

---

# Clustering

Caso di studio di Metodi Avanzati di  
Programmazione  
AA 2017-2018

---

# 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:*

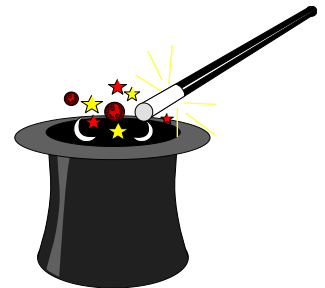
- una collezione  $D$  di transazioni dove, ogni transazione è un vettore di coppie attributo-valore (item);
- un intero  $k$ ;

*Lo scopo è:*

- partizionare  $D$  in  $k$  insiemi di transazioni  $D_1, \dots, D_k$ , tale che:

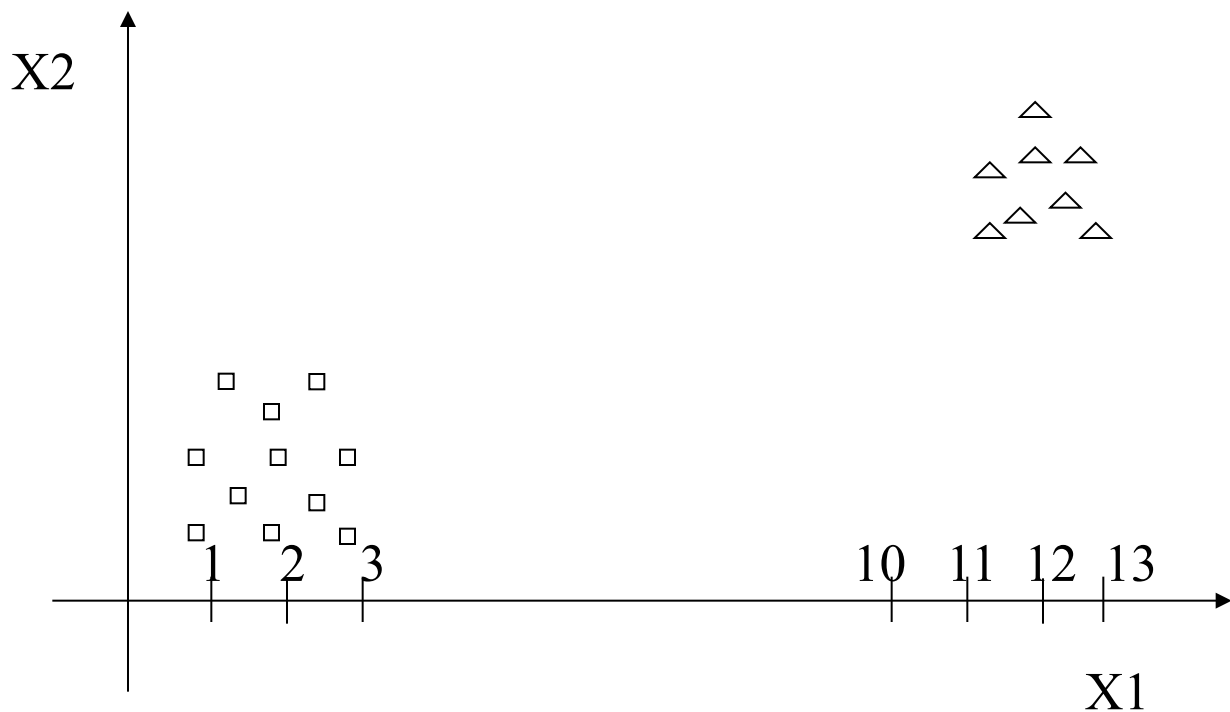
- $D_i$  ( $i=1, \dots, k$ ) è un segmento (selezione) omogenea di  $D$ ;

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



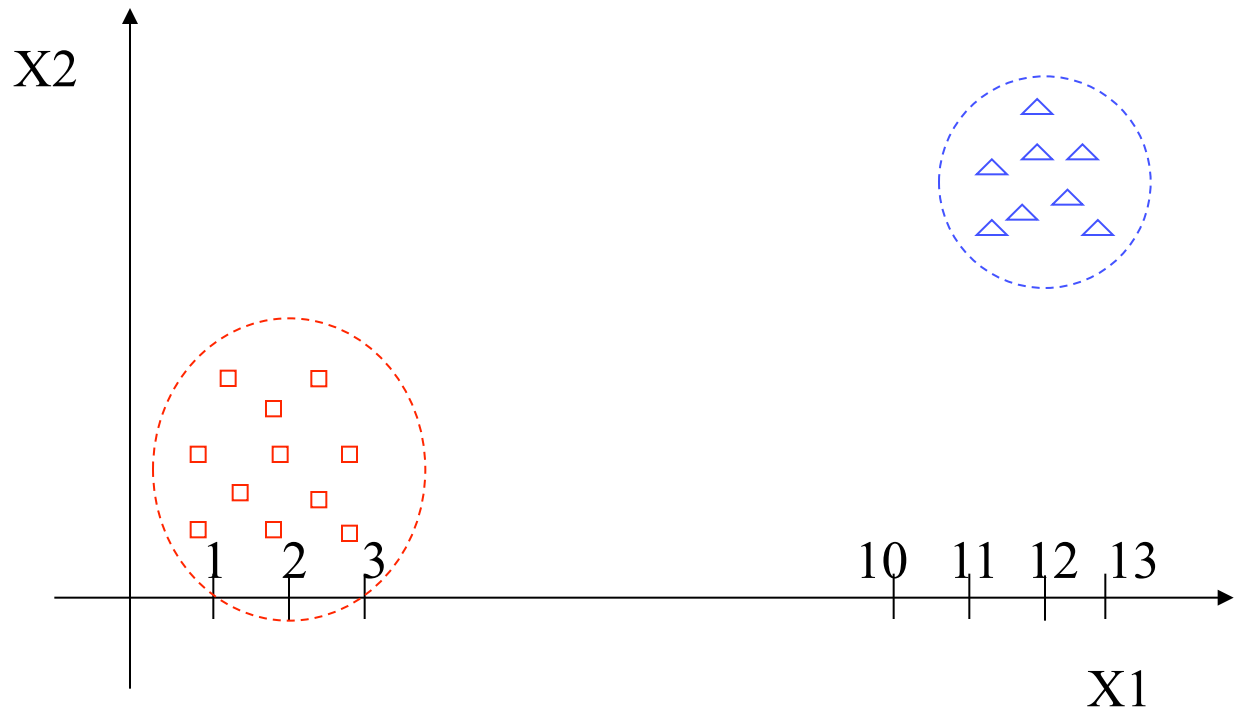
X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

# 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

# Clustering

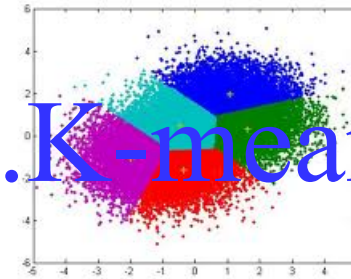


---

# 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 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 *clusterSet* una transazione casualmente scelta da  $D$

### **3. do**

**for** (*transazione*: $D$ )

3.1  $D_i = \text{cluster}(\text{clusterSet}, \text{transazione})$

3.2 sposta *transazione* nel segmento  $D_i$

**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 =  $\{D_1, D_2\}$

$D_1 = \{\}$

$D_2 = \{\}$

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<sub>1</sub>, D<sub>2</sub>}

c1 = (0.9, 1.2) : D1 = D1 ∪ c1

c2 = (2, 2.2) : D2 = D1 ∪ c2

X1	X2
0.9	1
<b>0.9</b>	<b>1.2</b>
1.3	2
1.2	3.7
1.9	1
<b>2</b>	<b>2.2</b>
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
<b>0.9</b>	<b>1.2</b>
1.3	2
1.2	3.7
1.9	1
<b>2</b>	<b>2.2</b>
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.*

"IDEE IN CORSO"



$$\text{EuclideanDist}((0.9, 1), (0.9, 1.2)) = 0.2$$

$$\text{EuclideanDist}((0.9, 1), (2, 2.2)) = 1.62$$

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

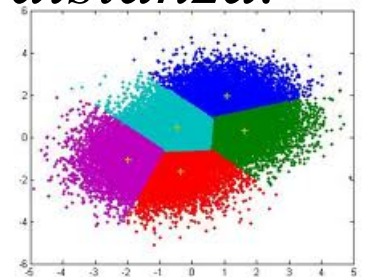
*L'appartenenza di una transazione a un cluster dipende dalla distanza della transazione dal centroide del cluster.*

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

clusterSet = { $D_1, D_2$ }

$D_1 = \{1, 2, 5, 8\}$

$D_2 = \{3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17\}$



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: 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<sub>1</sub>, D<sub>2</sub>}

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<sub>1</sub>, D<sub>2</sub>}

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

$c1 = (1.65, 1.05)$

$c2 = (8.03, 4.66)$

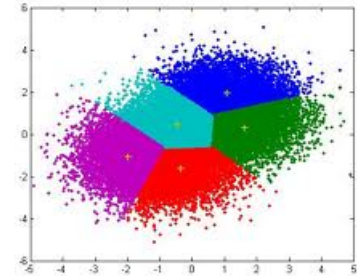


X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

## 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:

$c1 = (1.76, 1.98)$

$c2 = (11.9, 5.875)$

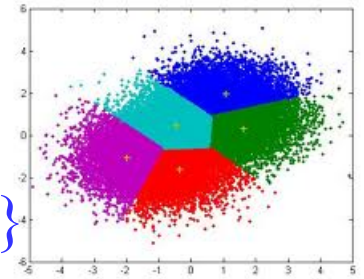
X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

## 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

$$c1 = (1.76, 1.98) \quad c2 = (11.9, 5.875)$$

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).

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

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



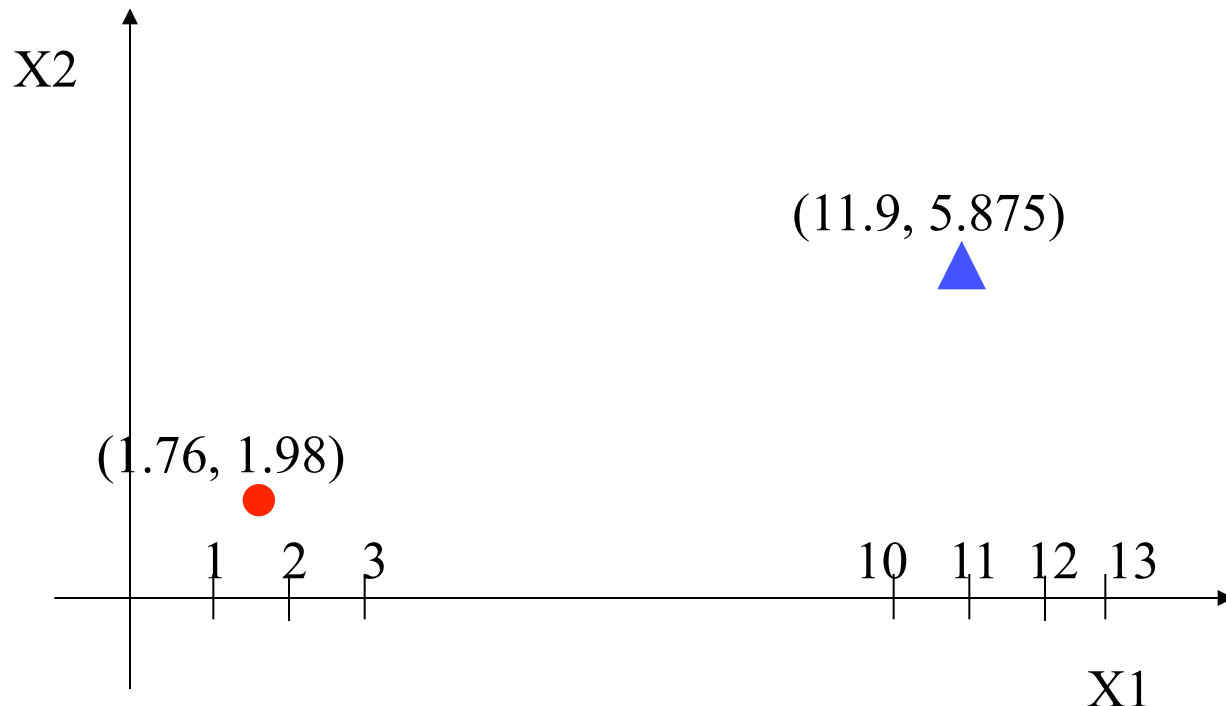
centroide

<b>F</b>	<b>Italiana</b>	<b>27.6</b>
----------	-----------------	-------------

### 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)

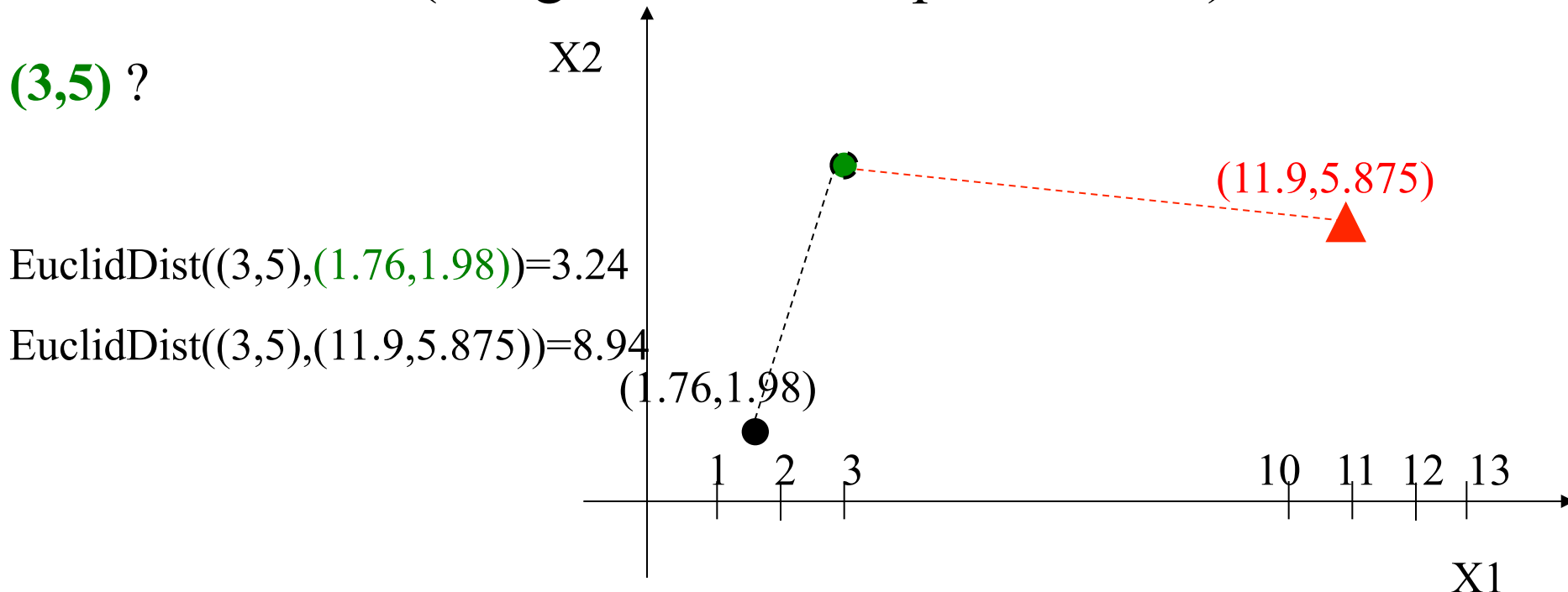




### 3.Cluster e/o centroidi: applicazioni reali

Vantaggi:

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



---

# 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 applet Java che consente di usufruire del servizio di scoperta remoto e visualizza la conoscenza (cluster) scoperta

---

# Tutor di Laboratorio

- Dott.ssa Lucrezia Macchia

---

# Istruzioni

1. Il progetto dello A.A. 2011-2012, 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).
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 sue parti (client-server, applet, accesso al db, serializzazione,...)