Clustering: Quality Threshold

Caso di studio di Metodi Avanzati di Programmazione

AA 2024-2025

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

Aree di Applicazione

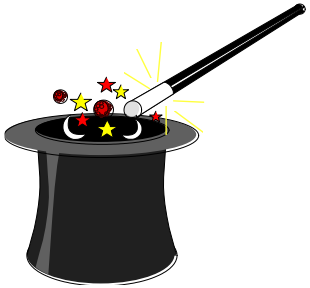
1. **previsione**
   utilizzo di valori noti per la previsione di quantità non note (es. stima del fatturato di un punto vendita sulla base delle sue caratteristiche)
2. **classificazione**
   individuazione delle caratteristiche che indicano a quale gruppo un certo caso appartiene (es. discriminazione tra comportamenti ordinari e fraudolenti)
3. **segmentazione** (o clustering)
   individuazione di gruppi con elementi omogenei all'interno del gruppo e diversi da gruppo a gruppo (es. individuazione di gruppi di consumatori con comportamenti simili)
4. **associazione**
   individuazione di elementi che compaiono spesso assieme in un determinato evento (es. prodotti che frequentemente entrano nello stesso carrello della spesa)
5. **sequenze**
   individuazione di una cronologia di associazioni (es. percorsi di visita di un sito web)

   …

Clustering

Dati:

- una collezione D di transazioni dove, ogni transazione è un vettore di coppie attributo-valore (item);
- un intero k;

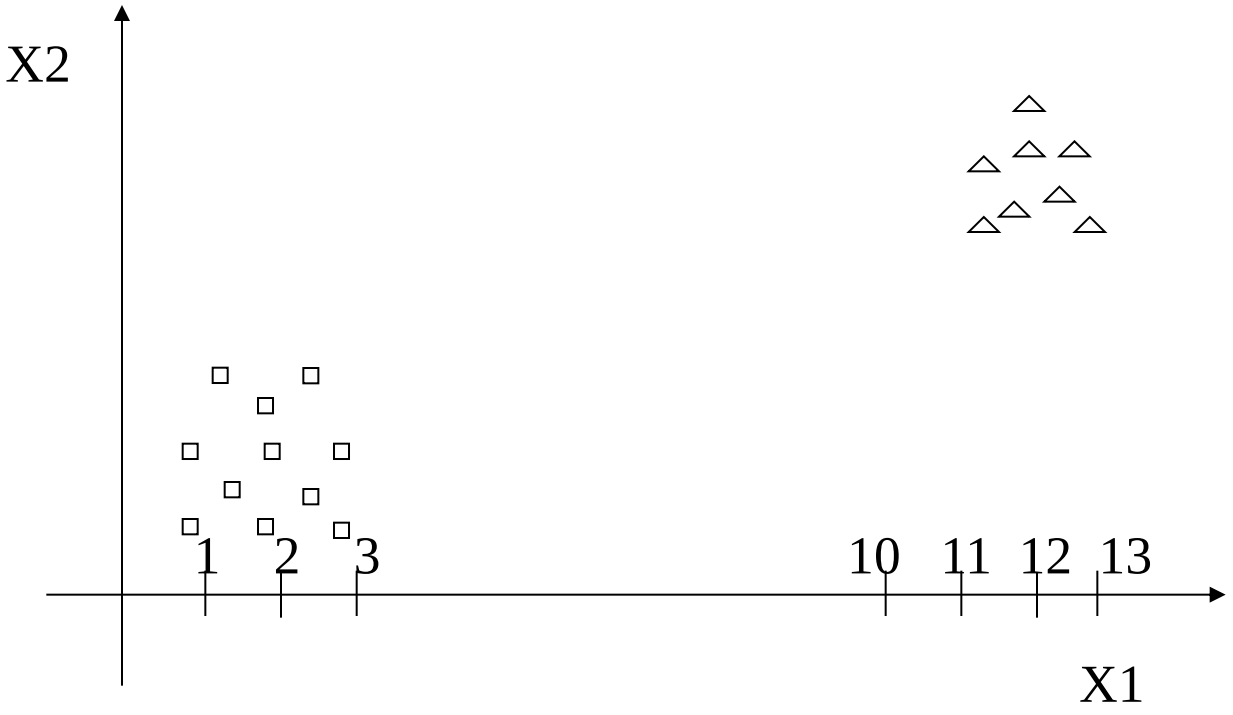


Lo scopo è:

- 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 $D_i \bigcap D_j = \Phi$.

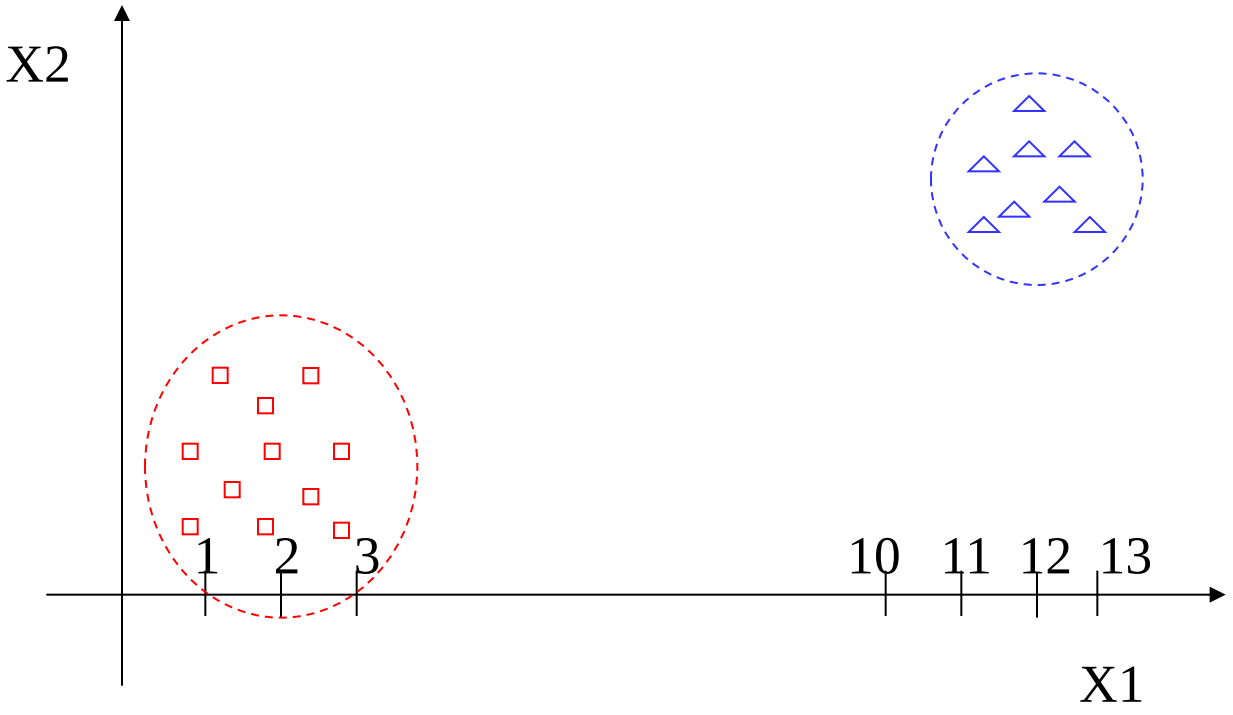
Clustering

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



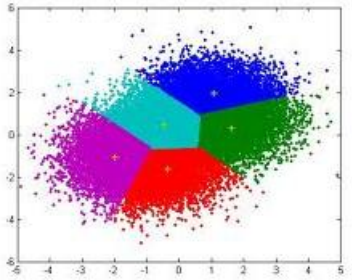
Clustering

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?
   - *Quality-Threshold (QR).*
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 per scoprire il cluster di appartenenza.*



QT

(Heyer, Kruglyak, & Yooseph 1999)

QT(D,radius)-:clusterSet

clusterSet: insieme di cluster D_i : ogni cluster D_i è un insieme di transazioni in D

Begin

1. Costruzione del cluster candidato **C** per ogni transazione **t** (centroide) in **D; C** raggruppa le transazioni **t'** di **D** che distano al più radius dal centroide candidato **t** (**distanza(t,t')<=radius**).
2. Scelta del cluster candidato più popoloso come cluster definitivo da aggiungere a **clusterSet** e rimozione delle transazioni in esso incluse da **D**
3. Se ci sono ancora transazioni in **D**, torna al passo 1

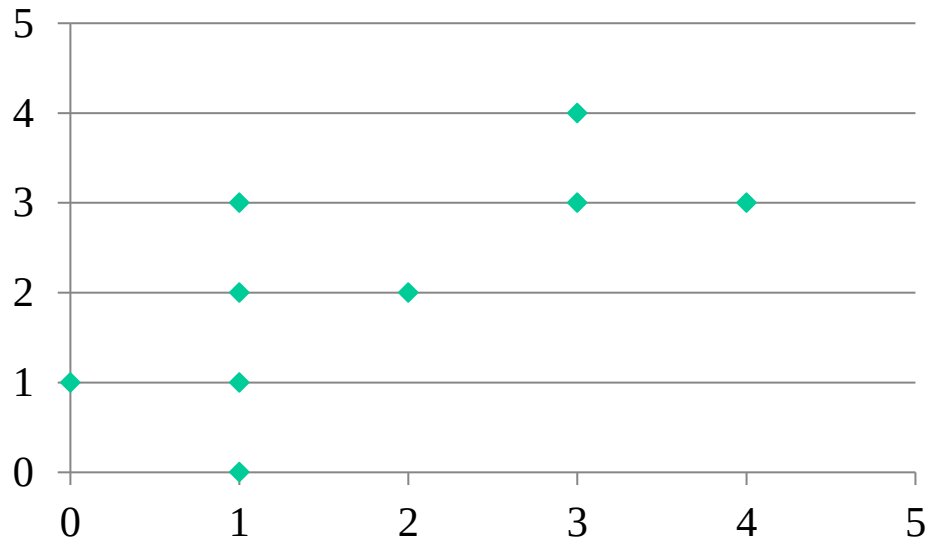
End

radius=1

QT: come?

PASSO 1: determinazione dei cluster candidati associati a ciascuna transazione

	X	Y
t1	1	2
t2	0	1
t3	1	0
t4	1	3
t5	2	2
t6	1	1
t7	3	3
t8	3	4
t9	4	3



radius=1

QT: come?

	X	Y
t1	1	2
t2	0	1
t3	1	0
t4	1	3
t5	2	2
t6	1	1
t7	3	3
t8	3	4
t9	4	3

PASSO 1: determinazione dei cluster candidati associati a ciascuna transazione

C(t1)={t1,t4,t5,t6}

EuclideanDist(t1,t1)=0

EuclideanDist(t1,t2)=1.414

EuclideanDist(t1,t3)=2

EuclideanDist(t1,t4)=1

EuclideanDist(t1,t5)=1

EuclideanDist(t1,t6)=1

EuclideanDist(t1,t7)=2.23

EuclideanDist(t1,t8)=2.82

EuclideanDist(t1,t9)=3.16

radius=1

QT: come?

	X	Y
t1	1	2
t2	0	1
t3	1	0
t4	1	3
t5	2	2
t6	1	1
t7	3	3
t8	3	4
t9	4	3

PASSO 1: determinazione dei cluster candidati associati a ciascuna transazione

C(t1)={t1,t4,t5,t6}

C(t2)={t2,t6}

C(t3)={t3,t6}

C(t4)={t1,t4}

C(t5)={t1,t5}

C(t6)={t1,t2,t3,t6}

C(t7)={t7,t8,t9}

C(t8)={t8}

C(t9)={t7,t9}

radius=1

QT: come?

	X	Y
t1	**1**	**2**
t2	0	1
t3	1	0
t4	**1**	**3**
t5	**2**	**2**
t6	**1**	**1**
t7	3	3
t8	3	4
t9	4	3

PASSO 2: scelta del cluster più popoloso e rimozione delle sue transazioni da D

C(t1)={t1,t4,t5,t6}

C(t2)={t2,t6}

C(t3)={t3,t6}

C(t4)={t1,t4}

C(t5)={t1,t5}

C(t6)={t1,t2,t3,t6,}

C(t7)={t7,t8,t9}

C(t8)={t8}

C(t9)={t7,t9}

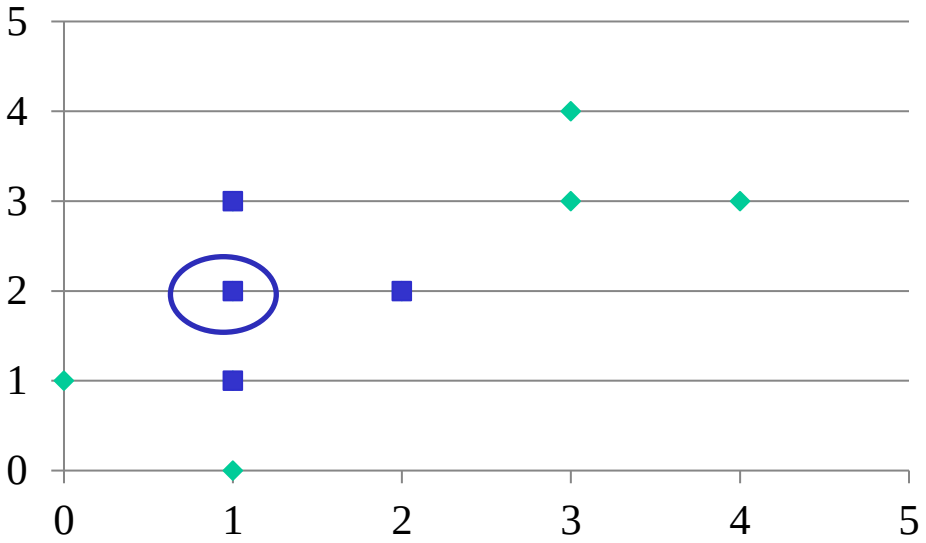
radius=1

QT: come?

PASSO 2: scelta del cluster più popoloso e rimozione delle sue transazioni da D

C(t1)={t1,t4,t5,t6}

	X	Y
t1	**1**	**2**
t2	0	1
t3	1	0
t4	**1**	**3**
t5	**2**	**2**
t6	**1**	**1**
t7	3	3
t8	3	4
t9	4	3



	X	Y
t2	0	1
t3	1	0
t7	3	3
t8	3	4
t9	4	3

radius=1

QT: come?

	X	Y
t2	0	1
t3	1	0
t7	3	3
t8	3	4
t9	4	3

PASSO 1: determinazione dei cluster candidati associati a ciascuna transazione

C(t2)={t2}

C(t3)={t3}

C(t7)={t7,t8,t9}

C(t8)={t7,t8}

C(t9)={t7,t9}

radius=1

QT: come?

PASSO 2: scelta del cluster più popoloso e rimozione delle sue transazioni da D

	X	Y
t2	0	1
t3	1	0
t7	**3**	**3**
t8	**3**	**4**
t9	**4**	**3**

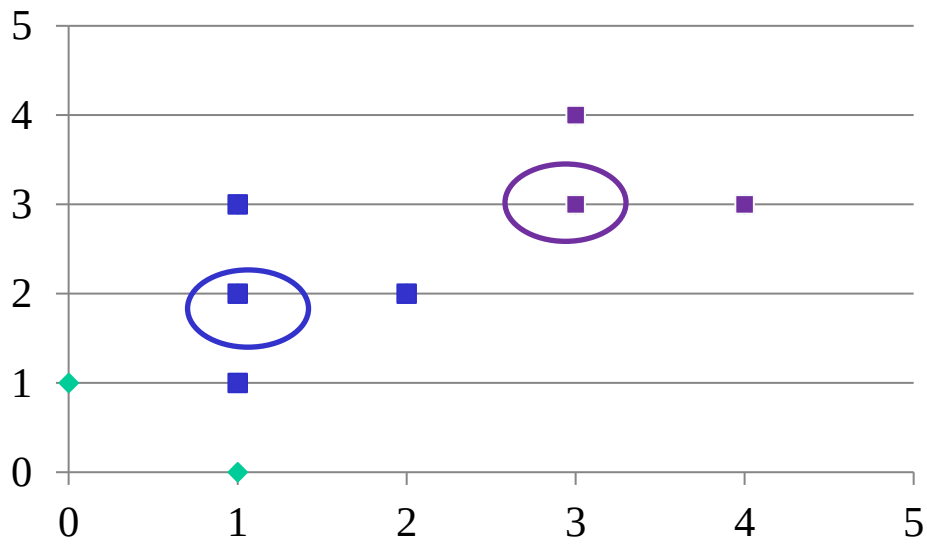
C(t2)={t2}

C(t3)={t3}

C(t7)={t7,t8,t9}

C(t8)={t7,t8}

C(t9)={t7,t9}



QT: come?

	X	Y
t2	0	1
t3	1	0

PASSI 1-2:

C(t2)={t2}

PASSI 1-2:

C(t3)={t3}

Cluster scoperti

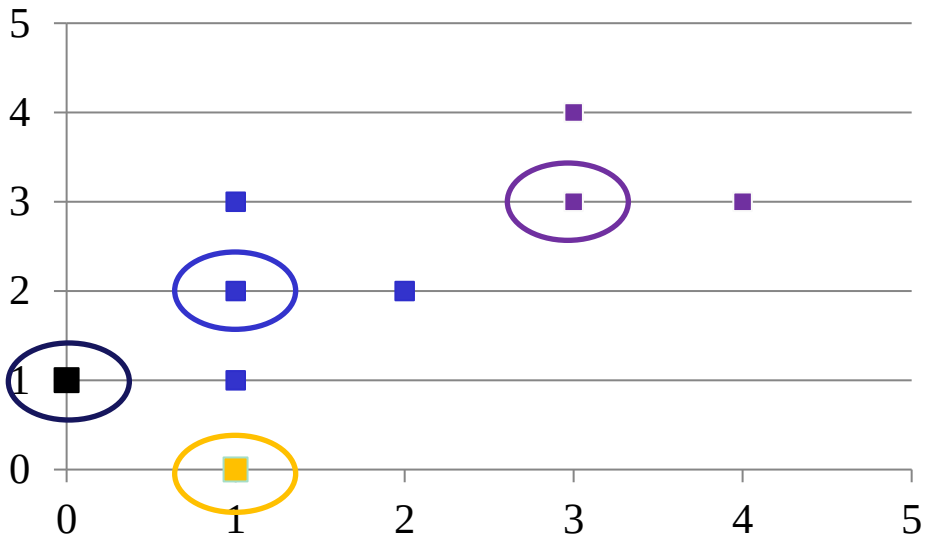
C(t1)={t1,t4,t5,t6}

C(t7)={t7,t8,t9}

C(t2)={t2}

C(t3)={t3}

QT: cluster e centroidi



Rappresentazione di un cluster

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

Cluster 1

	X	Y
t1	1	2
t2	0	1
t4	1	3
t5	2	2
t6	1	1

Cluster 2

	X	Y
t7	3	3
t8	3	4
t9	4	3

Cluster 3

	X	Y
t2	0	1

Cluster 4

	X	Y
t3	1	0

Rappresentazione di un cluster

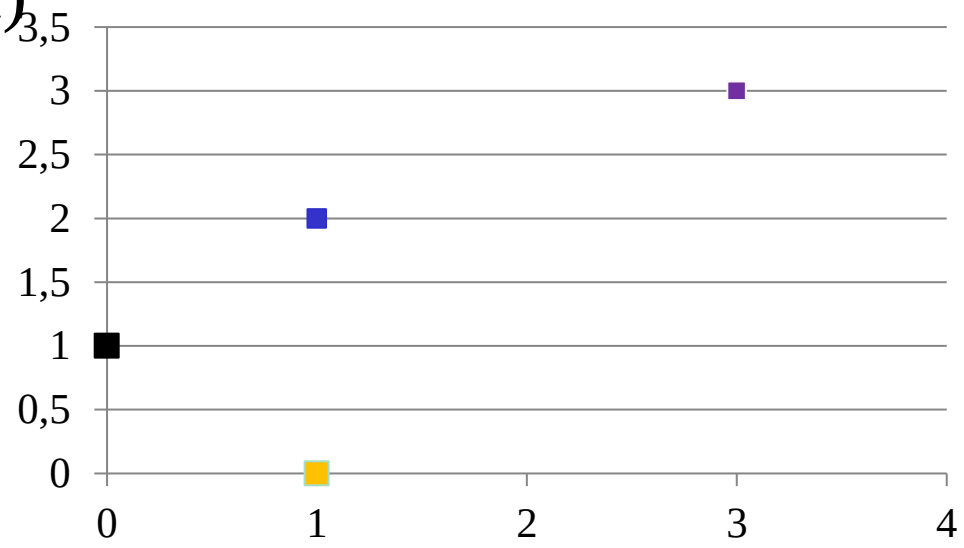
2) Descrizione intensionale (tramite i centroidi del cluster + Rggio).

Cluster 1	Cluster 2
(1, 2) Raggio=1	(3,3) Raggio =1
Cluster 3	Cluster 4
(0, 1) Raggio=1	(1,0) Raggio =1

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)

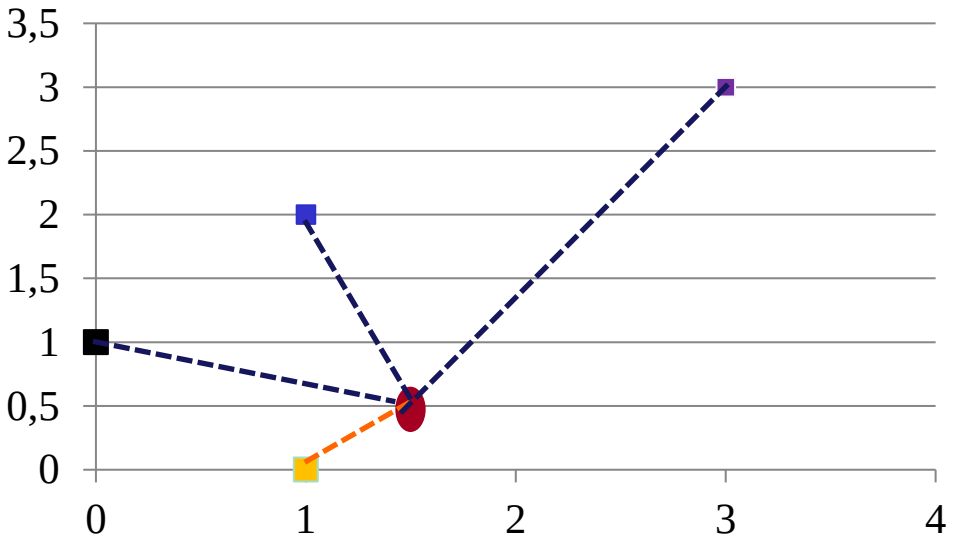


Cluster e/o centroidi: applicazioni reali

Vantaggi:

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

(1.5,0.5)??



Caso di studio

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

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

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

Istruzioni

1. Il progetto dello A.A. 2024-2025, denominato QT, è valido solo per coloro **che superano la prova scritta 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).
4. La valutazione del progetto avverrà alla sua consegna e sarà la stessa per ciascun componente del gruppo. Il voto massimo del progetto è 33. Il progetto deve essere valutato almeno 18, diversamente sarà necessario consegnare un nuovo progetto (con una nuova traccia)
5. Il voto finale sarà stabilito sulla base del voto attribuito allo scritto e al progetto (media aritmetica).

Istruzioni

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

Consultare il sillabo per la descrizione dei criteri applicati per la valutazione del progetto