Course ID 80250993 Dates: 9/16-12/23/2021 @ I-205
Tencent/VooV Meeting 521 4678 6257 (passwd 1205)
Broadcast and Reply: https://meeting.tencent.com/l/p11V4drTYoa4 (passwd 1205)

Part III
Unsupervised Learning

Xuegong Zhang
November 18, 2021

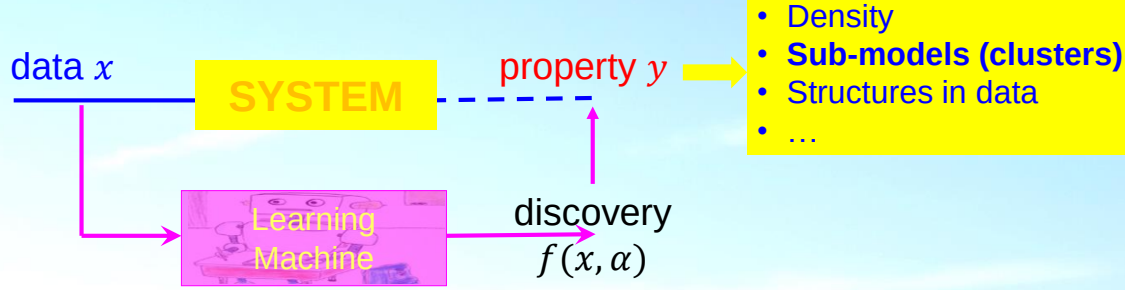
80250993 Machine Learning
@Tsinghua University

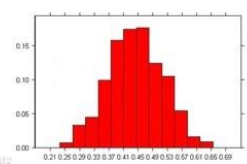
Lecture 13
Clustering:
Unsupervised Pattern Recognition

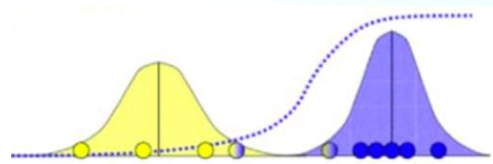
Xuegong Zhang
Nov. 18, 2021

13.1
The Task of Clustering and Partitioning

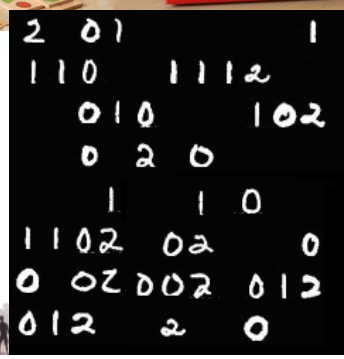
Unsupervised Learning

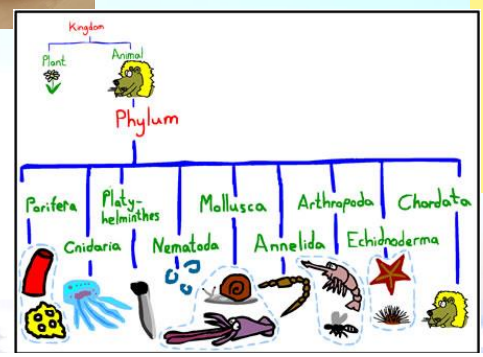


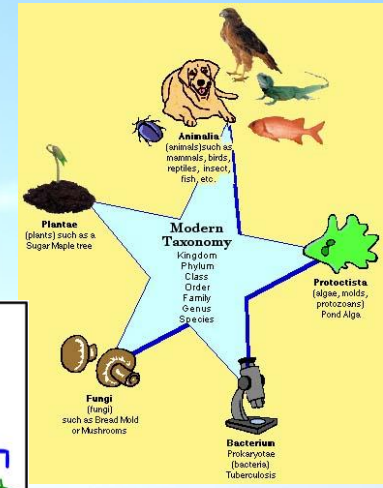




Another type of classification

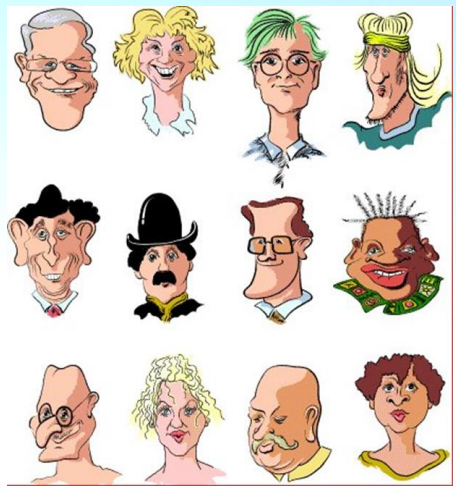






5

The task of unsupervised pattern recognition



- Are they all of the same class or of different classes?
- Classes 类 ➔ **Clusters 聚类**

What are clusters?

- Samples of the same cluster must have something in common, and those of different clusters must have some differences.
- Parametric approach — Generative models, Probabilistic Inference
  - A cluster is a probabilistic density function
  - Mixture Models
  → "model-based" methods
- Nonparametric approach — Deterministic Learning
  - A cluster is a region in the sample space
  → "distance-based", "model-free" methods

Model-free methods:
Dynamic Clustering

Basic idea:

- Divide samples into clusters according to some similarity/distance measure
  - Samples of same clusters are similar; Samples of different clusters are dissimilar.
    - Similarity or distance measure on the features
      —— Should be relevant for some question of interest
  - Define an objective function, and divide samples into clusters to optimize the objective function
- 3 Key Points
  - Distance/similarity measure
  - Objective function
  - Dividing/clustering algorithm

13.2
K-Means or C-Means Clustering

13.2.1
Standard K-Means Clustering

The C-means (K-means) Algorithm

- Euclidean distance (can be replaced by others)
- Objective function: minimal mean square error

$$J_e = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \|y - m_i\|^2 = \sum_{i=1}^{c} J_i$$

Γ_i: cluster i; sample counts: $N_i, \mathrm{i} = 1, \cdots, c$; cluster means: $m_i = \frac{1}{N_i} \sum_{y \in \Gamma_i} y$

- Intuition:
  - Use c (or k) means to represent all samples
  - Find the means to give minimal representation error
  - aka. *Vector Quantization*: using a code-book of length c $(\mathrm{or}\ k)$ to encode all samples with minimal error

How to optimize the objective function?

- For existing set of clusters, take sample $y \in \Gamma_k$,
- If y is movefrom Γ_k to Γ_j, Γ_k becomes $\widetilde{\Gamma_k}$, Γ_j becomes $\widetilde{\Gamma_j}$, and

$$\widetilde{m_k} = m_k + \frac{1}{N_k - 1}[m_k - y], \quad \tilde{J}_k = J_k - \frac{N_k}{N_k - 1}\|y - m_k\|^2$$

$$\widetilde{m_j} = m_j + \frac{1}{N_j + 1}[y - m_j], \quad \tilde{J}_j = J_j + \frac{N_j}{N_j + 1}\|y - m_j\|^2$$

Since

$$J_e = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \|y - m_i\|^2 = \sum_{i=1}^{c} J_i$$

If

$$\frac{N_j}{N_j + 1}\|y - m_j\|^2 < \frac{N_k}{N_k - 1}\|y - m_k\|^2$$

moving y from Γ_k to Γ_j will decrease J_e.

--- A greedy algorithm.

C-means (K-means) Algorithm

(1) Initialize $\Gamma_i, i = 1, \cdots, c$, calculate $m_i, i = 1, \cdots, c$ and J_e

(2) Pick up a sample y, e.g., $y \in \Gamma_i$

(3) If $N_i = 1$ go to step (2), otherwise continue

(4) Compute ρ_j:

$$\rho_j = \frac{N_j}{N_j + 1}\|y - m_j\|^2, \quad j \neq i$$

$$\rho_i = \frac{N_i}{N_i - 1}\|y - m_i\|^2$$

(5) Find the minimal ρ_j, say ρ_k. If $\rho_k < \rho_i$, move y from Γ_i to Γ_k

(6) Update $m_i, i = 1, \cdots, c$ and J_e

(7) If J_e has not changed in N iterations, stop; otherwise go to step (2).

C-means (K-means) Algorithm

initialize $m_i, i = 1, \cdots, c$

Do

assign samples to clusters with nearest $m_i, i = 1, \cdots, c$

recompute $m_i, i = 1, \cdots, c_i$

until (no change in $m_i, i = 1, \cdots, c$)

return $m_i, i = 1, \cdots, c$ and the cluster assignments of all samples

end

Example

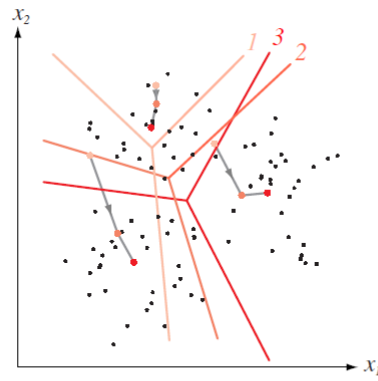
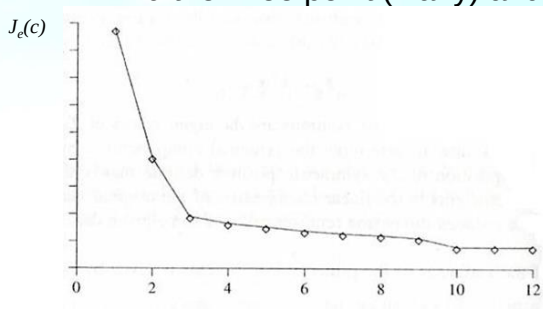


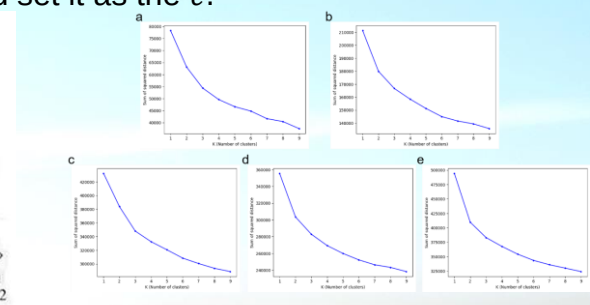
FIGURE 10.3. Trajectories for the means of the k-means clustering procedure applied to two-dimensional data. The final Voronoi tesselation (for classification) is also shown—the means correspond to the "centers" of the Voronoi cells. In this case, convergence is obtained in three iterations. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Initialization

- Usually by setting initial representative points and form initial clusters around them, e.g.,
  - Arbitrarily/randomly choose c points or set c centers
  - Randomly divide samples into c groups and use the means to start
  - Choose points with high local "density" (say, roughly estimated as the number of samples in a small neighborhood)
  - Find $c-1$ clusters and use the sample with the largest distance to all cluster centers as the cth representative. Start with 1 cluster.
  - Select according to domain knowledge/experience
  - …

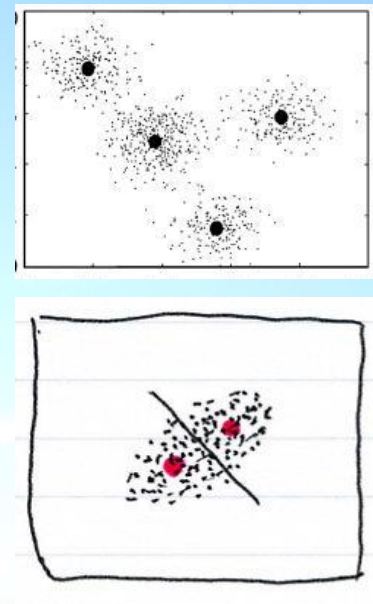
- $c = ?$ $(k = ?)$
  - Given based on domain knowledge
  - Try different choices and check the results
  - Decide through pre-experiments
    - Do clustering for $c = 1,2,3,\cdots$, draw the curve of $J_e(c)$ with regard to c, find the knee point (if any) and set it as the c.





--- This may not work, but it always worth looking at.

- Problems of C-means
  - Need to pre-select c
  - Minimal error division, with no reflection on the data distribution
  - It depends on initialization, and may fall in local minima.



13.2.2
The ISODATA algorithm

ISODATA algorithm

Iterative Self-Organizing Data Analysis Technique A

G.H. Ball & D.J. Hall, : ISODATA, a novel method of data analysis and pattern classification (Stanford Research Institute, 1965)

*With apologies for adding another acronym to the growing list, we have coined ISODATA to represent Iterative Self-Organizing Data Analysis Techniques A. (The "A" was added to make ISODATA pronouncable.) The classically-oriented can derive it from ISO, meaning "the same" or "like + Data.)

- Modified C-means
  - Batch correction
  - Allow for cluster splitting/merging
    → Provide some flexibility on the initial setting of number of clusters

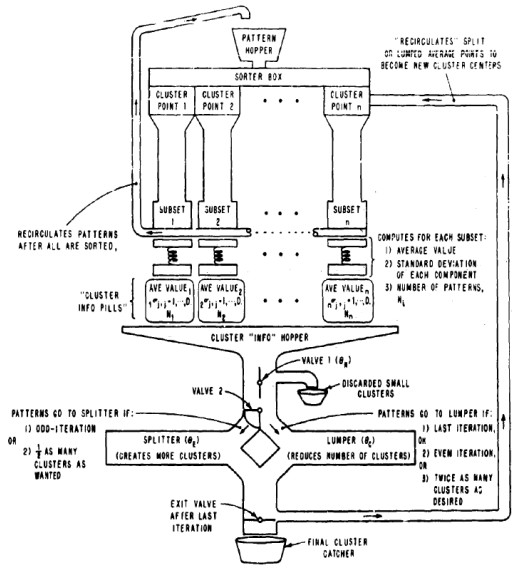


FIG. 1 A PICTORIAL DESCRIPTION OF ISODATA-POINTS

Algorithm ISODATA:

(1) Initialization: starting number of clusters c, cluster centers $m_i, i = 1, \cdots, c$, expected number of clusters k, lower limit on number of samples in a cluster θ_N, upper limit on iteration I, standard deviation threshold θ_s, merging threshold θ_c

(2) Assign all samples to their closest clusters $\Gamma_i, i = 1, \cdots, c$

(3) If $N_j < \theta_N$ for cluster Γ_j, remove this cluster and assign its samples to other close clusters, set $c = c - 1$

(4) Update the cluster centers $m_j = \frac{1}{N_j}\sum_{y \in \Gamma_j} y, j = 1, \cdots, c$

(5) Calculate the average distance of samples in a cluster to the cluster center

$$\bar{\delta_j} = \frac{1}{N_j}\sum_{y \in \Gamma_j} \|y - m_j\|, j = 1, \cdots, c$$

and the overall average distance of all clusters $\bar{\delta} = \frac{1}{N}\sum_{j=1}^{c} N_j \bar{\delta_j}$

(6) If this is the last iteration (decided by parameter I), then stop; otherwise

if $c \le k/2$, go to (7) for dividing;

if $c \ge 2k$ or the number of iteration is even, go to (8) for merging.

(7) Dividing:

7-1 For each cluster, compute standard deviation of each dimension $\sigma_j = [\sigma_{j1}, \sigma_{j2}, \cdots, \sigma_{jd}]^T$

$$\sigma_{ji} = \sqrt{\frac{1}{N}\sum_{y_k \in \Gamma_j} (y_{ki} - m_{jk})^2}, j = 1, \cdots, c; i = 1, \cdots, d$$

7-2 For each cluster j, find the dimension with largest deviation σ_{jmax}

7-3 If for $\sigma_{jmax}, j = 1, \cdots, c$, there exist a $\sigma_{jmax} > \theta_s$ (standard deviation threshold),

and $\bar{\delta}_j > \bar{\delta}$, $N_j > 2(\theta_N + 1)$, or $c \le k/2$,

then divide Γ_j to two new clusters with means m_j^+ and m_j^-, and set $c = c + 1$

$$m_j^+ = m_j + \gamma_j,\ m_j^- = m_j - \gamma_j$$

where $\gamma_j = k\sigma_{jmax},\ 0 < k < 1$.

(8) Merging

8-1 Compute distances between clusters

$$\delta_{ij} = \|m_i - m_j\|, i, j = 1, \cdots, c, i \ne j$$

8-2 Compare δ_{ij} with θ_c (merging threshold), order those smaller than θ_c

$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \cdots < \theta_c$$

8-3 Merge clusters m_{i_l} and m_{j_l}:

$$m_l = \frac{1}{N_{i_l} + N_{j_l}}\left[N_{i_l} m_{i_l} + N_{j_l} m_{j_l}\right]$$

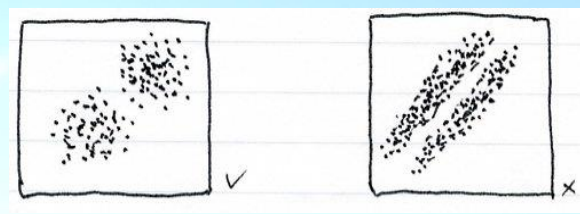
set $c = c - 1$.

(9) If this is the last iteration, stop; Otherwise go to (2). Threshold parameters can be adjusted if necessary.

13.2.3
K-Kernel Clustering

Clustering methods based on kernel similarities

- Shortcoming of C-means:
  - Using means to represent clusters, so only works for hypersphere-shaped clusters



- Solution:
  - Use kernels (instead of means) to represent clusters
  - $K_j = k(y, V_j)$ for cluster Γ_j
    - V_j is a set of parameters
    - K_j can be a function, a set of points, or a distribution model
  - C-means clustering → "C-kernels" clustering

- Define $\Delta(y, K_j)$ to be the distance of sample y to cluster Γ_j (kernel K_j), the objective function becomes

$$J_k = \sum_{i=1}^{c} \sum_{y \in \Gamma_j} \Delta(y, K_j)$$

Algorithm:

(1) Initialization: $K_j, j = 1, \cdots, c$

(2) Assign samples to clusters according to the rule:

$$\text{if } \Delta(y, K_j) = \min_{k=1,\cdots,c} \Delta(y, K_k), \text{ then } y \in \Gamma_j$$

(3) Update $K_j, j = 1, \cdots, c$. Stop if K_j does not change, otherwise go to (2).

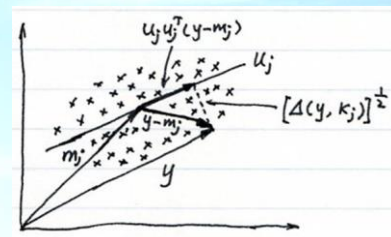
- C-means can be seen as a special case of this algorithm with K_j being m_j and Δ being the Euclidean distance.

- Examples of kernels
  - Gaussian kernel: $V_j = \{m_j, \hat{\Sigma}_j\}$

$$k_k(y, V_j) = \frac{1}{(2\pi)^{d/2}\left|\hat{\Sigma}_j\right|^{1/2}} \exp\left\{-\frac{1}{2}(y - m_j)^T \hat{\Sigma}_j^{-1}(y - m_j)\right\}$$

$$\Delta(y, k_j) = \frac{1}{2}(y - m_j)^T \hat{\Sigma}_j^{-1}(y - m_j) + \frac{1}{2}\log\left|\hat{\Sigma}_j\right|$$

Mahalanobis distance



  - Principle components

$$k(y, V_j) = U_j^T y$$

$U_j^T = [u_1, u_2, \cdots, u_{d_j}]$: Principle components of $\hat{\Sigma}_j$

$$\Delta(y, k_j) = \left[(y - m_j) - U_j U_j^T (y - m_j)\right]^T \left[(y - m_j) - U_j U_j^T (y - m_j)\right]$$

5-second break

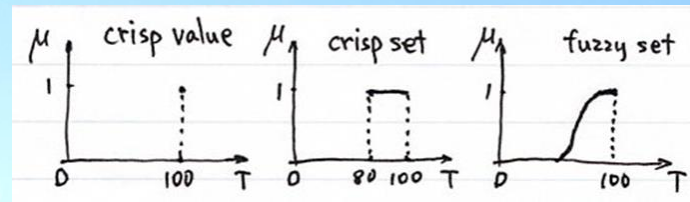
13.3
Fuzzy Clustering

Fuzzy Sets

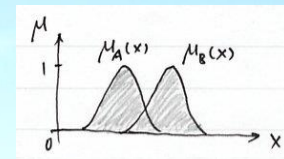
L.A. Zadeh, Fuzzy sets, *Information & Control*, v.8, pp.338-353, 1965

- Crisp sets:
  - Elements either belong to or does not belong to a set
- Fuzzy sets:
  - Elements belong to a set with a degree of membership
  - Proper for expressing concepts in natural languages
- Membership function $\mu_A(x)$:
  - the degree that x belongs to set A
  - $0 \le \mu_A(x) \le 1$
- Fuzzy set A: $A = \{(\mu_A(x_i), x_i)\}$ or $A = \bigcup_i \mu_i / x_i$

- Example: "boiled water"

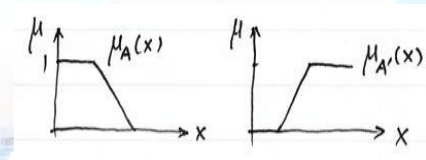


Fuzzy set operations:



- Union $C = A \cup B$: $\mu_c(x) = \max\{\mu_A(x), \mu_B(x)\}$
- Intersection $C = A \cap B$: $\mu_c(x) = \min\{\mu_A(x), \mu_B(x)\}$

- Complement $C = A'$: $\mu_c(x) = 1 - \mu_A(x)$



FCM (Fuzzy C-Means)

- What C-means does:
  - Assign n samples to c clusters, so that their average square errors to cluster means are minimal:

$$\min \quad J_e = \sum_{i=1}^{c} \sum_{y \in \Gamma_i} \|y - m_i\|^2$$

- FCM is to replace crisp clusters by fuzzy clusters:

$$\min \quad J_f = \sum_{j=1}^{c} \sum_{i=1}^{n} [\mu_j(x_i)]^b \|x_i - m_j\|^2$$

$\{\boldsymbol{x}_i, i = 1, \cdots, n\}$: the sample set, $m_j, j = 1, \cdots, c$: cluster centers:

$\mu_j(x_i)$: membership of x_i to cluster j

$b > 1$ is a parameter controlling the degree of fuzziness

- FCM

$$\min \ J_f = \sum_{j=1}^{c} \sum_{i=1}^{n} [\mu_j(x_i)]^b \|x_i - m_j\|^2$$

$$\sum_{j=1}^{c} \mu_j(\mathrm{x}_i) = 1, \quad i = 1, \cdots, n$$

Let $\partial J_f / \partial \mathrm{m}_j = 0$ and $\partial J_f / \partial \mu_j(\mathrm{x}_i) = 0$

we have

$$\mathrm{m}_j = \frac{\sum_{i=1}^{n} [\mu_j(x_i)]^b \mathrm{x}_i}{\sum_{i=1}^{n} [\mu_j(x_i)]^b}, \quad j = 1, \cdots, c$$

and

$$\mu_j(\mathrm{x}_i) = \frac{\left(1/\|\mathrm{x}_i - \mathrm{m}_j\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{c} (1/\|\mathrm{x}_i - \mathrm{m}_k\|^2)^{1/(b-1)}}, \ j = 1, \cdots, c, \ i = 1, \cdots, n$$

- FCM algorithm

(0) Set c and parameter b

(1) Initialize $m_i, i = 1, \cdots, c$

(2) Repeat till $\mu_j(x_i)$ stablized

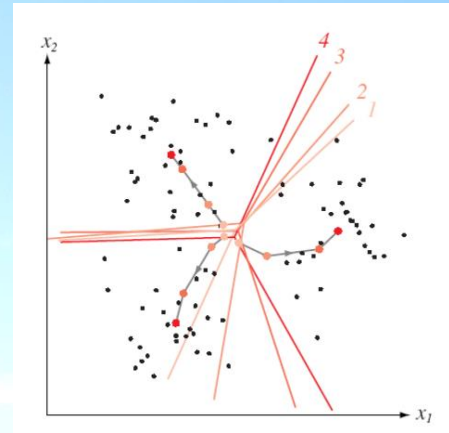
2.1 Use current m_j to compute $\mu_j(x_i)$:

$$\mu_j(\mathrm{x}_i) = \frac{\left(1/\|\mathrm{x}_i - \mathrm{m}_j\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{c}(1/\|\mathrm{x}_i - \mathrm{m}_k\|^2)^{1/(b-1)}}$$

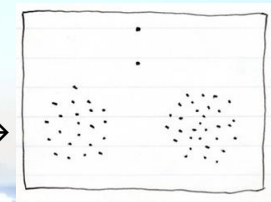
2.2 Use current $\mu_j(x_i)$ to compute m_j:

$$\mathrm{m}_j = \frac{\sum_{i=1}^{n}[\mu_j(x_i)]^b \mathrm{x}_i}{\sum_{i=1}^{n}[\mu_j(x_i)]^b}$$

(3) Defuzzify the clusters if necessary



- Since FCM requires $\sum_{j=1}^{c}\mu_j(x_i) = 1$, noisy samples that should not belong to any cluster (or should belong to a new cluster) cannot get very small memberships, e.g., →



AFC

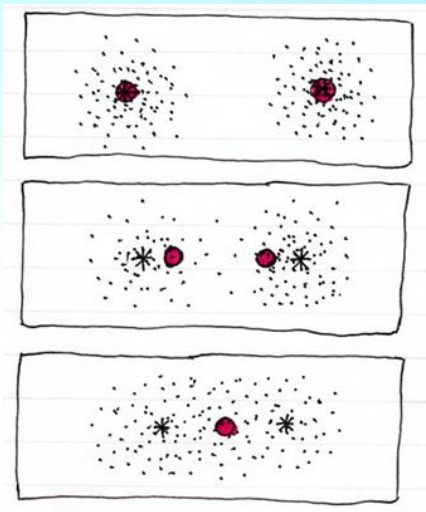
- Change the condition to $\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_j(x_i) = n$

then $$\mathrm{m}_j = \frac{\sum_{i=1}^{n}[\mu_j(x_i)]^b \mathrm{x}_i}{\sum_{i=1}^{n}[\mu_j(x_i)]^b}, \quad j = 1, \cdots, c$$

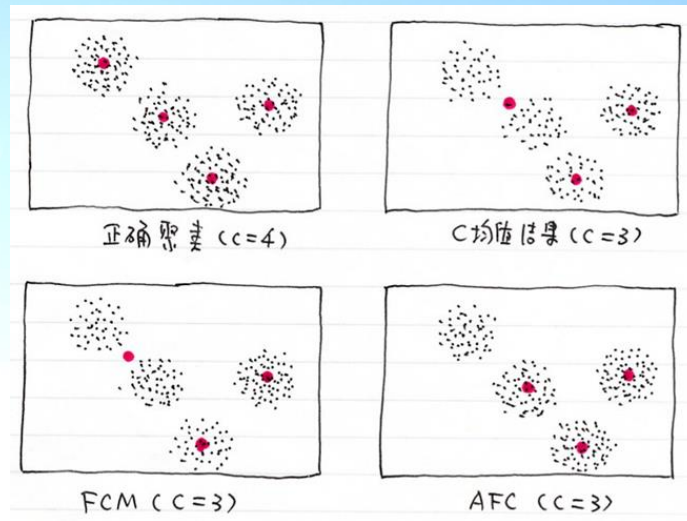
$$\mu_j(\mathrm{x}_i) = \frac{n\left(1/\|\mathrm{x}_i - \mathrm{m}_j\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{c}\sum_{l=1}^{n}(1/\|\mathrm{x}_l - \mathrm{m}_k\|^2)^{1/(b-1)}}, \; j = 1, \cdots, c, \; i = 1, \cdots, n$$

- AFC has the same algorithm procedures with FCM
  - Since $\mu_j(x_i)$ may be >1 in AFC, it's not a membership function in the strict sense, normalization can be taken if necessary.

• Examples (+: C-means, ×: FCM, ●: AFC)



When clusters are merging



When cluster counts are given wrong

单选题 1分

设置

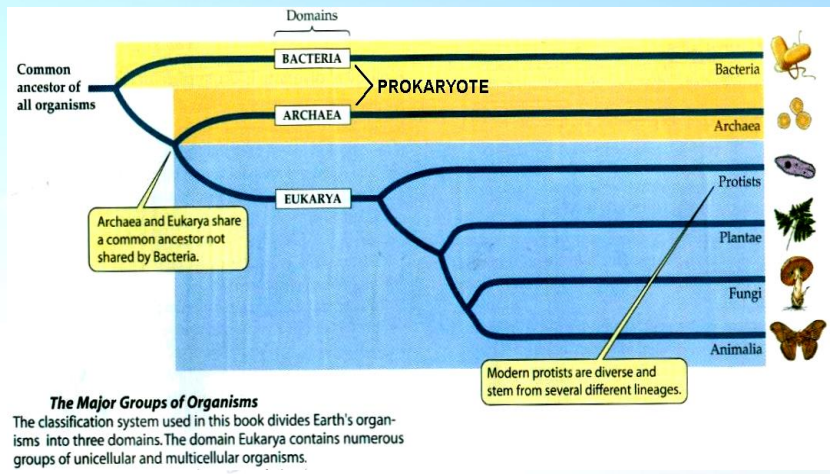
4-min break

A Yes, I'm back to my computer.

B Sorry, not yet.

提交

13.4
Hierarchical Clustering



The Major Groups of Organisms
The classification system used in this book divides Earth's organisms into three domains. The domain Eukarya contains numerous groups of unicellular and multicellular organisms.

Hierarchical Clustering

- Idea
  - Start from singletons, bottom-up
  - Merging clusters of highest similarity, till all in one
  - Cut the tree at a certain level to decide cluster counts
- Result

Hierarchical tree → **Dendrogram**

- Branch length: similarity between daughter nodes
- ➢ Caution: Order of nodes switchable as long as tree structure unchanged

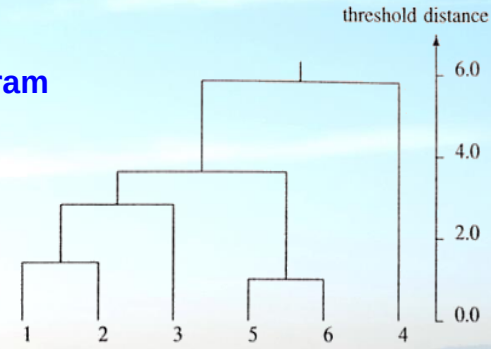
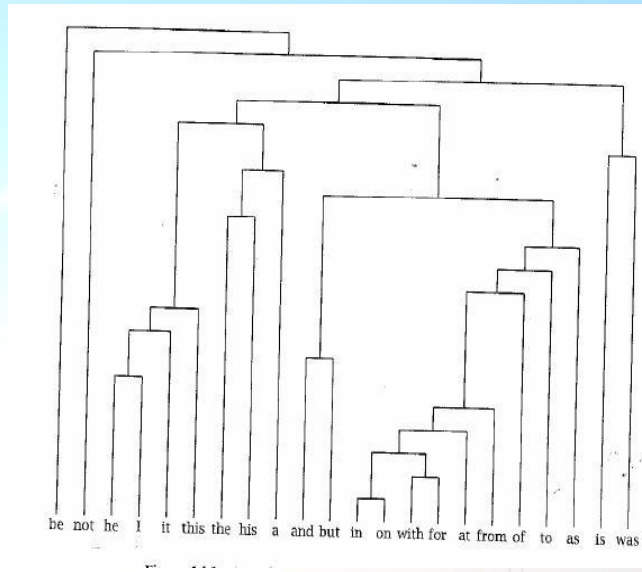
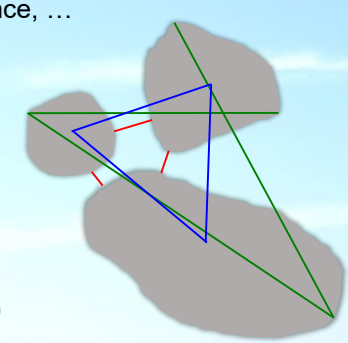


Figure 10.1 Dendrogram



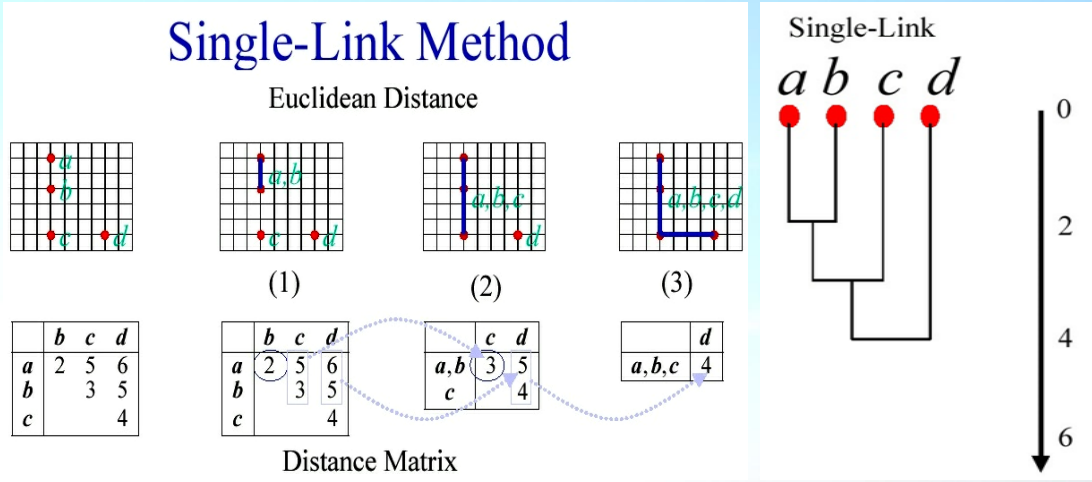
What are the key factors in hierarchical clustering?

- Distance / similarity measure
  - Distance between samples
    - e.g. Euclidean distance, correlation, City Block distance, …
  - Distance between clusters (linkage)
    - Single-linkage $\Delta(\Gamma_i, \Gamma_j) = \min_{\substack{y\in\Gamma_i \\ \tilde{y}\in\Gamma_j}} \delta(y, \tilde{y})$
    - Complete-linkage $\Delta(\Gamma_i, \Gamma_j) = \max_{\substack{y\in\Gamma_i \\ \tilde{y}\in\Gamma_j}} \delta(y, \tilde{y})$
    - Average-linkage $\Delta(\Gamma_i, \Gamma_j) = \delta(m_i, m_j)$

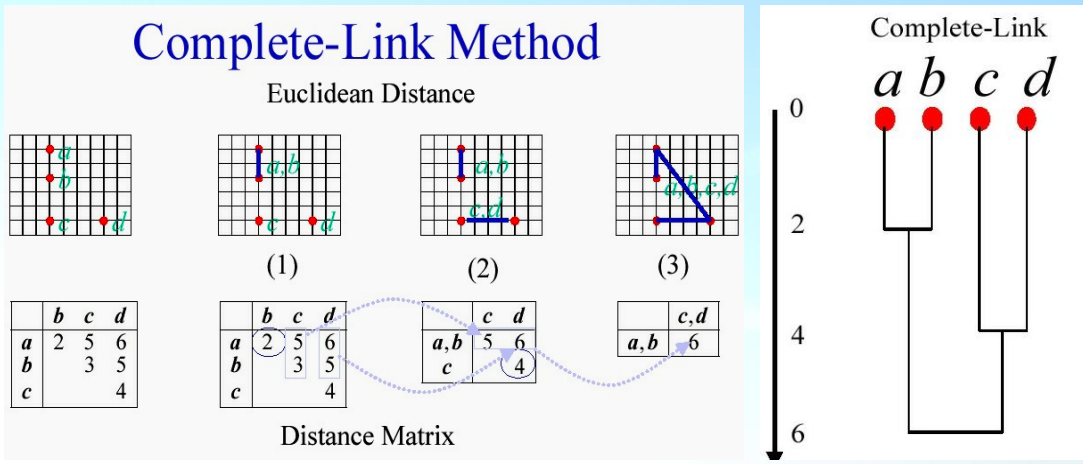


- The bottom-up algorithm (agglomerative algorithm)
  - (1) Initialization, every sample as one cluster
  - (2) Merge the two clusters with smallest distance
  - (3) Repeat (2), until all samples are merged into one reservoir

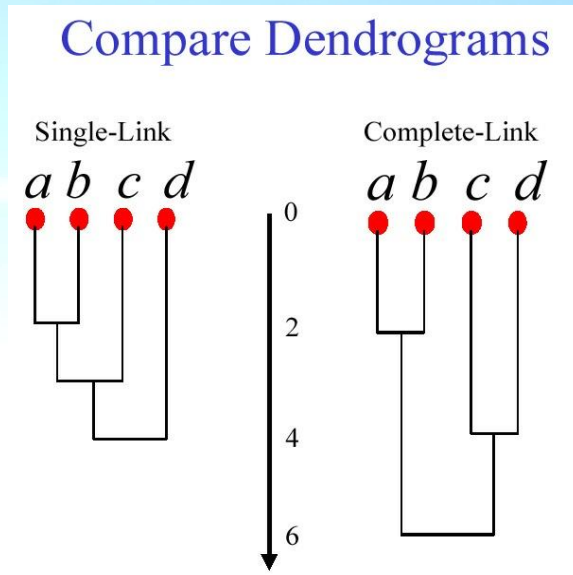
- The effect of linkage on results



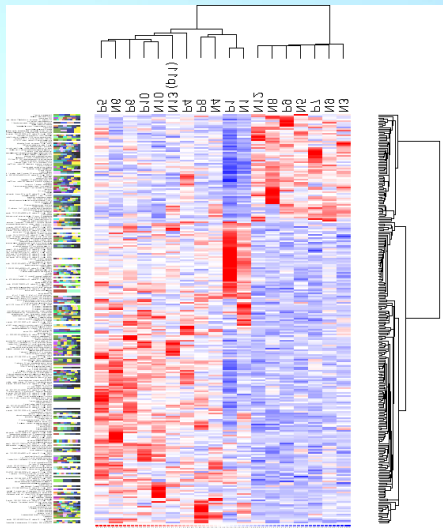
- The effect of linkage on results



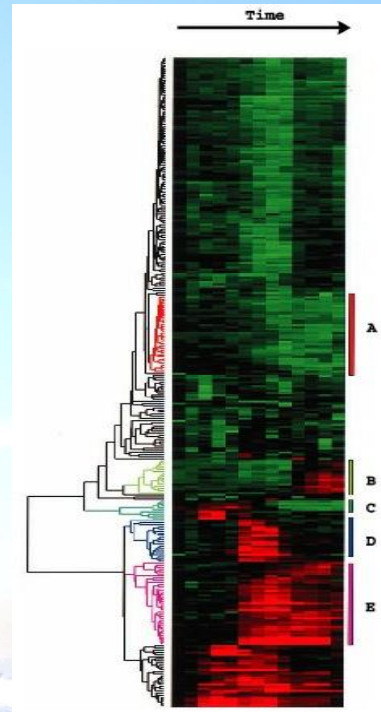
- The effect of linkage on results



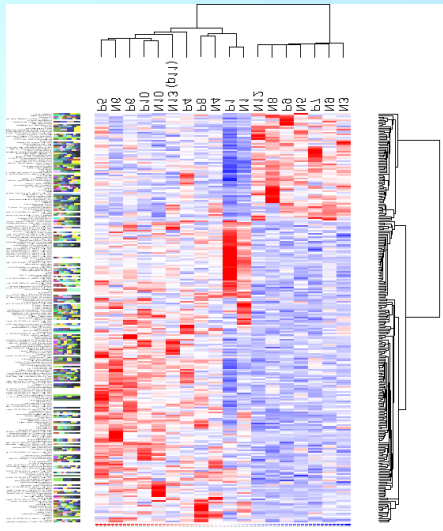
Hierarchical clustering and heatmaps



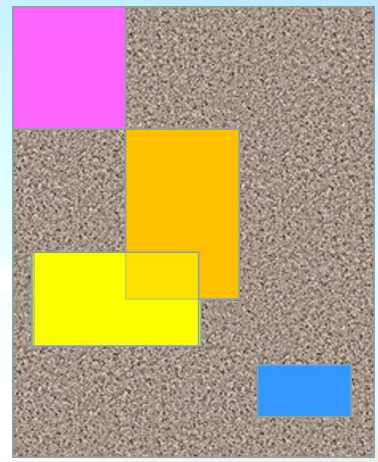
Two-way clustering



Two-way clustering vs. *biclustering*



Two-way clustering



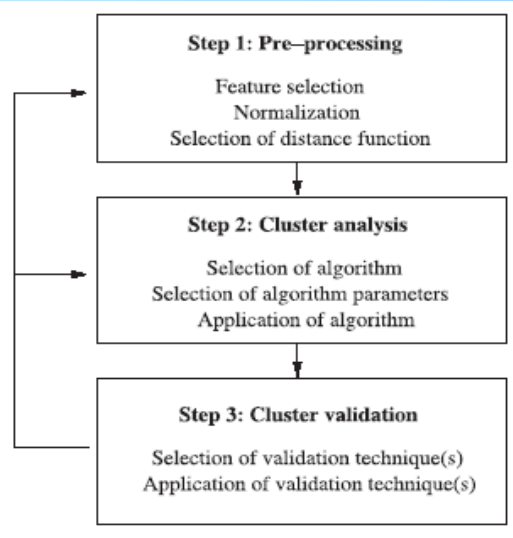
Biclustering

5-second break

13.5
Assessment of Clustering Results

Validating Unsupervised Learning Results

- External measures
  - According to some external knowledge
  - Caution for bias and subjectivity
- Internal measures
  - Quality of the clusters according to the data
  - Compactness and separation
  - Stability
  - ...



J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, *Bioinformatics*, **21**(15): 3201-3212, 2005

Comparing Clustering Results

- Confusion Matrix

Partition 1 / Partition 2	C1	C2	C3	C4
S1	23	38	122	0
S2	309	12	0	13
S3	0	0	3	98

- F-measure (the larger the closer)

$$F(C) = \sum_{t=1,\cdots,T} \frac{N_t}{N} \max_{k=1,\cdots,K} F(S_t, C_k)$$

$$F(S_t, C_k) = \frac{2P(S_t,C_k)R(S_t,C_k)}{P(S_t,C_k)+R(S_t,C_k)} \quad \text{or} \quad F(S_t, C_k) = \frac{(b^2+1)P(S_t,C_k)R(S_t,C_k)}{b^2P(S_t,C_k)+R(S_t,C_k)}$$

$$P(S_t, C_k) = \frac{N_{tk}}{N_k},\ R(S_t, C_k) = \frac{N_{tk}}{N_t}$$

If the comparison is equal

If S_t is regarded as the correct answer

Comparing Clustering Results

- Rand Index (the larger the closer)

$$R(U,V) = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

a, b, c, d: Numbers of pairs in the same/different clusters in partition U and in the same/different clusters in partition V.

a: same-same, b: diff-diff, c: same-diff, d: diff-same

- Adjusted Rand Index (ARI) (the larger the closer)

$$\text{ARI} = \frac{\text{Actual Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}}$$

$$R(U,V) = \frac{\sum_{lk}\binom{n_{lk}}{2} - \left[\sum_l \binom{n_{l.}}{2}\sum_k \binom{n_{.k}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_l \binom{n_{l.}}{2} + \sum_k \binom{n_{.k}}{2}\right] - \left[\sum_l \binom{n_{l.}}{2}\sum_k \binom{n_{.k}}{2}\right]/\binom{n}{2}}$$

Comparing Clustering Results

- Minkovski Score (the smaller the closer)

$$M(U,V) = \frac{\|C_U - C_V\|}{\|C_U\|}$$

C_U, C_V: cophenetic matrixes of partition U and V

	X1	X2	X3	.	.	.
X1	1	0	1			
X2	0	1	0			
X3	1	0	1			
.				1		
.					1	
.						1

cophenetic matrix of all samples

- Stability (the larger the better)
  - Data split into two halves, and both clustered
  - Check the agreement between the partitioning and prediction

Internal measures

R

- Compactness (homogeneity)

$$V(C) = \sqrt{\frac{1}{N}\sum_{C_k \in C}\sum_{i \in C_k}\delta(i, \mu_k)}$$

  - Intra-cluster variance
  - Smaller better

- Connectedness (connectivity)

$$\text{Conn}(C) = \sum_{i=1}^{N}\sum_{j=1}^{L} x_{i,nn_{i(j)}}$$

  - Degree that nearest L neighbors not clustered together
  - Smaller better

- Separation
  - Silhouette value:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}$$

    - Confidence for sample i be in the cluster comparing to closest other cluster
    - Larger better
  - Dunn Index

$$D(C) = \min_{C_k \in C}\left(\min_{C_l \in C}\frac{dist(C_k, C_l)}{\max_{C_m \in C} diam(C_m)}\right)$$

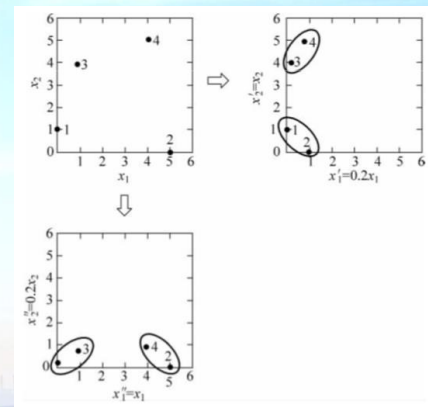
    - Ratio between the smallest inter-cluster distance and the largest intra-cluster distance
    - Larger better

J. Handl, J. Knowles, D.B. Kell, *Bioinformatics*, 2005

Discussion

D

- Special challenges for unsupervised clustering
  - Need to justify whether there are clusters in the data
  - Usually unknown cluster numbers
  - **Clustering results need to be annotated afterwards**
    - "Pretended" unsupervised tasks → with expected truth
    - "Real" unsupervised tasks → no grand truth
  - Scaling factor
  - A distribution vs. a closed region
    → fuzzy clustering
    →→ mixture models



Homework

- Computer exercises (Ex6)
  - Experiment on generating Gaussian distribution synthetic data and K-means clustering in different situations.
- Deadline:
  - Dec. 1 (Wednesday), 23:00

单选题 1分 设置

4-min break

A Yes, I'm back to my computer.

B Sorry, not yet.

提交