Clustering

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

AA 2022-2023

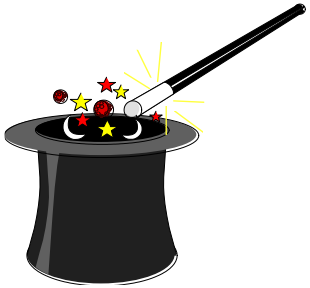
Data Mining

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 di coppie attributo-valore (item);
- un intero k;

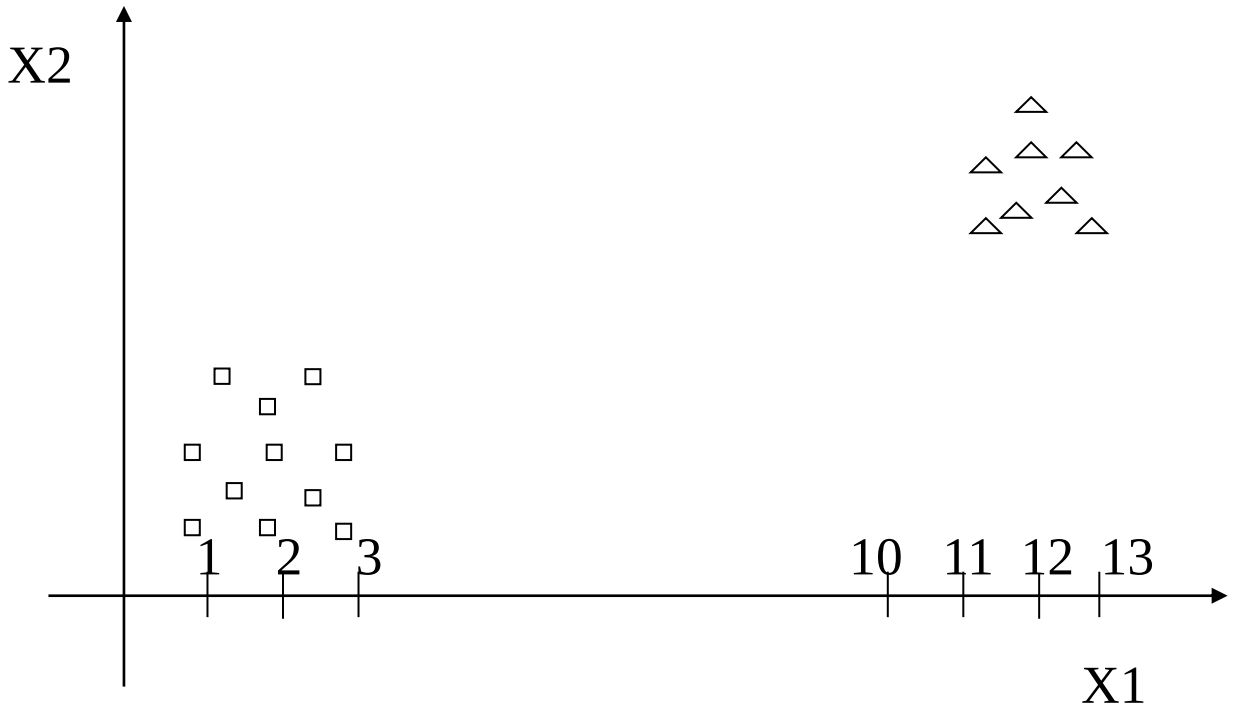


Lo scopo è:

- partizionare D in k insiemi di transazioni $D_1,\ldots D_k$, tale che:
  - D_i (i=1,…,k) è un segmento (selezione) omogenea di D;
  - $D = \bigcup_{i=1}^{k} D_i$ and $D_i \bigcap D_j = \Phi$.

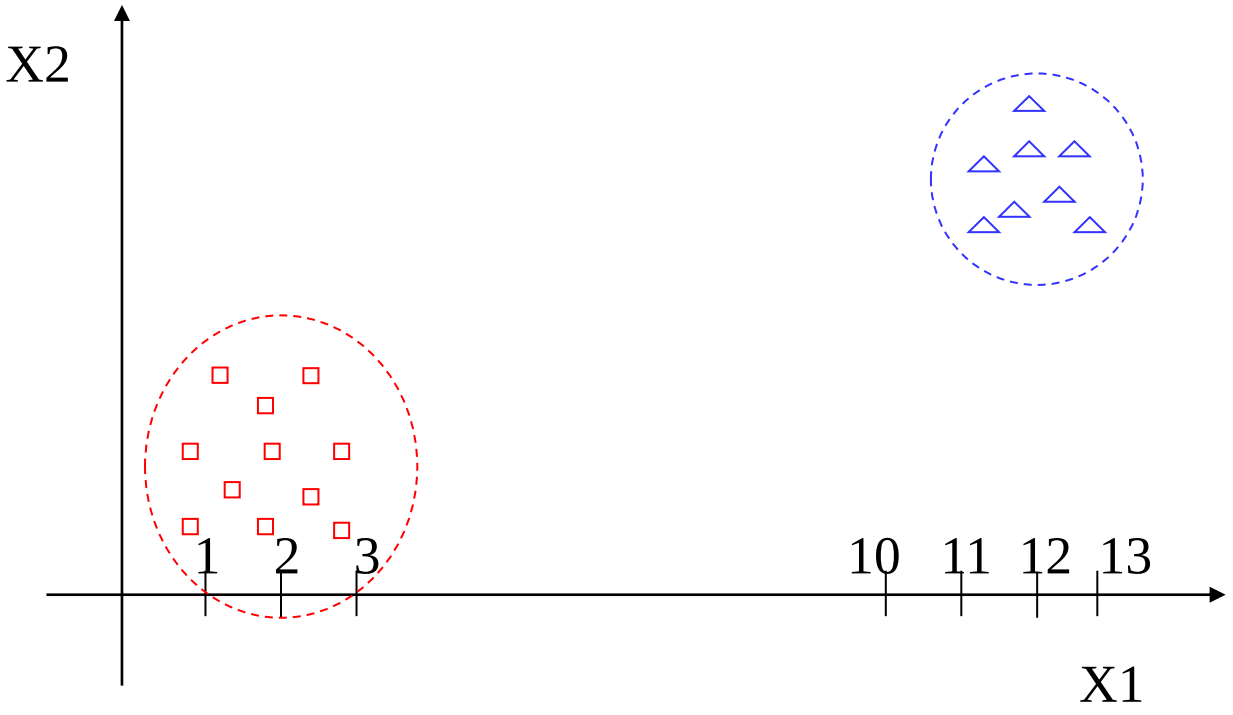
Clustering

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3



Clustering

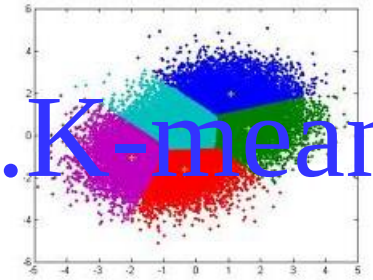
X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3



Problemi

1. Come eseguo il clustering?
   - *K-means.*
2. Come rappresento i cluster?
   - *Calcolare e memorizzare i centroidi dei cluster.*
3. Come uso i cluster in applicazioni reali?
   - *Minimizzare la distanza tra una transazione nuova e la rappresentazione dei cluster (centroidi) per scoprire il cluster di appartenenza.*

1.K-means



http://it.wikipedia.org/wiki/K-means

Kmeans(D,k)-:clusterSet

clusterSet: insieme di k segmenti D_i : ogni segmento D_i è un insieme di transazioni in *D*

begin

1. inizializza *clusterSet* con segmenti inizialmente vuoti
2. assegna a ciascun segmento di *clusteSet* una transazione casualmente scelta da *D*
3. **do**

    for(*transazione*:*D*)

        3.1 D_i=cluster(clusterSet,transazione)

        3.2 sposta *transazione* nel segmento D_i

        3.4 ricalcola I semi dei cluster come centroidi dei cluster

    while (almeno una transazione cambia cluster)

4. **return** *clusterSet;*

end

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 1: inizializzazione dei k sementi (insiemi vuoti)

clusterSet={D_1,D_2}

D1 = {}

D2 = {}

X1	X2
0.9	1
0.9	**1.2**
1.3	2
1.2	3.7
1.9	1
2	**2.2**
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 2: inizializzazione dei centroidi

Si scelgono k transazioni (centroidi) in maniera CASUALE e le si inseriscono nei segmenti: un centroide per segmento.

clusterSet={D_1,D_2}

c1= (0.9 , 1.2) : D1 = D1 ∪ c1

c2=(2,2.2) : D2 = D1 ∪ c2

X1	X2
0.9	1
0.9	**1.2**
1.3	2
1.2	3.7
1.9	1
2	**2.2**
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 3: assegno ciascuna transazione al *suo* cluster

L'appartenenza di una transazione ad un cluster dipende dalla distanza della transazione dal centroide del cluster.

Si sceglie di spostare la transazione nel cluster che minimizza tale distanza.

X1	X2
0.9	1
0.9	**1.2**
1.3	2
1.2	3.7
1.9	1
2	**2.2**
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 3: assegno ciascuna transazione al *suo* cluster

L'appartenenza di una transazione a in cluster dipende dalla distanza della transazione dal centroide del cluster.

Si sceglie di spostare la transazione nel cluster che minimizza tale distanza.

EuclideanDist((0.9,1),(0.9,1.2))=0.2

EuclideanDist((0.9,1),(2, 2.2))=1.62

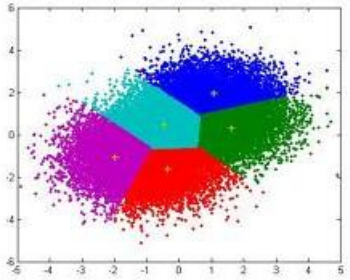
X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 3: assegno ciascuna transazione al *suo* cluster

L'appartenenza di una transazione a in cluster dipende dalla distanza della transazione dal centroide del cluster.

Si sceglie di spostare la transazione nel cluster che minimizza tale distanza.



clusterSet={D_1,D_2}

D_1={1,2,5,8}

D_2={3,4,6,7,9,10,11,12,1314,15,16,17}

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 4: ricalcolo i centroidi dei cluster

Il centroide è una transazione fittizia del segmento che ad ogni attributo associa il valore medio (moda) calcolato sul segmento

clusterSet={D_1,D_2}

c1= (1.65, 1.05) dove:

$$\frac{0.9+0.9+1.9+2.9}{4}=1.65$$

$$\frac{1+1.2+1+1}{4}=1.05$$

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 4: ricalcolo i centroidi dei cluster

Il centroide è una transazione fittizia del segmento che ad ogni attributo associa il valore medio (moda) calcolato sul segmento

clusterSet={D_1,D_2}

c1= (1.65, 1.05)

c2=(8.03, 4.66)

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 5: ci sono transazioni che hanno cambiato cluster di appartenenza?

ripeto PASSO 3 con

c1= (1.65, 1.05)

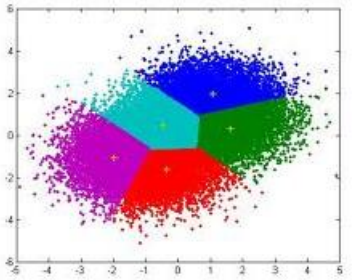
c2=(8.03, 4.66)

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 3: assegno ciascuna transazione al cluster più vicino

clusterSet={D_1,D_2}



D_1 = {1,2,3,4,5,6,7,8,9}

D_2 = {10,11,12,13,14,15,16,17}

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 4: calcolo i centroidi dei nuovi cluster

clusterSet={D1,D2}

c1= (1.76,1.98)

c2=(11.9, 5.875)

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 5: ci sono transazioni che hanno cambiato il cluster di appartenenza? SI

ripeto PASSO 3 con:

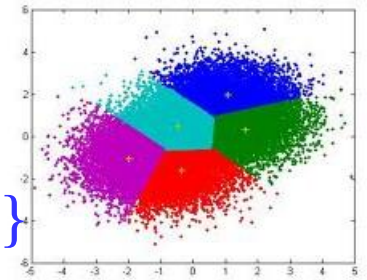
c1= (1.76,1.98)

c2=(11.9, 5.875)

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: come?

PASSO 3: assegno ciascuna transazione al cluster più vicino.



$D_1 = \{1,2,3,4,5,6,7,8,9\}$

$D_2 = \{10,11,12,13,14,15,16,17\}$

PASSO 4: calcolo i centroidi dei nuovi cluster

c1= (1.76,1.98) c2=(11.9, 5.875)

PASSO 5: ci sono transazioni che hanno cambiato il cluster di appartenenza? No!!

2.Rappresentazione di un cluster

1) Descrizione estensionale (elenco delle transazioni nel cluster).

Cluster 1

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1

Cluster 2

X1	X2
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

Rappresentazione di un cluster

2) Descrizione intensionale (tramite i centroidi del cluster).

$$X_{centroide} = \begin{cases} \dfrac{\sum_{(\ldots,x_i,\ldots)\in cluster} x_i}{|cluster|} & \text{se X è attributo numerico} \\ \underset{x_{ii}}{\operatorname{argmax}}\ frequency(x_i, cluster) & \text{se X è attributo categorico} \end{cases}$$

Cluster 1

(1.76, 1.98)

Cluster 2

(11.9, 5.875)

Calcolo di un centroide: come?

Genere	Nazionalità	Età
F	Italiana	25
F	Italiana	27
F	Italiana	34
F	Inglese	23
M	Americana	29

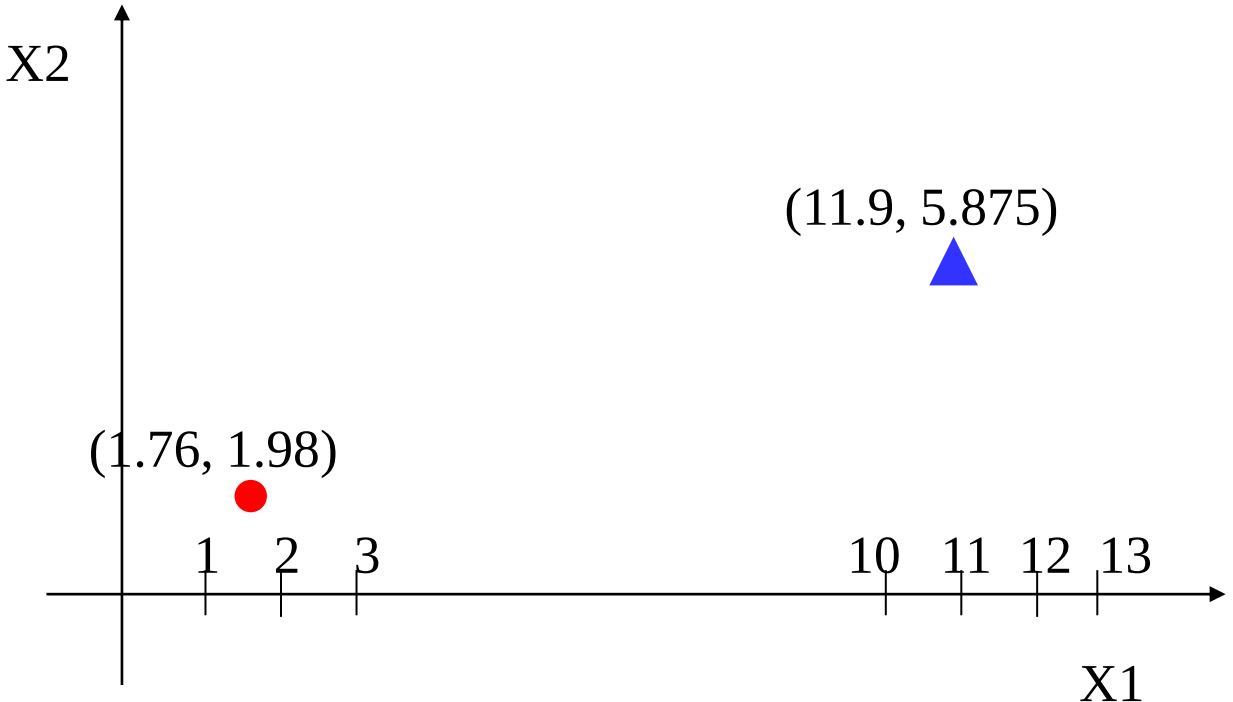
centroide

F	Italiana	27.6

3.Cluster e/o centroidi: applicazioni reali

Vantaggi:

1. Compatta in termini di spazio di memoria (memorizzo una singola transazione piuttosto che un insieme di transazioni)



3.Cluster e/o centroidi: applicazioni reali

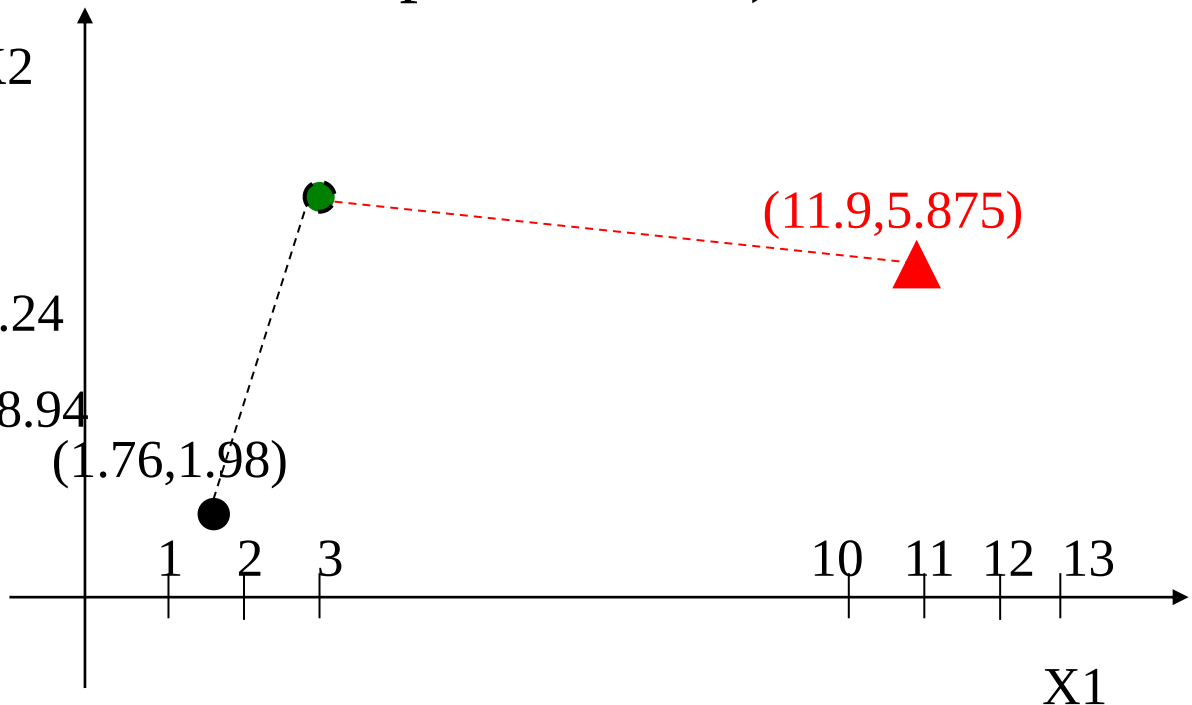
Vantaggi:

2. Posso usare i centroidi dei cluster per individuare il segmento a cui plausibilmente appartiene una nuova transazione (scelgo il centroide più vicino!!)

(3,5) ?

EuclidDist((3,5),(1.76,1.98))=3.24

EuclidDist((3,5),(11.9,5.875))=8.94



4.Qualità dei cluster: Silhouette

https://en.wikipedia.org/wiki/Silhouette_(clustering)

Valida la consistenza del modello di cluster con i dati

misura quanto è consistente il dato con il cluster a cui è

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \qquad -1 \leq s(i) \leq 1$$

Con:

- b(i) la più piccola distanza media di i dagli esempi di qualsiasi altro cluster diverso da quello in cui è inserito i
- a(i) la distanza media di i dai punti raggruppati nel medesimo cluster di i

4.Qualità dei cluster: come scegliere k?

k= 2 calcolo avg Silhouette

k=3 calcolo avg Silhouette

…

Scelgo k che rappresenta un massimo locale.

Caso di studio

Progettare e realizzare un sistema client-server denominato “K-MEANS”.

Il server include funzionalità di data mining per la scoperta di cluster di dati.

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. 2022-23, denominato K-MEANS, 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, interfaccia 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)