

# Clustering

Caso di studio di Metodi Avanzati di  
Programmazione (Corso A)

AA 2023-2024

# Data Mining

Lo scopo del **data mining** è l'*estrazione* (semi) automatica di *conoscenza* nascosta in voluminose basi di dati al fine di renderla disponibile e direttamente utilizzabile



# Clustering

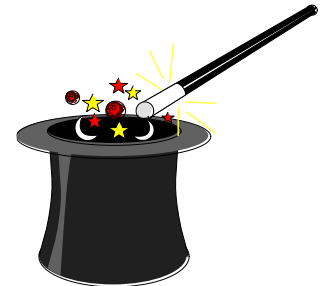
## *Dati:*

- una collezione  $D$  di transazioni dove, ogni transazione è un vettore valori misurati per una collezione di attributi numerici;
- un intero  $k$ ;

## *Lo scopo è:*

- partizionare  $D$  in  $k$  insiemi di transazioni  $D_1, \dots, D_k$ , tale che:
  - $D_i$  ( $i=1, \dots, k$ ) è un segmento (selezione) omogenea di  $D$ ;

- $D = \bigcup_{i=1}^k D_i$  and  $D_i \cap D_j = \Phi$  .



# Distanza tra esempi

- Assumendo che un esempio è un vettore di numeri reali
- La distanza tra due esempi è calcolata tramite la distanza Euclidea
- Esempio,  $t1=[1,3,0]$  ;  $t2=[0,1,0]$   
$$D(t1,t2)=(1-0)^2+(3-1)^2+(0-0)^2=1+4+0=5$$

# Distanza tra cluster

- Dati due cluster di esempi,

Per esempio  $C1 = \{[1,0,3]; [0,0,0]\}$ ,  $C2 = \{[5,0,2]; [0,0,1], [1,1,1]\}$

Distanza single-link

$$D(C1, C2) = \min_{(t1 \in C1, t2 \in C2)} dist(t1, t2)$$

Distanza average-link:

$$D(C1, C2) = \frac{\sum_{t1 \in C1, t2 \in C2} d(t1, t2)}{\#(C1 \times C2)}$$

# Single-link

$$D(C1, C2) = \min_{(t1 \in C1, t2 \in C2)} dist(t1, t2)$$

- Dati due cluster di esempi,

Per esempio  $C1 = \{t1=[1,0,3]; t2=[0,0,0]\}$ ,  $C2 = \{t3=[5,0,2]; t4=[0,0,1]; t5=[1,1,1]\}$

$$D(t1, t3) = 16 + 0 + 1 = 17$$

$$D(t1, t4) = 1 + 0 + 4 = 5$$

$$D(t1, t5) = 0 + 1 + 4 = 5$$

$$D(t2, t3) = 25 + 0 + 4 = 29$$

$$D(t2, t4) = 0 + 0 + 1 = 1$$

$$D(t2, t5) = 1 + 1 + 1 = 3$$

min=1

# Average-link

$$D(C1, C2) = \frac{\sum_{t1 \in C1, t2 \in C2} d(t1, t2)}{\#(C1 \times C2)}$$

- Dati due cluster di esempi,

Per esempio  $C1 = \{t1=[1,0,3]; t2=[0,0,0]\}$ ,  $C2 = \{t3=[5,0,2]; t4=[0,0,1]; t5=[1,1,1]\}$

$$D(t1, t3) = 1 + 6 + 0 + 1 = 17$$

$$D(t1, t4) = 1 + 0 + 4 = 5$$

$$D(t1, t5) = 0 + 1 + 4 = 5$$

$$D(t2, t3) = 2 + 5 + 0 + 4 = 29$$

$$D(t2, t4) = 0 + 0 + 1 = 1$$

$$D(t2, t5) = 1 + 1 + 1 = 3$$

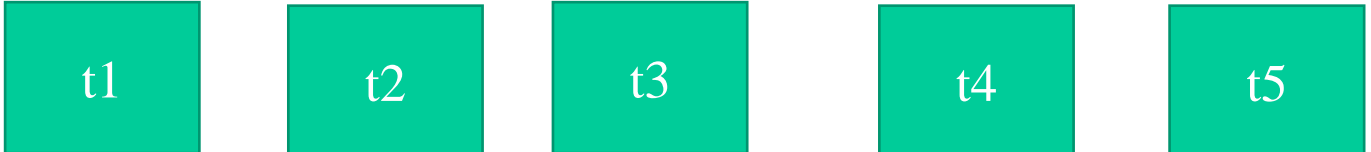
$$\text{avg} = (17 + 5 + 5 + 29 + 1 + 3) / 6 = 10.0$$

# Clustering agglomerativo

- Con distanza tra cluster single-link

Livello 0: un cluster per ogni esempio

	X	Y
t1	1	2
t2	0	1
t3	1	3
t4	2	2
t5	2	1





# Clustering agglomerativo

	X	Y
t1	1	2
t2	0	1
t3	1	3
t4	2	2
t5	2	1

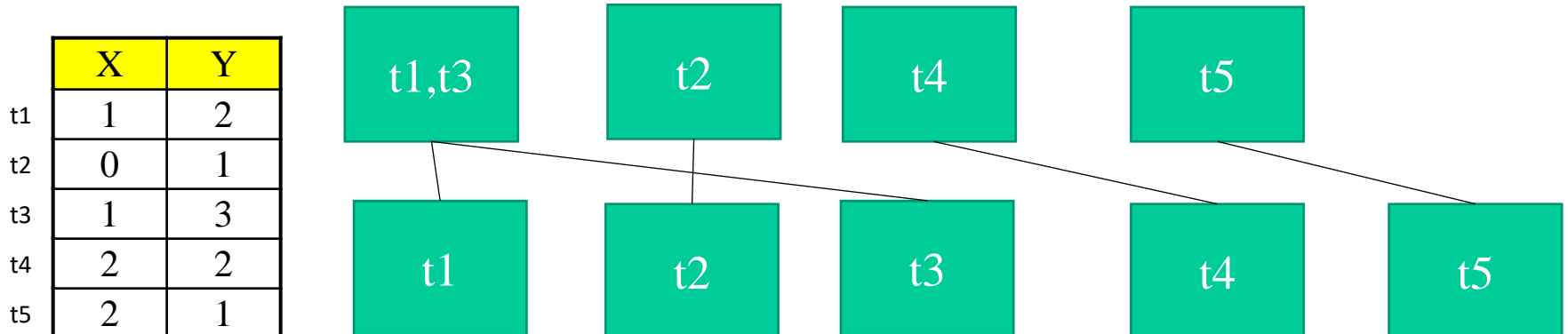
Distanza Euclidea	t1	t2	t3	t4	t5
t1		2	1	1	2
t2			5	5	4
t3				2	5
t4					1
t5					

# Clustering agglomerativo

Distanza Euclidea	t1	t2	t3	t4	t5
t1		2	1	1	2
t2			5	5	4
t3				2	5
t4					1
t5					

# Clustering agglomerativo

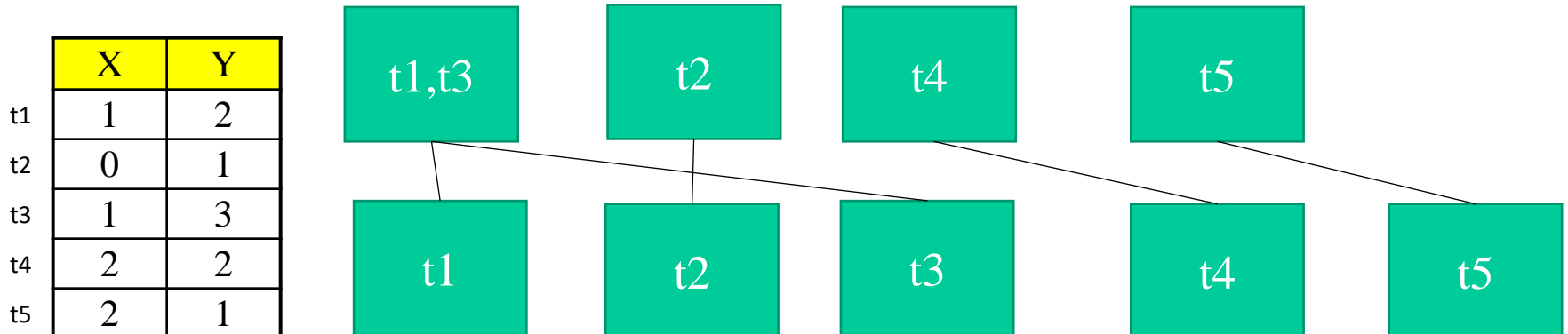
- Con distanza tra cluster single-link  
fondo i due cluster più vicini al livello 0 e creo il nuovo cluster set al livello 1



# Clustering agglomerativo

- Con distanza tra cluster single-link

Livello 1: cerco i due cluster a livello 1 più vicini



# Clustering agglomerativo

$\text{singleLink}(\{t1, t3\}, \{t2\}) = \min(d(t1, t2), d(t3, t2)) = \min(2, 5) = 2$

$\text{singleLink}(\{t1, t3\}, \{t4\}) = 1$

$\text{singleLink}(\{t1, t3\}, \{t5\}) = 2$

$\text{singleLink}(\{t2\}, \{t4\}) = 5$

$\text{singleLink}(\{t2\}, \{t5\}) = 4$

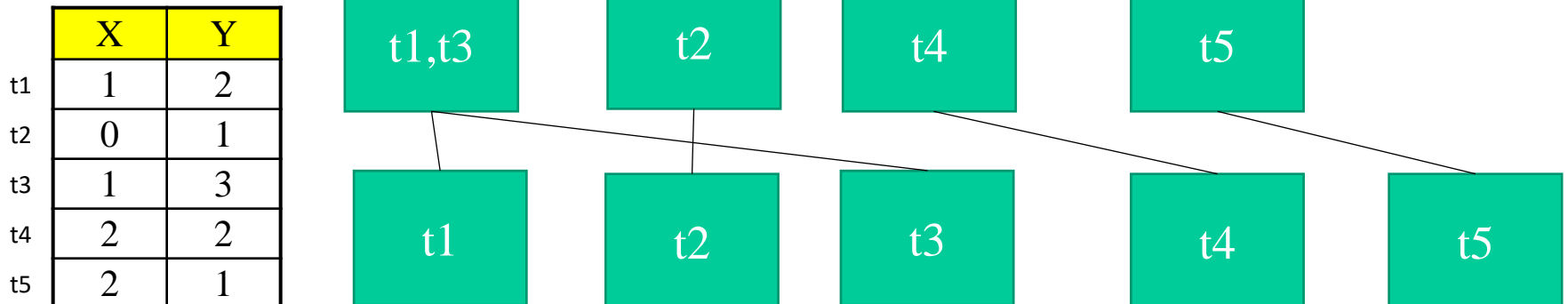
$\text{singleLink}(\{t4\}, \{t5\}) = 1$

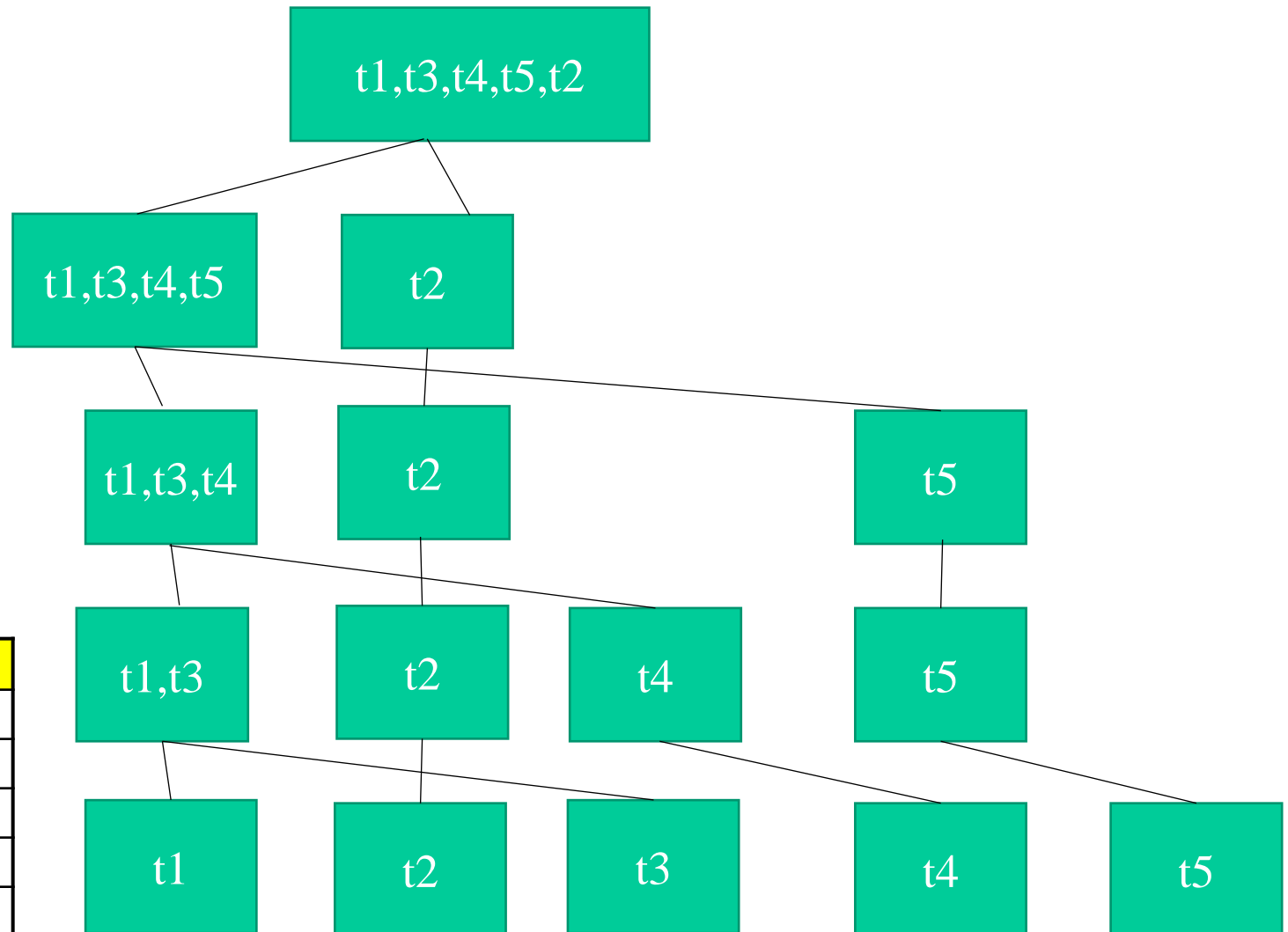
Distanza Euclidea	t1	t2	t3	t4	t5
t1		2	1	1	2
t2			5	5	4
t3				2	5
t4					1
t5					

# Clustering agglomerativo

## Con distanza tra cluster single-link

fondo i due cluster più vicini al livello 1 e creo il nuovo cluster set al livello 2





	X	Y
t1	1	2
t2	0	1
t3	1	3
t4	2	2
t5	2	1

## Caso di studio

Progettare e realizzare un sistema **client-server** denominato “H-CLUS”.

Il server include funzionalità di **data mining** per la scoperta di un dendrogramma di cluster di dati con algoritmo di clustering agglomerativo.

Il client è un applicativo Java che consente di usufruire del servizio di scoperta remoto e visualizza la conoscenza (cluster) scoperta



# Istruzioni

1. Il progetto dello A.A. 2023-24, denominato H-CLUS, è valido solo per coloro che superano la prova scritta o prove in itinere entro il corrente A.A.
2. Ogni progetto può essere svolto da gruppi di **al più TRE** (3) studenti.
3. Coloro i quali superano la prova scritta devono consegnare il progetto **ENTRO** la data prevista per la corrispondente prova orale (da sito web degli appelli del corso di laurea). La verbalizzazione avrà luogo in data successiva alla consegna (la data verrà comunicata su esse3 dopo la consegna del progetto).
4. La discussione del progetto avverrà alla sua consegna, *ad personam* per ciascun componente del gruppo. Il voto massimo della prova scritta è 33. Un voto superiore a 30 equivale a 30 e lode.
5. Il voto finale sarà stabilito sulla base del voto attribuito allo scritto e al progetto.



## Istruzioni

Non si riterrà sufficiente, e come tale non sarà corretto, un progetto non sviluppato in tutte le su parti (client-server, client, accesso al db, serializzazione,...)

# Valutazione

Diagramma delle classi (2 punti)

JavaDoc (3 punti)

Guida di installazione (con Jar+ Bat+ Script SQL) (2 punti)

Guida utente con esempi di test (2 punti)

Sorgente del sistema (14 punti)

Estensioni del progetto svolto in laboratorio (10 punti)