omega_i)$使得 l 取最大值，且满足

$$P(\omega_i) \geq 0, \ i = 1, \ldots, c$$

$$\sum_{i=1}^{c} P(\omega_i) = 1$$

16

16

中山大學

先验概率未知时的最大似然估计

$\hat{P}(\omega_i)$表示$P(\omega_i)$的最大似然估计，则（习题6）

$$\hat{P}(\omega_i)=\frac{1}{n}\sum_{k=1}^{n}\hat{P}\left(\omega_i\middle|x_k,\hat{\theta}\right)$$

$$\sum_{k=1}^{n}\hat{P}\left(\omega_i\middle|x_k,\hat{\theta}\right)\nabla_{\theta_i}\ln p\left(x_k\middle|\omega_i,\hat{\theta}_i\right)=0$$

其中$\hat{P}(\omega_i\mid\mathbf{x}_k,\hat{\boldsymbol{\theta}})=\dfrac{p(\mathbf{x}_k\mid\omega_i,\hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{c}p(\mathbf{x}_k\mid\omega_j,\hat{\boldsymbol{\theta}}_j)\hat{P}(\omega_j)}$

17

17

中山大學

10.4 对混合正态密度的应用

- **分量密度** $p(\mathbf{x}\mid\omega_i,\theta_i)\sim N(\mu_i,\Sigma_i)$
- **三种情况:**

Case	μ_i	Σ_i	$P(\omega_i)$	c
1	?	√	√	√
2	?	?	?	√
3	?	?	?	?

18

18

情况1：未知均值向量

中山大學

$$\ln p(\mathbf{x}\mid\omega_i,\boldsymbol{\mu}_i)=-\ln\left[(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}\right]-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

$$\nabla_{\boldsymbol{\mu}_i}\ln p(\mathbf{x}\mid\omega_i,\boldsymbol{\mu}_i)=\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

$$\sum_{k=1}^{n}P(\omega_i\mid\mathbf{x}_k,\hat{\boldsymbol{\mu}})\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_i)=0,\quad \hat{\boldsymbol{\mu}}=(\hat{\boldsymbol{\mu}}_1,\ldots,\hat{\boldsymbol{\mu}}_c)^t$$ 根据(8)式得到

$$\hat{\boldsymbol{\mu}}_i=\frac{\sum_{k=1}^{n}P(\omega_i\mid\mathbf{x}_k,\hat{\boldsymbol{\mu}})\mathbf{x}_k}{\sum_{k=1}^{n}P(\omega_i\mid\mathbf{x}_k,\hat{\boldsymbol{\mu}})}$$

$P(\omega_i\mid x_k,\hat{\mu})$是来自第$i$类和具有值$x_k$的那些样本的一部分，$\hat{\mu}_i$实质上是来自第i类样本的平均值。

19

19

情况2：所有参数未知

中山大學

- 对协方差矩阵没有约束
- 令 $p(x\mid\mu,\sigma^2)$ 表示一个由两分量组成的混合密度:

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

20

20

中山大學

假设 $\mu = x_1$, 则:

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

对其他样本:

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

从而,

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

似然函数可以任意地大，参数解是奇异的。

21

21

中山大學

□ **增加一个假设**

只取似然函数的局部最优点中对应最大有界值的那一个，假设似然函数在这个点附近的特性足够好，则有如下迭代算法:

Iterative scheme

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

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

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

22

22

中山大學

其中:

$$\hat{P}(\omega_i \mid \mathbf{x}_k, \hat{\boldsymbol{\theta}}) = \frac{p(\mathbf{x}_k \mid \omega_i, \hat{\boldsymbol{\theta}}_i)\hat{P}(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}_k \mid \omega_j, \hat{\boldsymbol{\theta}}_j)\hat{P}(\omega_j)}$$

$$= \frac{\left|\hat{\boldsymbol{\Sigma}}_i\right|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\right)^t \hat{\boldsymbol{\Sigma}}_i^{-1}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\right)\right]\hat{P}(\omega_i)}{\sum_{j=1}^{c}\left|\hat{\boldsymbol{\Sigma}}_j\right|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j\right)^t \hat{\boldsymbol{\Sigma}}_j^{-1}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j\right)\right]\hat{P}(\omega_j)}$$

23

23

中山大學

k-均值聚类

$\hat{P}(\omega_i \mid \mathbf{x}_k, \hat{\theta})$ 随着马氏距离的平方 $\left(\mathbf{x}_k - \hat{\mu}_i\right)^t \hat{\Sigma}_i^{-1}\left(\mathbf{x}_k - \hat{\mu}_i\right)$ 的减少而增大，用近似的方法，通过计算 $\left\|\mathbf{x}_k - \hat{\mu}_i\right\|^2$

找到最接近 $\mathbf{x}_k$ 的类中心 $\hat{\mu}_m$ ，并取 $\hat{P}(\omega_i \mid \mathbf{x}_k, \hat{\theta})$ 的近似:

$$\hat{P}(\omega_i \mid \mathbf{x}_k, \hat{\boldsymbol{\theta}}) = \begin{cases} 1 & \text{if } i = m \\ 0 & \text{otherwise} \end{cases}$$

迭代计算

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

24

24

中山大學

K-均值算法--动态聚类的算法

① 先选定某种距离作为样本间的相似性的度量;

② 确定评价聚类结果的准则函数;

③ 给出某种初始分类，用迭代法找出使准则函数取极值的最好的聚类结果。

25

25

中山大學

k-均值聚类

算法1（*k*-均值聚类）

1. **begin initialize** $n, c, \mu_1, \mu_2, \ldots, \mu_c$
2. **do** 按照最近邻 μ_i分类n个样本
3. 重计算 μ_i
4. **until** 不再改变 μ_i
5. **return** $\mu_1, \mu_2, \ldots, \mu_c$
6. **end**

26

26

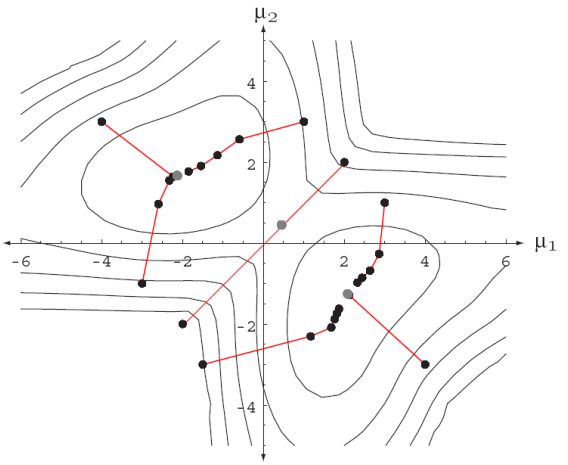
k-均值聚类

- 复杂度$O(ndcT)$
- 在实践中迭代次数通常小于样本的数量
- 从这个算法中得到的结果既可以作为最终答案，也可以作为进一步计算的初始值

27

27

k-均值聚类



$\hat{\mu}_1 \approx -2.176$

$\hat{\mu}_2 \approx 1.684$

极大似然方法结果

$\hat{\mu}_1 \approx -2.130$

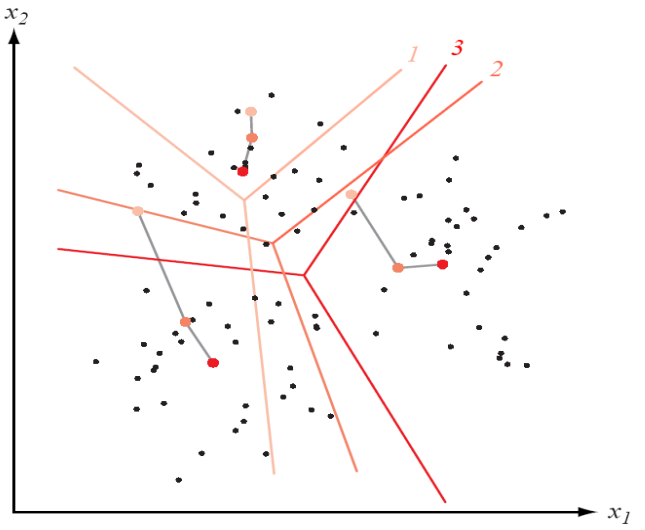
$\hat{\mu}_2 \approx 1.688$

28

28

k-均值聚类

- k-均值聚类用于二维数据的迭代轨迹。



29

29

10.5 无监督贝叶斯学习

- 与 ML 估计类似, Bayesian 估计技术也能用于无监督情况。假设如下:
  - 类别数c已知;
  - 先验概率已知;
  - 类条件概率密度的数学形式已知，但θ参数未知;
  - 关于θ的先验知识由概率密度$p(q)$表示;
  - 剩下的关于q的知识都存在于样本集中。

30

30

- 直接计算后验概率密度 $p(\theta|D)$

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D)P(\omega_i \mid D)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D)P(\omega_j \mid D)} = \frac{p(\mathbf{x} \mid \omega_i, D)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D)P(\omega_j)}$$

注意：其中借助参数向量 θ 来表示类条件概率。

$$p(\mathbf{x} \mid \omega_i, D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid \omega_i, D)d\boldsymbol{\theta} = \int p(\mathbf{x} \mid \boldsymbol{\theta}, \omega_i, D)\, p(\boldsymbol{\theta} \mid \omega_i, D)d\boldsymbol{\theta}$$

$$= \int p(\mathbf{x} \mid \omega_i, \boldsymbol{\theta}_i)\, p(\boldsymbol{\theta} \mid D)d\boldsymbol{\theta}$$

对 $p(x|\omega_i)$ 的最好估计是通过对 $p(x \mid \omega_i, \theta_i)$ 在 θ_i上的加权积分得到。

- 估计的好坏决定于$p(\theta \mid D)$。

31

31

- 根据 Bayes 准则

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- 假设样本互相独立

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

或者利用递归 (用D^n表示D中前面n个样本集合)

$$p(\boldsymbol{\theta} \mid D^n) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D^{n-1}, \omega_i)}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D^{n-1}, \omega_i)d\boldsymbol{\theta}}$$

- 如果 $p(\theta)$ 在$p(D|\theta)$达到峰值的附近接近均匀分布，则$p(\theta|D)$也会在同样区域达到峰值。

32

32

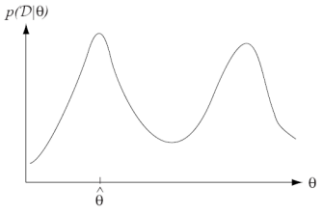
- 如果在 $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$ 的附近出现最主要的尖峰，则

$$p(\mathbf{x}\mid\omega_i,D)\cong p(\mathbf{x}\mid\omega_i,\hat{\boldsymbol{\theta}})\cong p(\mathbf{x}\mid\omega_i,\hat{\boldsymbol{\theta}}_i)$$

且

$$P(\omega_i\mid\mathbf{x},D)\cong\frac{p(\mathbf{x}\mid\omega_i,\hat{\boldsymbol{\theta}}_i)P(\omega_i)}{\sum_{j=1}^{c}p(\mathbf{x}\mid\omega_j,\hat{\boldsymbol{\theta}}_j)P(\omega_j)}$$

- 因此, ML 估计具有合理性。
- 当数据量非常大的时候，ML估计和贝叶斯方法会取得近似一致的效果。
- 在小样本集情况下，逼近效果不理想。
- ML 方法更易实现。



33

33

- 有监督学习和无监督学习之间的最明显不同: 可辨别性、计算复杂性。
- 可辨别性:
  - 对有监督学习, 缺少可辨别性表明求出的参数向量并不唯一，不带来严重问题。
  - 对无监督学习，缺少可辨别性，混合密度就不能分解为各种真实的分量。

    $\Rightarrow p(\mathrm{x}\mid D^n)$ 仍然收敛到$p(\mathrm{x})$, 但$p(\mathrm{x}\mid\omega_i, D^n)$ 一般不会收敛到$p(\mathrm{x}\mid\omega_i)$, 这是理论上的障碍.

- 计算复杂度
  - 对有监督学习，如果能找到充分的统计量，就会得到解析的解。
  - 对无监督学习，计算 $p(D\mid\theta)$异常复杂。

34

34

中山大學

- 另外一种比较有监督和无监督学习的方法是用到混合密度，得到:

$$p(\boldsymbol{\theta}\mid D^n)=\frac{p(\mathbf{x}_n\mid\boldsymbol{\theta})p(\boldsymbol{\theta}\mid D^{n-1})}{\int p(\mathbf{x}_n\mid\boldsymbol{\theta})p(\boldsymbol{\theta}\mid D^{n-1})d\boldsymbol{\theta}}=$$

$$=\frac{\sum_{j=1}^{c}p(\mathbf{x}_n\mid\omega_j,\boldsymbol{\theta}_j)P(\omega_j)}{\sum_{j=1}^{c}\int p(\mathbf{x}_n\mid\omega_j,\boldsymbol{\theta}_j)P(\omega_j)p(\boldsymbol{\theta}\mid D^{n-1})d\boldsymbol{\theta}}p(\boldsymbol{\theta}\mid D^{n-1})$$

- 考虑 $P(\omega_1)=1$ 的情况，所有样本来自于类别ω_1, 此正好对应有监督学习，上式可化简为

35

35

中山大學

$$p(\boldsymbol{\theta}\mid D^n)=\frac{p(\mathbf{x}_n\mid\omega_1,\boldsymbol{\theta}_1)}{\int p(\mathbf{x}_n\mid\omega_1,\boldsymbol{\theta}_1)p(\boldsymbol{\theta}\mid D^{n-1})d\boldsymbol{\theta}}p(\boldsymbol{\theta}\mid D^{n-1})$$

- 比较这两个方程, 观察增加一个样本对θ估计的影响.
- 忽略用来归一化的分母。最主要区别是:
  - 对有监督学习 SL, 通过先验密度P(θ)和分量密度$p(x_n/\omega_1,\theta_1)$的乘积来获得后验密度。
  - 对非监督学习,由参数先验密度和混合密度的乘积来获得后验概率密度:

$$\sum_{j=1}^{c}p(\mathbf{x}_n\mid\omega_j,\boldsymbol{\theta}_j)P(\omega_j)$$

- 假设样本x_n来自于 ω_1, 无监督学习由于不知道样本所属类别而减少了x_n对 θ的影响.

36

36

Example 2: Unsupervised learning of Gaussian data

As an example, consider the one-dimensional, two-component mixture with $p(x|\omega_1) \sim N(\mu, 1), p(x|\omega_2, \theta) \sim N(\theta, 1)$, where $\mu, P(\omega_1)$ and $P(\omega_2)$ are known. Here we have

$$p(x|\theta) = \underbrace{\frac{P(\omega_1)}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\mu)^2\right]}_{\omega_1} + \underbrace{\frac{P(\omega_2)}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-\theta)^2\right]}_{\omega_2},$$

and we seek the mean of the second component.

Then after one observation $(x = x_1)$ we have

$$\begin{aligned} p(\theta|x_1) &= \alpha p(x_1|\theta)p(\theta) \\ &= \left\{ \begin{array}{ll} \alpha'\{P(\omega_1)\exp[-\frac{1}{2}(x_1-\mu)^2]+ & \\ \quad P(\omega_2)\exp[-\frac{1}{2}(x_1-\theta)^2]\} & a \le \theta \le b \\ 0 \qquad \text{otherwise} & \end{array} \right\}, \end{aligned}$$

37

37

With the addition of a second sample x_2, $p(\theta|x_1)$ changes to

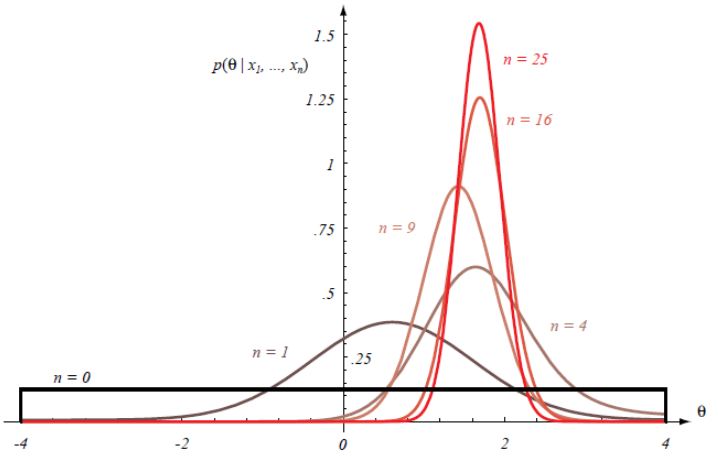
$$\begin{aligned} p(\theta|x_1, x_2) &= \beta p(x_2|\theta)p(\theta|x_1) \\ &= \left\{ \begin{array}{ll} \beta'\{P(\omega_1)P(\omega_1)\exp\left[-\frac{1}{2}(x_1-\mu)^2 - \frac{1}{2}(x_2-\mu)^2\right] & \\ \quad +[P(\omega_1)P(\omega_2)\exp\left[-\frac{1}{2}(x_1-\mu)^2 - \frac{1}{2}(x_2-\theta)^2\right] & \\ \quad +[P(\omega_2)P(\omega_1)\exp\left[-\frac{1}{2}(x_1-\theta)^2 - \frac{1}{2}(x_2-\mu)^2\right] & \\ \quad +[P(\omega_2)P(\omega_2)\exp\left[-\frac{1}{2}(x_1-\theta)^2 - \frac{1}{2}(x_2-\theta)^2\right]\} & \\ & a \le \theta \le b \\ 0 & \text{otherwise.} \end{array} \right. \end{aligned}$$

38

38

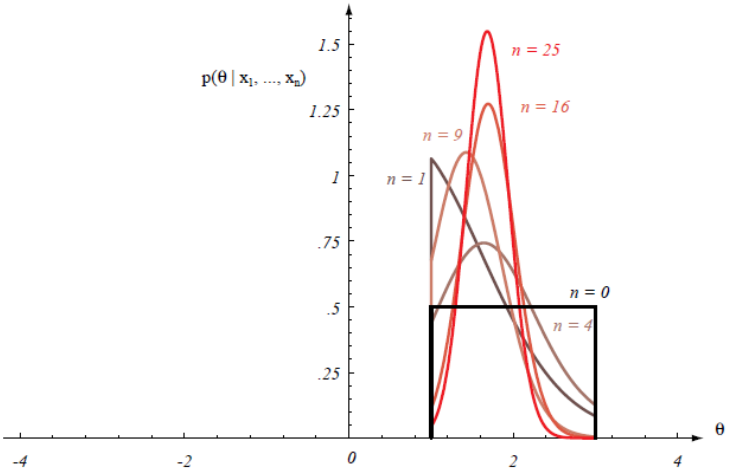
It is possible to use the relation

$$p(\theta|\mathcal{D}^n) = \frac{p(x_n|\theta)p(\theta|\mathcal{D}^{n-1})}{\int p(x_n|\theta)p(\theta|\mathcal{D}^{n-1})\ d\theta}$$



39

39



40

40

10.5.3 判定导向的近似解

- 无监督学习不会因为很难找到解析解而被放弃，它实在太重要。幸好人们找到了很多可以得到近似解的方法。
- 一种方法就是，用先验信息设计一个分类器，然后用这个分类器对样本的判定标识样本进行分类。即，判定导向法 **decision-directed.**
- 缺点：如果初始分类器不好，导致分类器向着错误的方向发展。有缺陷的解总是比没有解好。

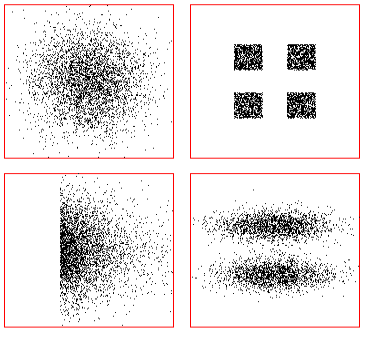
41

41

10.6 数据描述和聚类

- 多维模式的结构对聚类非常重要
- 如果我们知道数据来自某一特定分布，这些数据能被一组紧致参数描述(*充分统计量*)

- 如果将样本考虑成符合某特定分布，然而事实上并不符合该分布，则该数据描述容易误导人（如图）



二阶统计量不足以揭示数据集的空间结构.

42

42

中山大學

相似性度量—自然分组方法

- **问题:**
  - 怎样度量样本之间的相似性?
  - 怎样衡量对样本集的一种划分的好坏?
- 最明显的度量：样本距离。
- 目的：同类样本间距明显小于异类样本间距。

43

43

中山大學

- **欧氏距离: 距离阈值 d_0**

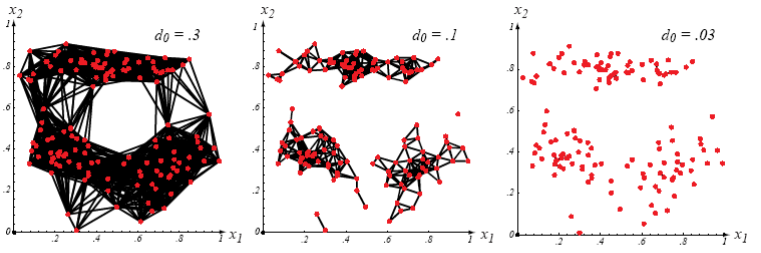
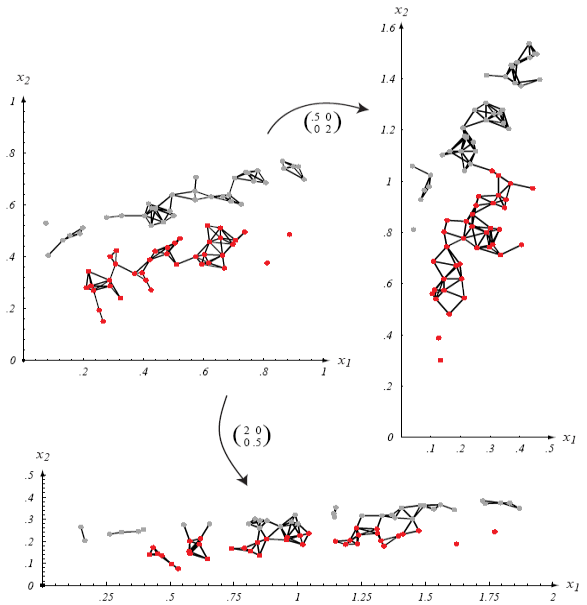


Figure 10.6: The distance threshold affects the number and size of clusters. Lines are drawn between points closer than a distance d_0 apart for three different values of d_0 — the smaller the value of d_0, the smaller and more numerous the clusters.

- 欧氏距离对特征空间的平移和旋转不变，但是一般上对线性变换或其他会扭曲距离关系的变换不能保证不变性。

44

44



- 坐标轴的尺度变换对聚类的影响

45

45

Is this kind of normalization necessarily desirable?

- If the data fall into well-separated clusters (left), normalization by a whitening transform for the full data may reduce the separation, and hence be undesirable (right). Such a whitening normalization may be appropriate if the full data set arises from a single fundamental process (with noise), but inappropriate if there are several different processes.

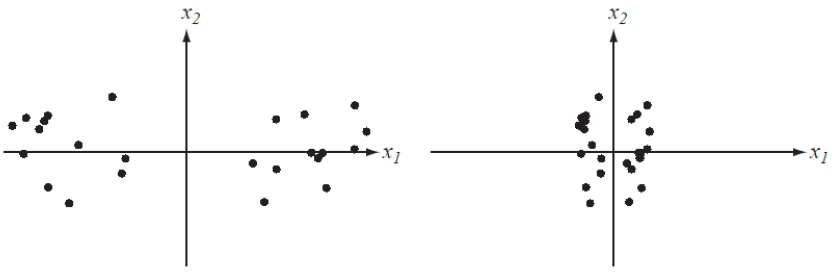


Figure 10.9

46

46

中山大學

- 为达到不变性，一般对数据进行归一化。 例如，化为零均值和单位方差可以得到位移和缩放的不变性，主成份变换可以达到旋转不变性。

- Minkowsky 度量（见Ch.4）

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

其中 $q \geq 1$：

$q = 1 \Rightarrow$ Manhattan （city block） 度量

$q = 2 \Rightarrow$ 欧氏距离

- 除了距离，还有非度量的相似性函数： s(x,x’)

47

47

中山大學

- 内积

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

- 对二值特征

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

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

Tanimoto 距离

48

48

10.7 聚类的准则函数

- **问题二: 如何评估聚类结果?**
- **准则函数**
  - 误差平方和准则和相应变化形式。
  - 有关最小方差准则
  - 散布准则
- **误差平方和**
  - n_i 为 D_i 中的样本数， m_i 为对应样本均值:

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

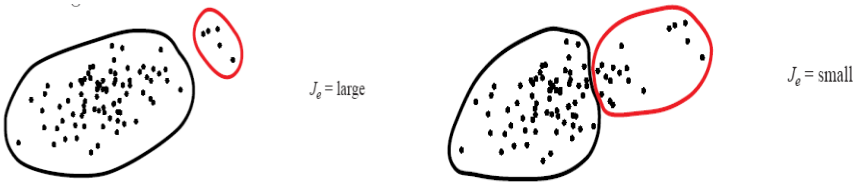
49

49

- 误差平方和:

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

- 最优划分为使得 J_e 最小的划分。
- 当数据点能划分成能很好区分的几类，而且类内数据又很稠密时，该准则表现较好； 当不同聚类所包含的样本个数相差较大时，将一个大的类别分割反而可能具有更小的误差平方和，如图，“出格点”



50

50

散布准则

- 用于多重判别分析, i.e., S_W 和 S_B

$$S_T = S_B + S_W$$

仅仅取决于样本集(与具体划分方式无关)

- 大致上，两个量之间存在一定的互补关系。
- 散布矩阵最简单的标量度量是迹（对角线上元素的和），代表的是散布半径的平方，因为它正比于数据在各个坐标轴方向上的方差的和。
- 与误差平方和准则是完全等价。

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

51

51

- $tr[S_T] = tr[S_W] + tr[S_B]$ 且 $tr[S_T]$ 与划分无关。
- 最小化 $J_e = tr[S_W]$, 同时最大化类间准则：

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

其中 m 是整体样本均值:

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

52

52

聚类中用到的均值向量和散度矩阵

中山大學

Table 10.1: Mean vectors and scatter matrices used in clustering criteria.

	Depend on cluster center? Yes	Depend on cluster center? No	
Mean vector for the ith cluster		×	$\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{x}$ (54)
Total mean vector		×	$\mathbf{m} = \frac{1}{n}\sum_{\mathcal{D}}\mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i\mathbf{m}_i$ (55)
Scatter matrix for the ith cluster	×		$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$ (56)
Within-cluster scatter matrix	×		$\mathbf{S}_W = \sum_{i=1}^{c}\mathbf{S}_i$ (57)
Between-cluster scatter matrix	×		$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t$ (58)
Total scatter matrix		×	$\mathbf{S}_T = \sum_{\mathbf{x}\in\mathcal{D}}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t$ (59)

53

53

中山大學

散布准则

- 基于迹的准则：$tr\ \mathbf{S}_W = \sum_{i=1}^{c} tr\ \mathbf{S}_i = \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}\|\mathbf{x}-\mathbf{m}_i\|^2 = J_e.$
- 基于行列式的准则：

$$J_d = |\mathbf{S}_W| = \left|\sum_{i=1}^{c}\mathbf{S}_i\right|.$$

- 基于不变量的准则：

$$tr\mathbf{S}_W^{-1}\mathbf{S}_B = \sum_{i=1}^{d}\lambda_i. \qquad J_f = tr\mathbf{S}_T^{-1}\mathbf{S}_W = \sum_{i=1}^{d}\frac{1}{1+\lambda_i}$$

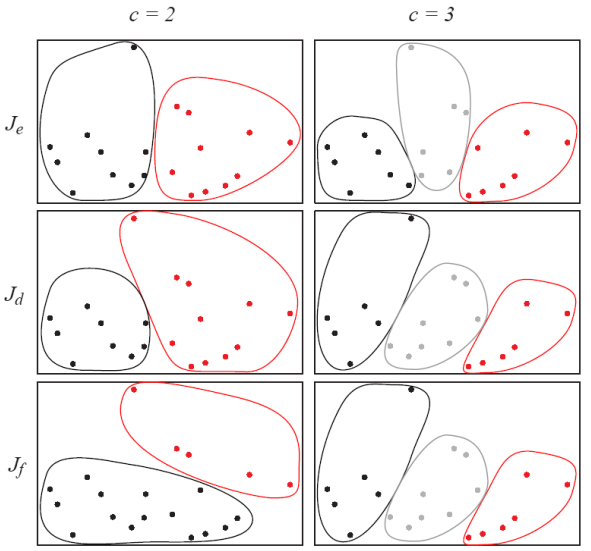
54

54

例2：聚类准则

中山大學

sample	x_1	x_2
1	-1.82	0.24
2	-0.38	-0.39
3	-0.13	0.16
4	-1.17	0.44
5	-0.92	0.16
6	-1.69	-0.01
7	0.33	-0.17
8	-0.71	-0.21
9	1.27	-0.39
10	-0.16	-0.23
11	0.41	0.91
12	1.70	0.48
13	0.92	-0.49
14	2.41	0.32
15	1.48	-0.23
16	-0.34	1.88
17	0.83	0.23
18	0.62	0.81
19	-1.42	-0.51
20	0.67	-0.55



55

55

§ 10.8 Iterative optimization

中山大學

- Once a criterion function has beem selected, clustering becomes a problem of discrete optimization.
- As the sample set is finite there is a finite number of possible partitions, and the optimal one can be always found by exhaustive search.
- Most frequently, it is adopted an iterative optimization procedure to select the optimal partitions
- The basic idea lies in starting from a reasonable initial partition and “move” samples from one cluster to another trying to minimize the criterion function.
- In general, this kinds of approaches guarantee local, not global, optimization.

56

56

- Let us consider an iterative procedure to minimize the sum-of-squared-error criterion J_e

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

where J_i is the effective error per cluster.

- It can be proved that if a sample $\hat{\mathbf{x}}$ currently in cluster D_i is tentatively moved in D_j, the change of the errors in the 2 clusters is

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

- Hence, the transfer is advantegeous if the decrease in J_i is larger than the increase in J_j

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

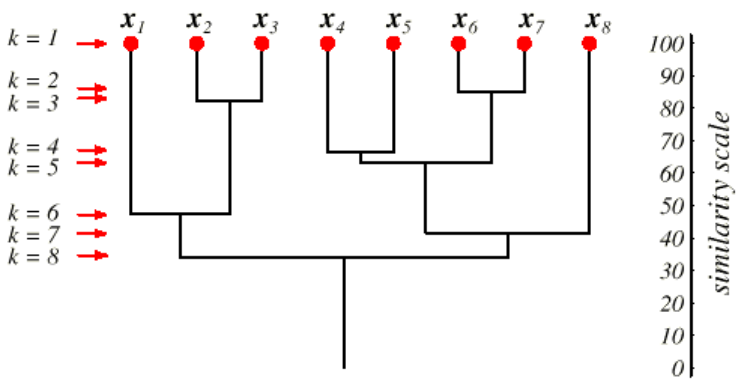
Algorithm 3 (Basic iterative minimum-squared-error clustering)

```
begin initialize n, c, m_1, m_2, ..., m_c
        do randomly select a sample x̂;
            i ← arg min_{i'} ||m_{i'} − x̂||      (classify x̂)
            if n_i ≠ 1 then compute
                                ρ_j = { n_j/(n_j+1) ||x̂ − m_j||^2   j ≠ i
                                      { n_j/(n_j−1) ||x̂ − m_i||^2   j = i
                                if ρ_k ≤ ρ_j for all j then transfer x̂ to D_k
                                    recompute J_e, m_i, m_k
            until no change in J_e in n attempts
        return m_1, m_2, ..., m_c
end
```

- This procedure is a sequential version of the *k-means* algorithm, with the difference that *k-means* waits until n samples have been reclassified before updating, whereas the latter updates each time a sample is reclassified.
- This procedure is more prone to be trapped in local minima, and depends from the order of presentation of the samples, but it is *online!*
- Starting point is always a problem:
  - Random centers of clusters
  - Repetition with different random initialization
  - *c*-cluster starting point as the solution of the (*c-1*)-cluster problem plus the sample farthest from the nearer cluster center

§ 10.9 Hierarchical Clustering

- Given any two samples x and x', they will be grouped together *at some level*, and if they are grouped a level *k*, they remain grouped for all higher levels
- Hierarchical clustering ⇒ tree representation called *dendrogram*



- The similarity values may help to determine if the grouping are natural or forced, but if they are evenly distributed no information can be gained
- Another representation is based on set, e.g., on the Venn diagrams

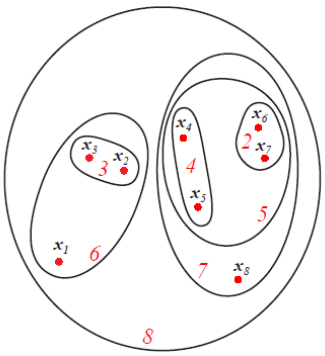


Figure 10.11: A set or Venn diagram representation of two-dimensional data (which was used in the dendrogram of Fig. 10.10) reveals the hierarchical structure but not the quantitative distances between clusters. The levels are numbered in red.

61

61

- Hierarchical clustering can be divided in *agglomerative* (合并) and *divisive*（分裂）.
- Agglomerative (bottom up, clumping): start with *n* singleton cluster and form the sequence by merging clusters
- Divisive (top down, splitting): start with all of the samples in one cluster and form the sequence by successively splitting clusters

62 Pattern Classification, Chapter 10

62

62

Agglomerative hierarchical clustering

Algorithm 4 (Agglomerative hierarchical clustering)

```
begin initialize c, ĉ ← n, D_i ← {x_i}, i = 1, ..., n
        do ĉ ← ĉ − 1
            Find nearest clusters, say, D_i and D_j
            Merge D_i and D_j
        until c = ĉ
    return c clusters
end
```

- **The procedure terminates when the specified number of cluster has been obtained, and returns the cluster as sets of points, rather than the mean or a representative vector for each cluster**

- At any level, the distance between nearest clusters can provide the dissimilarity value for that level
- To find the nearest clusters, one can use

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

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

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

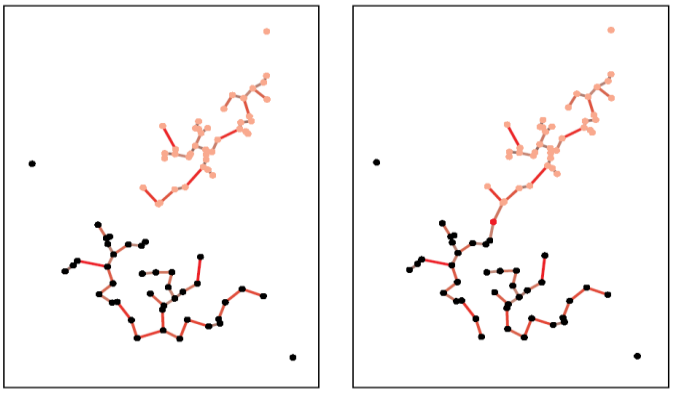
$$d_{mean}(D_i, D_j) = \|\mathbf{m}_i - \mathbf{m}_j\|$$

which behave quite similar of the clusters are hyperspherical (超球面) and well separated.

- The computational complexity is $O(cn^2d^2)$, $n>>c$

Nearest-neighbor algorithm

- The use of d_{min} as a distance measure and the agglomerative clustering generate a *minimal spanning tree*

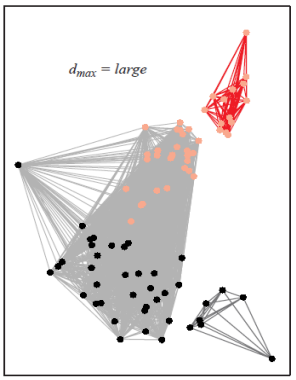


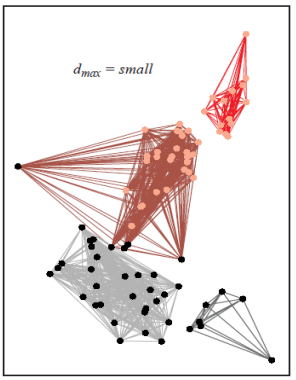
- 一旦有一个新样本点，如右图中的红点儿，重新运行，结果大不相同
- *Chaining* effect: defect of this distance measure (right)

The farthest neighbor algorithm

- When d_{max} is used, the algorithm is called the *farthest neighbor* algorithm
- If it is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called *complete-linkage* algorithm（全连接算法）
- This method discourages the growth of elongated clusters
- In the terminology of the graph theory, every cluster constitutes a complete subgraph, and the distance between two clusters is determined by the most distant nodes in the 2 clusters

- When two clusters are merged, the graph is changed by adding edges between every pair of nodes in the 2 clusters





- All the procedures involving minima or maxima are sensitive to outliers. The use of d_{mean} or d_{avg} are natural compromises

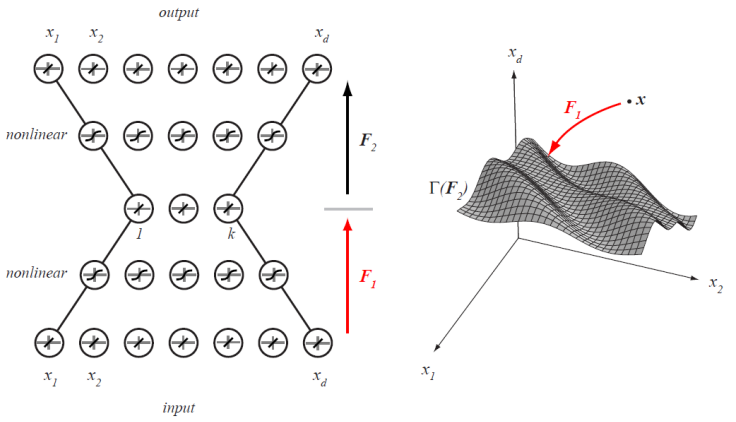
The problem of the number of clusters

- Typically, the number of clusters is known.
- When it's not, there are several ways of proceed.
- When clustering is done by extremizing a criterion function, a common approach is to repeat the clustering with c=1, c=2, c=3, etc.
- Another approach is to state a threshold for the creation of a new cluster; this is adapt to on line cases but depends on the order of presentation of data.
- These approaches are similar to *model selection* procedures, typically used to determine the topology and number of states (e.g., clusters, parameters) of a model, given a specific application.

中山大學

§ 10.13 Component Analysis

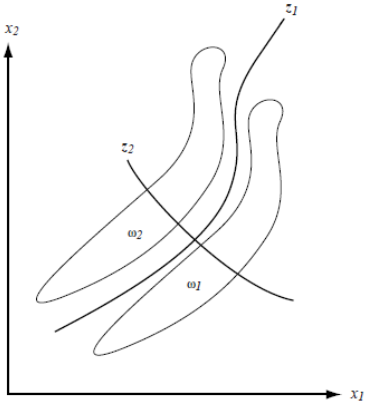
- 五层神经网络，包含两个非线性，训练成为一个自动编解码器(auto-encoder)



中山大學

主成分分析及非线性成分分析

主分量比较明显，但噪声的存在使得最大的非线性成分是沿着z1的曲线方向



中山大學

- 强化学习

- Reinforcement Learning, 是指从环境状态到行为映射的学习，以使系统行为从环境中获得的累计奖励值最大的一种机器学习。
- 基于评判的学习，介于有标记和无标记之间的状态；
- 例如，当某个样本的分类结果被评判为正确时，那么就允许更新权向量，否则就拒绝更新。
- 在智能控制机器人及分析预测等领域有很多应用

71

71

中山大學

本章小结

- **Unsupervised learning and clustering seek to extract information from unlabeled samples.**
- **If the underlying distribution comes from a mixture of component densities described by a set of unknown parameters θ, then θ can be estimated by Bayesian or maximum-likelihood methods.**

72

72

中山大學

本章小结

- A more general approach is to define some measure of similarity between two clusters, as well as a global criterion such an a sum –squared-error or trace of a scatter matrix.
- Because there are only occasionally analytic methods for computing the clustering which optimizes the criterion, a number of greedy iterative algorithms can be used , such as *k*-means and fuzzy *k*– means clustering.