# Lecture 10

## Clustering

**Mercoledì, 21 Febbraio 2007**

**Giuseppe Manco**

*Readings:*

Chapter 8, Han and Kamber

Chapter 14, Hastie , Tibshirani and Friedman



ICAR CNR
Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

# Outline

- Introduction
- K-means clustering
- Hierarchical clustering: COBWEB

# Apprendimento supervisionato

- **Dati**
  - Un insieme X di istanze su dominio multidimensionale
  - Un funzione target *c*
  - Il linguaggio delle ipotesi *H*
  - Un insieme di allenamento *D* = {<x,c(x)>}

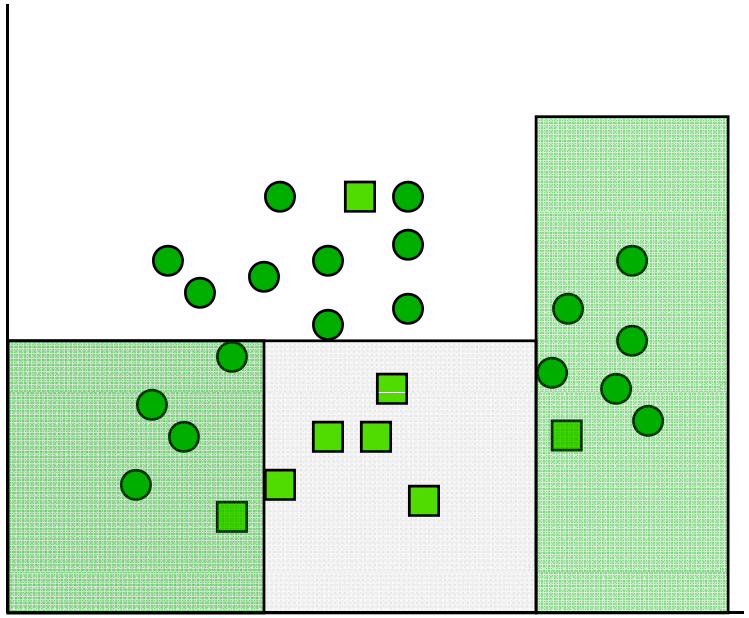
- **Determinare**
  - L'ipotesi $h \in H$ tale che $h(x) = c(x)$ per ogni $x \in D$
  - Che sia consistente con il training set

# Supervised vs. Unsupervised Learning

- Supervised learning (classificazione)
  - Supervisione: il training set contiene l'etichetta che indica la classe da apprendere
  - I nuovi dati sono classificati sulla base di quello che si apprende dal training set
- Unsupervised learning (clustering)
  - L'etichetta di classe è sconosciuta
  - Le istanze sono fornite con l'obiettivo di stabilire se vi sono raggruppamenti (classi) tra i dati

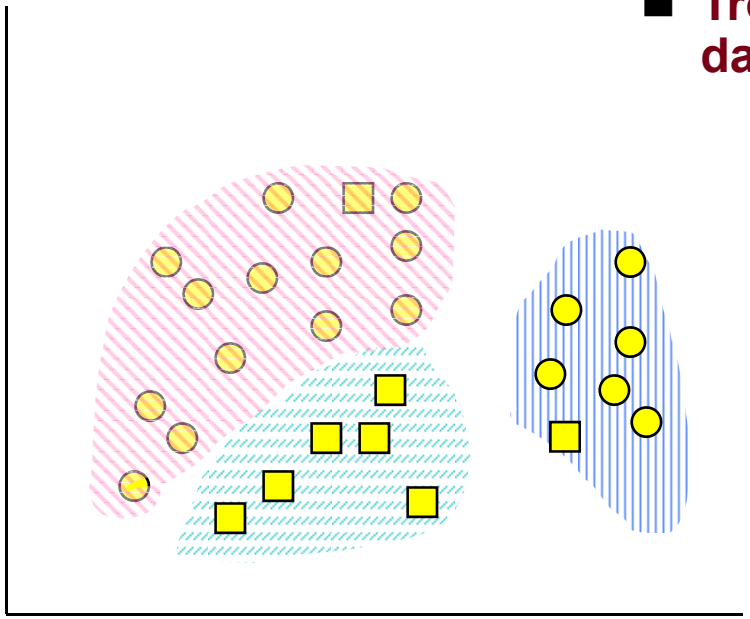
# Classificazione, Clustering

- **Classificazione: Apprende un metodo per predire la classe dell'istanza da altre istanze pre-classificate**



# Clustering

- Trova raggruppamenti “naturali” nei dati non etichettati



- **Applicazioni tipiche**
  - **Tool stand-alone to get insight into data distribution**
  - Passo di preprocessing **per altri algoritmi**

# Clustering, clusters

- Raggruppamento di dati in classi (clusters) che abbiano una significatività
  - alta similarità intra-classe
  - Bassa similarità inter-classe
- Qualità di uno schema di clustering
  - La capacità di ricostruire la struttura nascosta
    - similarità

# Similarità, distanza

- **Distanza d(x,y)**
  - **Misura la "dissimilarità tra gli oggetti"**
- **Similarità s(x,y)**
  - **S(x,y) ≈1/d(x,y)**
  - **Esempio**

$$s(x, y) = e^{-d(x,y)}$$

- **Proprietà desiderabili**

$$d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

$$d(\mathbf{x}_i, \mathbf{x}_i) = 0$$

$$d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$$

$$d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_k, \mathbf{x}_j)$$

- **Definizione application-dependent**
  - **Può richiedere normalizzazione**
  - **Diverse definizioni per differenti tipi di attributi**

# Esempio

| Istanza | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $I_1$ | 0 | 0 | 0 |
| $I_2$ | 1 | 0 | 0 |
| $I_3$ | 2 | 0 | 0 |
| $I_4$ | 2.5 | 2 | 0 |
| $I_5$ | 3 | 0 | 0 |
| $I_6$ | 1 | 2 | 1 |
| $I_7$ | 1.5 | 0 | 1 |
| $I_8$ | 2 | 2 | 1 |
| $I_9$ | 3 | 2 | 1 |
| $I_{10}$ | 4 | 2 | 1 |

$$\begin{bmatrix} d(I_1,I_1) & d(I_1,I_2) & \dots & d(I_1,I_{10}) \\ d(I_2,I_1) & d(I_2,I_2) & & \vdots \\ \vdots & & & \vdots \\ \vdots & & \cdot & \vdots \\ \vdots & & & \\ d(I_{10},I_1) & d(I_{10},I_2) & & d(I_{10},I_{10}) \end{bmatrix}$$

# Similarità e dissimilarità tra oggetti

- La distanza è utilizzata per misurare la similarità (o dissimilarità) tra due istanze
- Distanza di Minkowski (Norma $L_p$):

$$dist(x, y) = \left( \sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

  - Dove $\underline{x} = (x_1, x_2, \ldots, x_d)$ e $\underline{y} = (y_1, y_2, \ldots, y_d)$ sono due oggetti d-dimensionali, e p è un numero primo
- se p = 1, d è la distanza Manhattan

$$dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

- Se p=∞

$$dist(x, y) = \sup_{1 \le i \le d} |x_i - y_i|$$

# Similarità e dissimilarità tra oggetti [2]

- **se p = 2, d è la distanza euclidea:**

$$dist(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

  - **Proprietà**
    - **Translation-invariant**
    - **Scale variant**
- **Varianti**
  - **Distanza pesata**

$$dist(x, y) = \left( \sum_{i=1}^{d} w_i |x_i - y_i|^p \right)^{1/p}$$

  - **Distanza Mahalanobis**

$$dist(x, y) = (x - y)^T \Sigma^{-1} (x - y)$$

# Esempio

Euclidea

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3,2016 | 3 | 2,4495 | 1,8028 | 3 | 3,7417 | 4,5826 |
| 1 | 0 | 1 | 2,5 | 2 | 2,2361 | 1,118 | 2,4495 | 3 | 3,7417 |
| 2 | 1 | 0 | 2,0616 | 1 | 2,4495 | 1,118 | 2,2361 | 2,4495 | 3 |
| 3,2016 | 2,5 | 2,0616 | 0 | 2,0616 | 1,8028 | 2,4495 | 1,118 | 1,118 | 1,8028 |
| 3 | 2 | 1 | 2,0616 | 0 | 3 | 1,8028 | 2,4495 | 2,2361 | 2,4495 |
| 2,4495 | 2,2361 | 2,4495 | 1,8028 | 3 | 0 | 2,0616 | 1 | 2 | 3 |
| 1,8028 | 1,118 | 1,118 | 2,4495 | 1,8028 | 2,0616 | 0 | 2,0616 | 2,5 | 3,2016 |
| 3 | 2,4495 | 2,2361 | 1,118 | 2,4495 | 1 | 2,0616 | 0 | 1 | 2 |
| 3,7417 | 3 | 2,4495 | 1,118 | 2,2361 | 2 | 2,5 | 1 | 0 | 1 |
| 4,5826 | 3,7417 | 3 | 1,8028 | 2,4495 | 3 | 3,2016 | 2 | 1 | 0 |

mahalanobis

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,9 | 3,6 | 7,65 | 8,1 | 4,5 | 7,65 | 5,4 | 8,1 | 12,6 |
| 0,9 | 0 | 0,9 | 5,85 | 3,6 | 5,4 | 5,85 | 4,5 | 5,4 | 8,1 |
| 3,6 | 0,9 | 0 | 5,85 | 0,9 | 8,1 | 5,85 | 5,4 | 4,5 | 5,4 |
| 7,65 | 5,85 | 5,85 | 0 | 7,65 | 7,65 | 18 | 5,85 | 5,85 | 7,65 |
| 8,1 | 3,6 | 0,9 | 7,65 | 0 | 12,6 | 7,65 | 8,1 | 5,4 | 4,5 |
| 4,5 | 5,4 | 8,1 | 7,65 | 12,6 | 0 | 7,65 | 0,9 | 3,6 | 8,1 |
| 7,65 | 5,85 | 5,85 | 18 | 7,65 | 7,65 | 0 | 5,85 | 5,85 | 7,65 |
| 5,4 | 4,5 | 5,4 | 5,85 | 8,1 | 0,9 | 5,85 | 0 | 0,9 | 3,6 |
| 8,1 | 5,4 | 4,5 | 5,85 | 5,4 | 3,6 | 5,85 | 0,9 | 0 | 0,9 |
| 12,6 | 8,1 | 5,4 | 7,65 | 4,5 | 8,1 | 7,65 | 3,6 | 0,9 | 0 |

| Istanza | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $I_1$ | 0 | 0 | 0 |
| $I_2$ | 1 | 0 | 0 |
| $I_3$ | 2 | 0 | 0 |
| $I_4$ | 2.5 | 2 | 0 |
| $I_5$ | 3 | 0 | 0 |
| $I_6$ | 1 | 2 | 1 |
| $I_7$ | 1.5 | 0 | 1 |
| $I_8$ | 2 | 2 | 1 |
| $I_9$ | 3 | 2 | 1 |
| $I_{10}$ | 4 | 2 | 1 |

manhattan

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 4,5 | 3 | 4 | 2,5 | 5 | 6 | 7 |
| 1 | 0 | 1 | 3,5 | 2 | 3 | 1,5 | 4 | 5 | 6 |
| 2 | 1 | 0 | 2,5 | 1 | 4 | 1,5 | 3 | 4 | 5 |
| 4,5 | 3,5 | 2,5 | 0 | 2,5 | 2,5 | 4 | 1,5 | 1,5 | 2,5 |
| 3 | 2 | 1 | 2,5 | 0 | 5 | 2,5 | 4 | 3 | 4 |
| 4 | 3 | 4 | 2,5 | 5 | 0 | 2,5 | 1 | 2 | 3 |
| 2,5 | 1,5 | 1,5 | 4 | 2,5 | 2,5 | 0 | 2,5 | 3,5 | 4,5 |
| 5 | 4 | 3 | 1,5 | 4 | 1 | 2,5 | 0 | 1 | 2 |
| 6 | 5 | 4 | 1,5 | 3 | 2 | 3,5 | 1 | 0 | 1 |
| 7 | 6 | 5 | 2,5 | 4 | 3 | 4,5 | 2 | 1 | 0 |

# Similarità e dissimilarità tra oggetti [3]

- **Similarità del coseno**

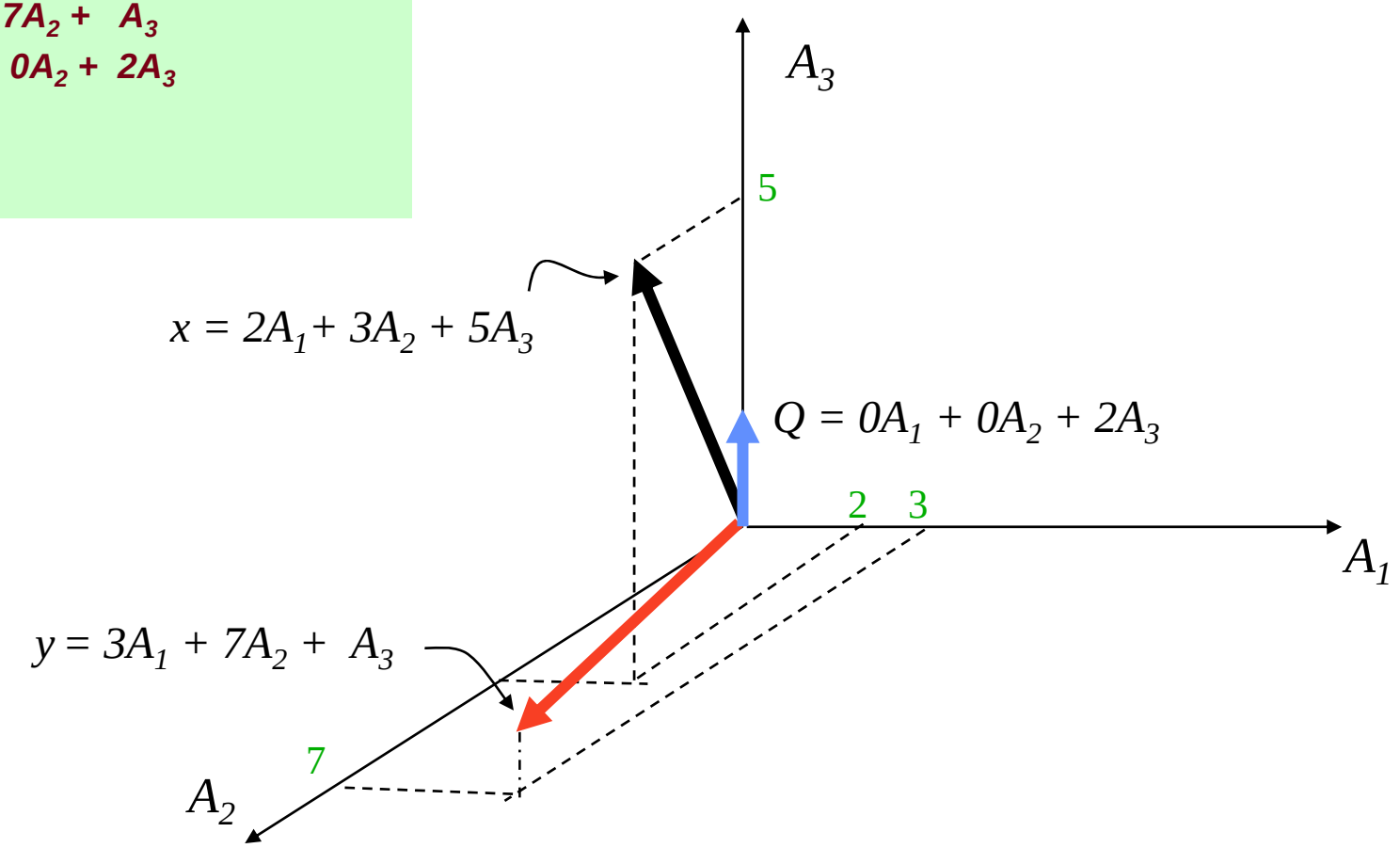
$$sim(x, y) = \frac{x^T y}{\|x\| \|y\|}$$

  - **Proprietà**
    - **Translation variant**
    - **Scale invariant**
- **Similarità di Jaccard (Tanimoto)**

$$sim(x, y) = \frac{x^T y}{\|x\|^2 + \|y\|^2 - x^T y}$$

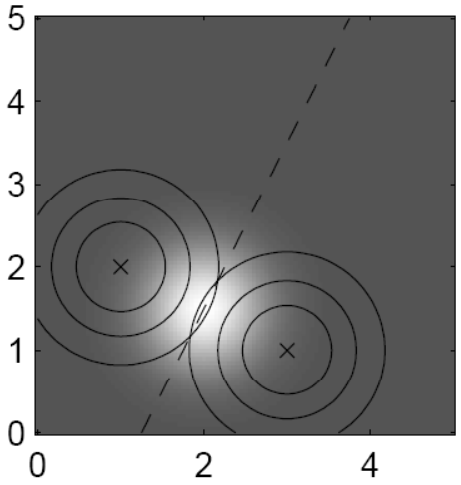
# Rappresentazione grafica

$x = 2A_1 + 3A_2 + 5A_3$
$y = 3A_1 + 7A_2 + A_3$
$Q = 0A_1 + 0A_2 + 2A_3$

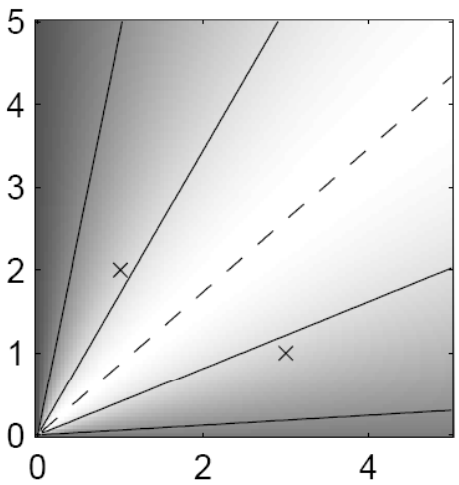


# Rappresentazione grafica

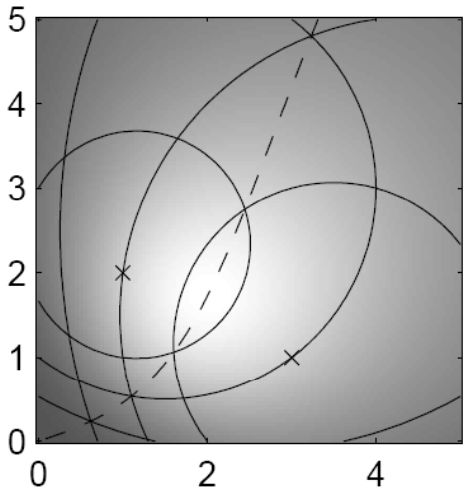
euclidea




Coseno




Jaccard



# Esempio

Euclidea

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3,2016 | 3 | 2,4495 | 1,8028 | 3 | 3,7417 | 4,5826 |
| 1 | 0 | 1 | 2,5 | 2 | 2,2361 | 1,118 | 2,4495 | 3 | 3,7417 |
| 2 | 1 | 0 | 2,0616 | 1 | 2,4495 | 1,118 | 2,2361 | 2,4495 | 3 |
| 3,2016 | 2,5 | 2,0616 | 0 | 2,0616 | 1,8028 | 2,4495 | 1,118 | 1,118 | 1,8028 |
| 3 | 2 | 1 | 2,0616 | 0 | 3 | 1,8028 | 2,4495 | 2,2361 | 2,4495 |
| 2,4495 | 2,2361 | 2,4495 | 1,8028 | 3 | 0 | 2,0616 | 1 | 2 | 3 |
| 1,8028 | 1,118 | 1,118 | 2,4495 | 1,8028 | 2,0616 | 0 | 2,0616 | 2,5 | 3,2016 |
| 3 | 2,4495 | 2,2361 | 1,118 | 2,4495 | 1 | 2,0616 | 0 | 1 | 2 |
| 3,7417 | 3 | 2,4495 | 1,118 | 2,2361 | 2 | 2,5 | 1 | 0 | 1 |
| 4,5826 | 3,7417 | 3 | 1,8028 | 2,4495 | 3 | 3,2016 | 2 | 1 | 0 |

Coseno

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0,21 | 0 | 0,59 | 0,16 | 0,33 | 0,19 | 0,12 |
| 0 | 0 | 0 | 0,21 | 0 | 0,59 | 0,16 | 0,33 | 0,19 | 0,12 |
| 0 | 0,21 | 0,21 | 0 | 0,21 | 0,17 | 0,35 | 0,06 | 0,03 | 0,045 |
| 0 | 0 | 0 | 0,21 | 0 | 0,59 | 0,16 | 0,33 | 0,19 | 0,12 |
| 0 | 0,59 | 0,59 | 0,17 | 0,59 | 0 | 0,43 | 0,04 | 0,12 | 0,19 |
| 0 | 0,16 | 0,16 | 0,35 | 0,16 | 0,43 | 0 | 0,26 | 0,18 | 0,15 |
| 0 | 0,33 | 0,33 | 0,06 | 0,33 | 0,04 | 0,26 | 0 | 0,022 | 0,054 |
| 0 | 0,19 | 0,19 | 0,039 | 0,19 | 0,12 | 0,18 | 0,02 | 0 | 0,008 |
| 0 | 0,12 | 0,12 | 0,04 | 0,12 | 0,19 | 0,15 | 0,054 | 0,008 | 0 |

| Istanza | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $I_1$ | 0 | 0 | 0 |
| $I_2$ | 1 | 0 | 0 |
| $I_3$ | 2 | 0 | 0 |
| $I_4$ | 2.5 | 2 | 0 |
| $I_5$ | 3 | 0 | 0 |
| $I_6$ | 1 | 2 | 1 |
| $I_7$ | 1.5 | 0 | 1 |
| $I_8$ | 2 | 2 | 1 |
| $I_9$ | 3 | 2 | 1 |
| $I_{10}$ | 4 | 2 | 1 |

Jaccard

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0,66667 | 0,28571 | 0,42857 | 0,16667 | 0,54545 | 0,25 | 0,25 | 0,22222 |
| 0 | 0,66667 | 0 | 0,54054 | 0,85714 | 0,25 | 0,70588 | 0,44444 | 0,5 | 0,47059 |
| 0 | 0,28571 | 0,54054 | 0 | 0,6383 | 0,66667 | 0,38462 | 0,87805 | 0,90196 | 0,81159 |
| 0 | 0,42857 | 0,85714 | 0,6383 | 0 | 0,25 | 0,58065 | 0,5 | 0,64286 | 0,66667 |
| 0 | 0,16667 | 0,25 | 0,66667 | 0,25 | 0 | 0,37037 | 0,875 | 0,66667 | 0,5 |
| 0 | 0,54545 | 0,70588 | 0,38462 | 0,58065 | 0,37037 | 0 | 0,48485 | 0,46809 | 0,4058 |
| 0 | 0,25 | 0,44444 | 0,87805 | 0,5 | 0,875 | 0,48485 | 0 | 0,91667 | 0,76471 |
| 0 | 0,25 | 0,5 | 0,90196 | 0,64286 | 0,66667 | 0,46809 | 0,91667 | 0 | 0,94444 |
| 0 | 0,22222 | 0,47059 | 0,81159 | 0,66667 | 0,5 | 0,4058 | 0,76471 | 0,94444 | 0 |

# Attributi binari

- **Distanza di Hamming**

$$dist(x,y) = \sum_{i=1}^{d} |x_i - y_i| = \sum_{i=1}^{d} \delta(x_i, y_i)$$

$$\delta(x_i, y_i) = \begin{cases} 1 & \text{se } x_i = y_i \\ 0 & \text{altrimenti} \end{cases}$$

  - **Distanza Manhattan quando i valori possibili sono 0 o 1**
- **In pratica, conta il numero di mismatches**

# Attributi binari

- **Utilizzando la tabella di contingenza**

| | | Oggetto y | | |
|---|---|---|---|---|
| | | 1 | 0 | *totale* |
| **Oggetto *x*** | 1 | $a$ | $b$ | $a+b$ |
| | 0 | $c$ | $d$ | $c+d$ |
| | *totale* | $a+c$ | $b+d$ | $p$ |

- **Coefficiente di matching (invariante, se le variabili sono *simmetriche*):**

$$d(x,y) = \frac{b+c}{a+b+c+d}$$

- **Coefficiente di Jaccard (noninvariante se le variabili sono *asimmetriche*):**

$$d(x,y) = \frac{b+c}{a+b+c}$$

# Dissimilarità tra attributi binari

- **Esempio**

| Name | Gender | Fever | Cough | Test-1 | Test-2 | Test-3 | Test-4 |
|---|---|---|---|---|---|---|---|
| Jack | M | Y | N | P | N | N | N |
| Mary | F | Y | N | P | N | P | N |
| Jim | M | Y | P | N | N | N | N |

  - gender è simmetrico
  - Tutti gli altri sono asimmetrici
  - Poniamo Y e P uguale a 1, e N a 0

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

# Variabili Nominali

- Generalizzazione del meccanismo di variabili binarie
- Metodo 1: Matching semplice
  - *m*: # di matches, *p*: # di attributi nominali

$$d(x, y) = \frac{p - m}{p}$$

- metodo 2: binarizzazione
- Metodo 3: Jaccard su insiemi

# Combinare dissimilarità

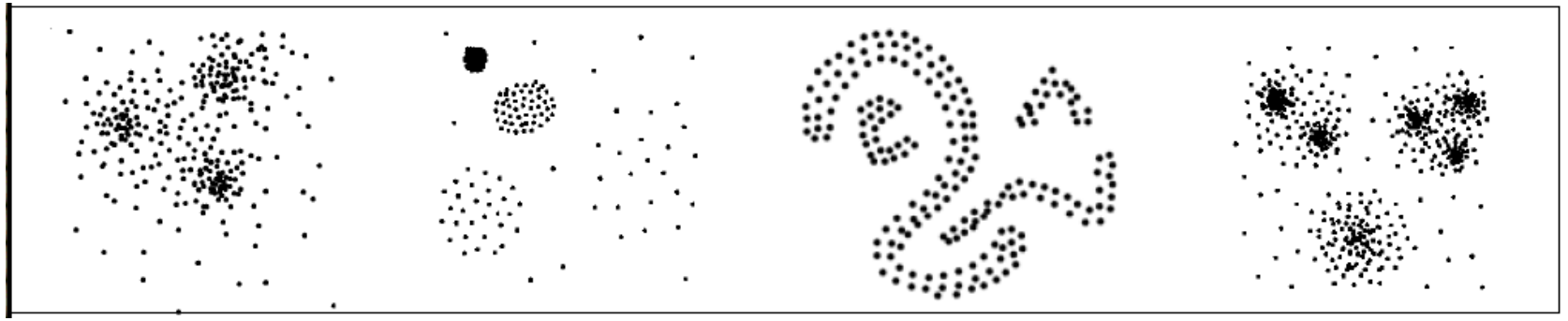
- **x,y**
  - **$x_R$, $y_R$: attributi numerici**
  - **$x_n$, $y_n$: attributi nominali**

$$dist(x, y) = \alpha dist(x_R, y_R) + \beta dist(x_n, y_n)$$

  - **Va garantita la stessa scala**

# Metodi di clustering

- **Una miriade di metodologie differenti:**
  - **Dati numerici/simbolici**
  - **Deterministici/probabilistici**
  - **Partizionali/con overlap**
  - **Gerarchici/piatti**
  - **Top-down/bottom-up**



# Clustering per enumerazione

- **Utilizzando come criterio**

$$w(C) = \sum_{x \in C} \sum_{y \in C} dist(x, y)$$

- **Il numero di possibili clusterings sarebbe**

$$S(n,k) = \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-i} \binom{k}{i} \times i^n$$

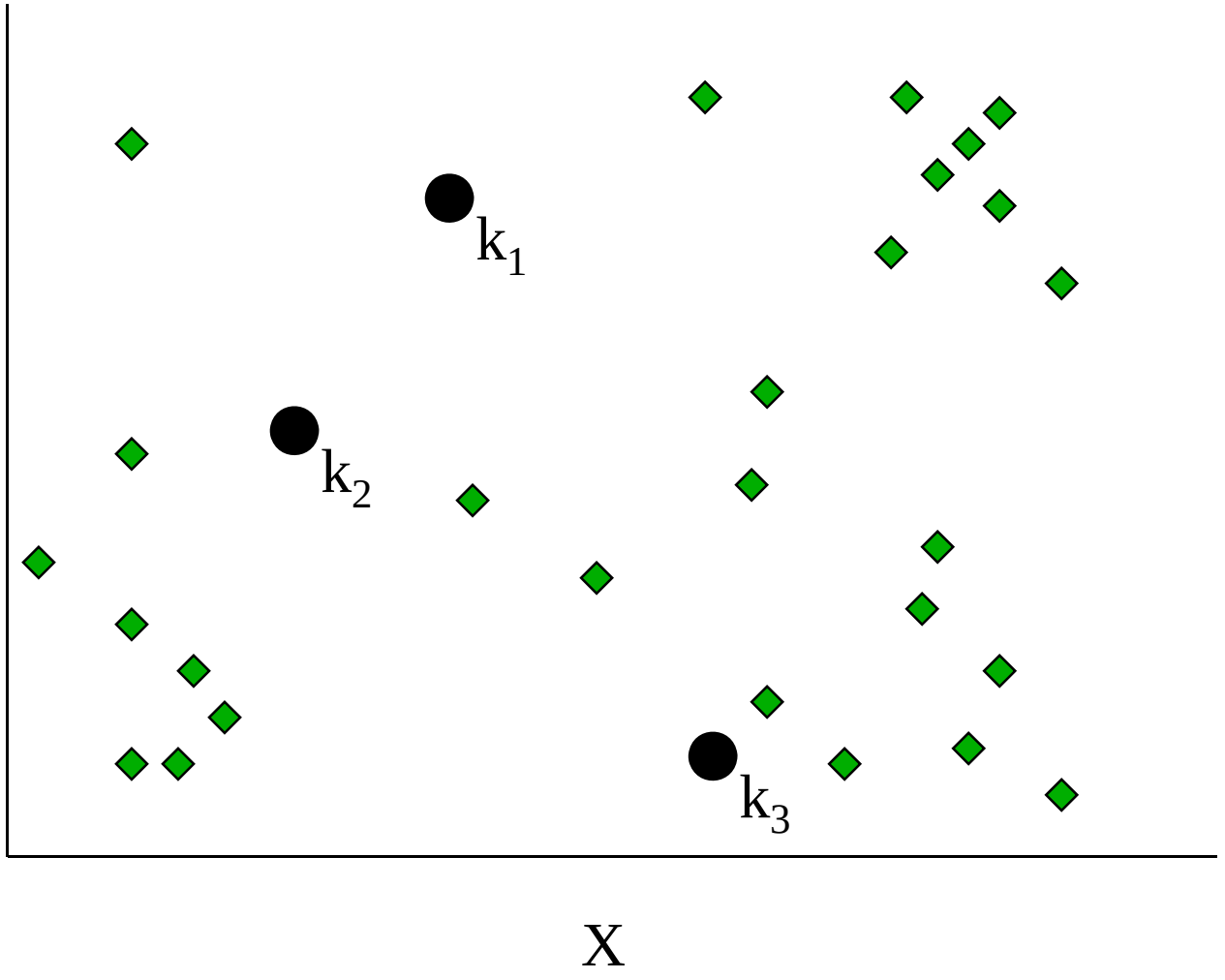
- **S(10,4)=34.105**

# L'algoritmo più semplice: K-means

- **Algorithm *K-Means*(*D*, *k*)**
  - ***m*** ← ***D.size*** // numero di istanze
  - **FOR *i*** ← **1 TO k DO**
    - $\mu_i$ ← **random** // scegli un punto medio a caso
  - **WHILE (condizione di terminazione)**
    - **FOR *j*** ← **1 TO *m* DO** // calcolo del cluster membership
    - **h** ← $\text{argmin}_{1 \le i \le k} \text{dist}(x_j, \mu_i)$
      ***C[h]*** ←$x_j$
    - **FOR *i*** ← **1 TO k DO**
      - $\mu_i \leftarrow \frac{1}{n_i} \sum_{x_j \in C[i]} x_j$
  - **RETURN *Make-Predictor* (*w*, *P*)**

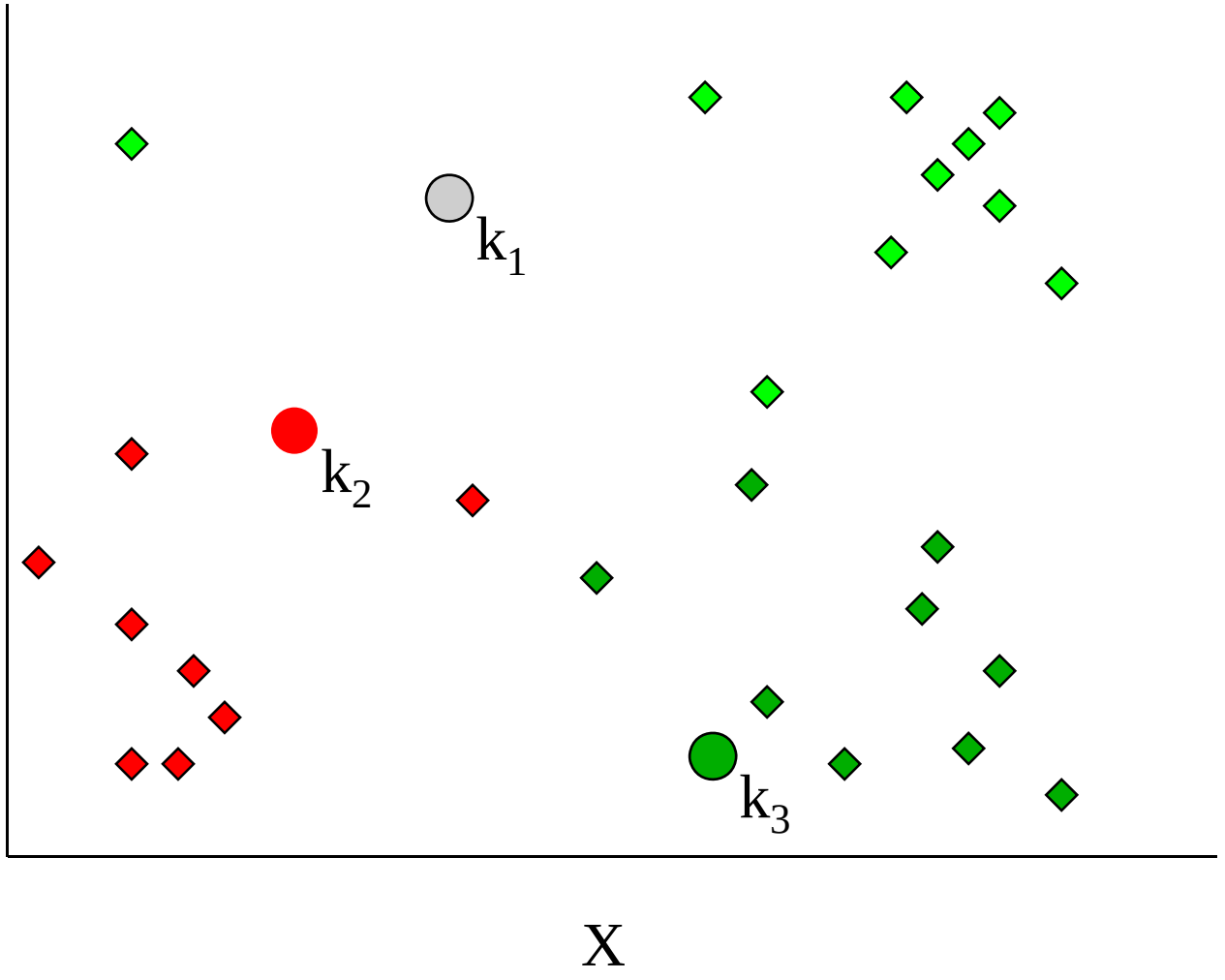
# Esempio [1]

Scegliamo
3 centri
iniziali



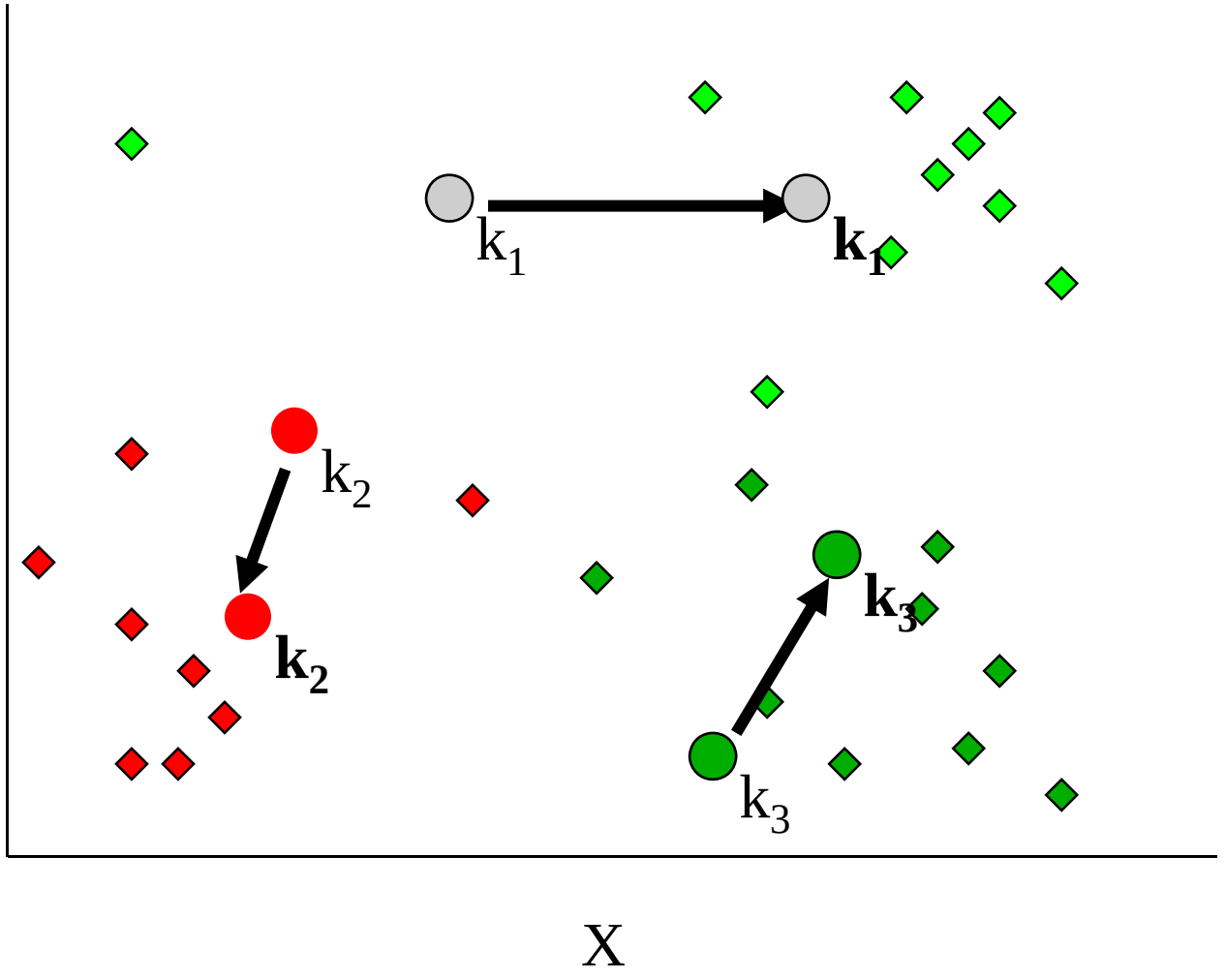
# Esempio [2]

Assegniamo ogni punto al cluster più vicino



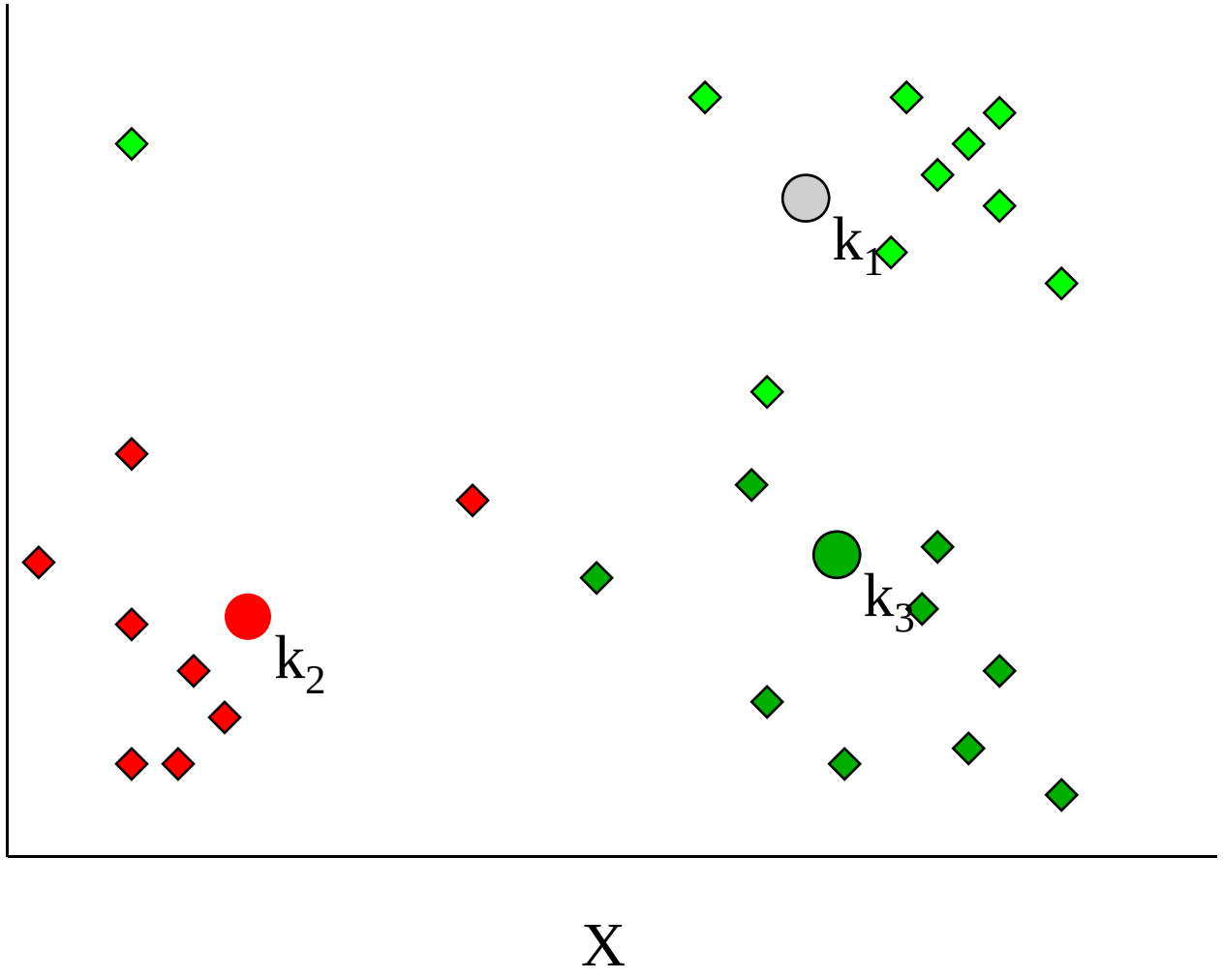
# Esempio [3]

Ricalcoliamo il centroide

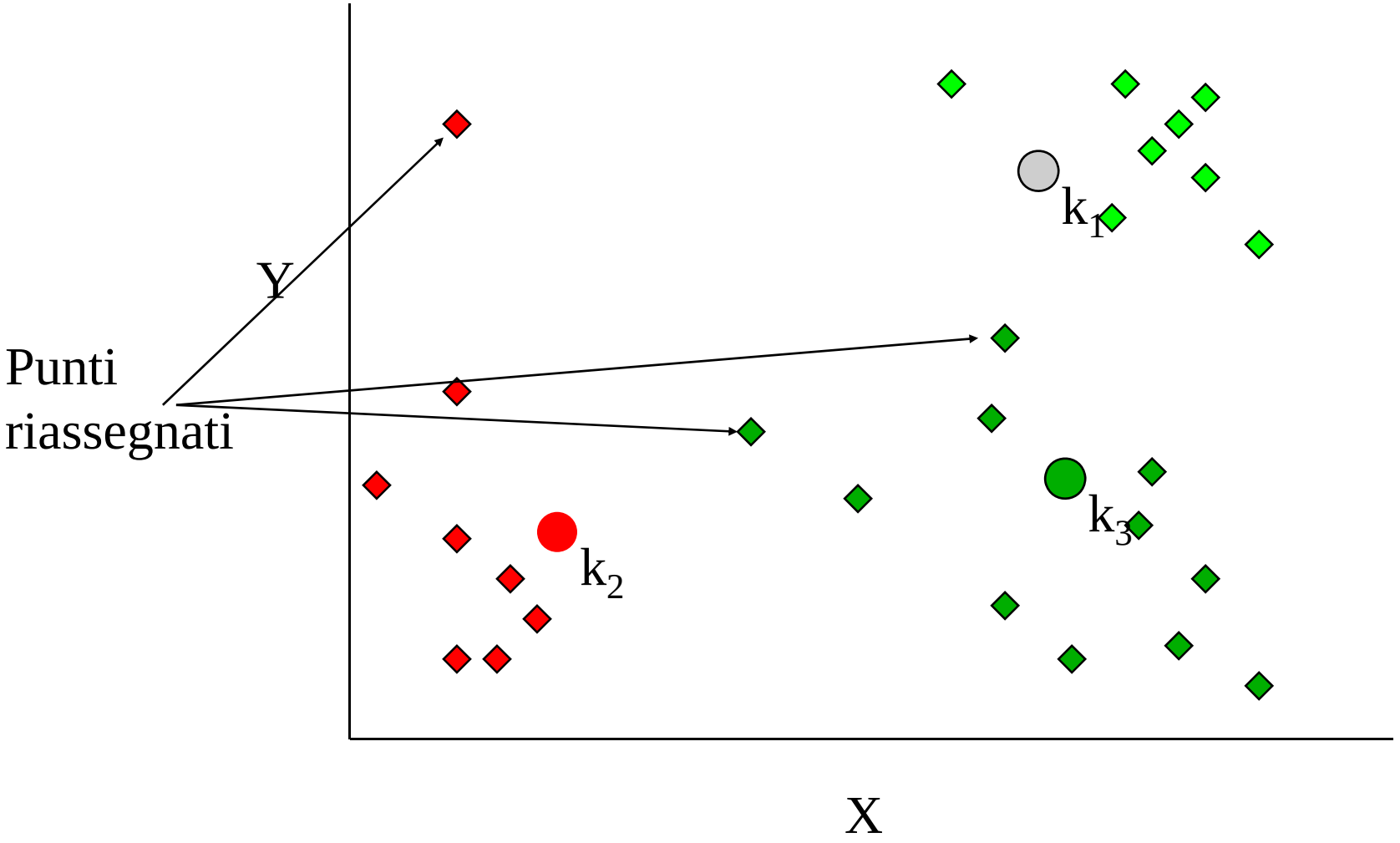


# Esempio [4]

Riassegniamo i punti ai clusters

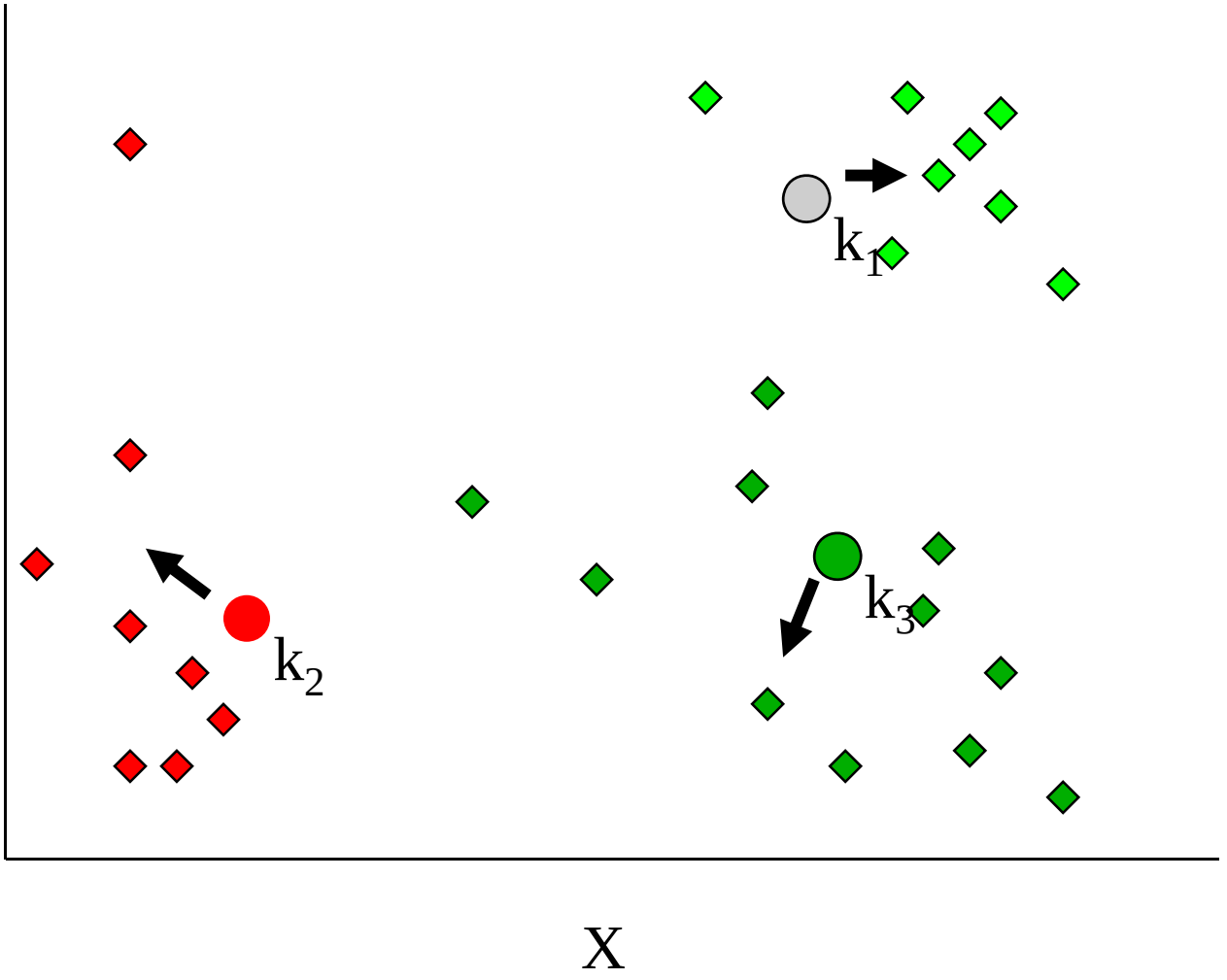


# Esempio [5]

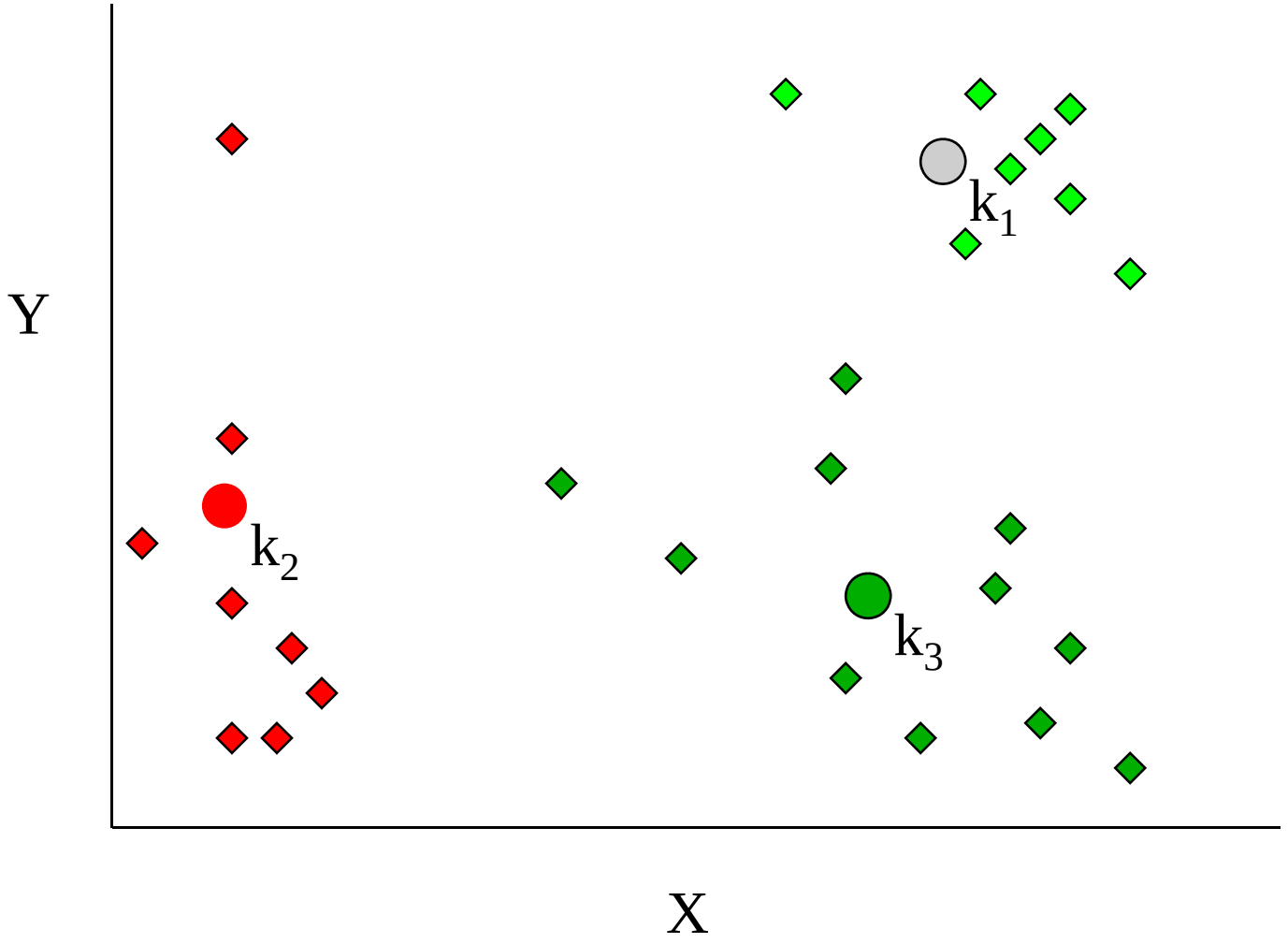


# Esempio [5]

Ricalcoliamo i centroidi



# Esempio [7]



# Quale criterio di stop?

- **Cos'è il centroide?**

$$\mu_c = \arg\min_{y \in X} \sum_{x \in C} dist(x, y)$$

- **Misura della compattezza di un cluster**

$$\cos t(C) = \sum_{x \in C} dist(x, \mu_c)$$

- **Misura della compattezza del clustering**

$$\cos t(S = [C_1, ..., C_k]) = \sum_{C_i \in S} \cos t(C_i)$$

- **Teorema**
  - **Ad una generica iterazione t,**

$$\cos t(S^t) \geq \cos t(S^{t+1})$$

# Il metodo K-Means

- **Vantaggi:** *Relativamente efficiente*: *O*(*tkn*), dove *n* è il numero di oggetti, k il numero di clusters e t il numero di iterazioni. Di solito, *k*, *t* << *n*.
  - Al confronto: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$
- Converge ad un ottimo locale. L'ottimo globale può essere trovato ricorrendo a varianti (ad esempio, basati su algoritmi genetici)
- **Punti critici**
  - Applicabile solo quando il punto medio ha senso
    - Attributi nominali?
  - K va dato in input
    - Qual'è il numero ottimale?
  - Incapace di gestire outliers e rumore
  - Non adatto a scoprire forme non convesse

# Varianti

- Selezione dei centri iniziali
- Strategie per calcolare i centroidi
- L'algoritmo *k-modes* (Huang'98)
  - Dati categorici
  - Usiamo le mode invece dei centroidi
  - Usiamo Hamming
  - L'update si basa sul cambiamento di frequenze
  - Modelli misti: *k-prototype*

# Varianti

- K-medoids **– invece del centroide, usa il punto mediano**
  - **La media di 1, 3, 5, 7, 9 è**
  - **La media di 1, 3, 5, 7, 1009 è**
  - **La mediana di 1, 3, 5, 7, 1009 è**

- **Problema: come calcolare il medoide?**

$$m_c = \arg\min_{y \in C} \sum_{x \in C} dist(x, y)$$

# Algoritmo PAM

1. Seleziona k oggetti $m_1$, …, $m_k$ arbitrariamente da D
   - $m_1$, …, $m_k$ rappresentano i medoidi
2. Assegna ogni oggetto $x \in D$ al cluster j ($1 \leq j \leq k$) che ne minimizza la distanza dal medoide
   - Calcola $cost(S)_{current}$
3. Per ogni coppia (m,x)
   1. Calcola $cost(S)_{x \leftrightarrow m}$
4. Seleziona la coppia (m,x) per cui $cost(S)_{x \leftrightarrow m}$ è minimale
5. Se $cost(S)_{x \leftrightarrow m} < cost(S)_{current}$
   - Scambia m con x
6. Ritorna al passo 3 (2)

# Problemi con PAM

- Più robusto del k-means in presenza di rumore e outliers
  - Un medoide è meno influenzato del centroide dalla presenza di outliers
- Efficiente su pochi dati, non scala su dati di grandi dimensioni.
  - $O(k(n-k)^2)$ per ogni iterazione
- Alternative basate sul campionamento
  - CLARA(Clustering LARge Applications)
  - CLARANS

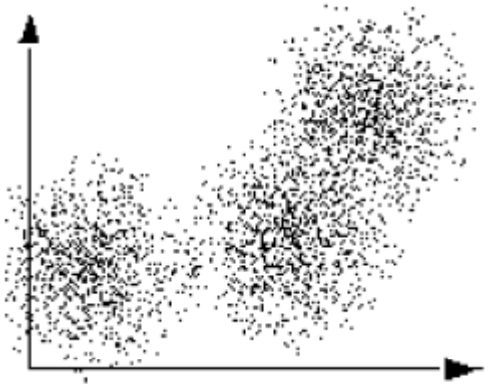
# Varianti al K-Medoid

- **CLARA [Kaufmann and Rousseeuw,1990]**
  - Parametro addizionale: *numlocal*
  - Estrae numlocal campioni dal dataset
  - Applica PAM su ogni campione
  - Restituisce il migliore insieme di medoidi come output
  - <u>Svantaggi:</u>
  - L'efficienza dipende dalla dimensione del campione
  - Non necessariamente un clustering buono sul campione rappresenta un clustering buono sull'intero dataset
    - Sample bias
- **CLARANS [Ng and Han, 1994)**
  - Due ulteriori parametri: *maxneighbor e numlocal*
  - Vengono valutate al più *maxneighbor* coppie (m,x)
  - La prima coppia (*m,x*) per cui $cost(S)_{x \leftrightarrow m} < cost(S)_{current}$ è scambiata
    - Evita il confronto con la coppia minimale
  - Si ripete la progedura *numlocal* volte per evitare il minimo locale.
- **runtime(CLARANS) < runtime(CLARA) < runtime(PAM)**

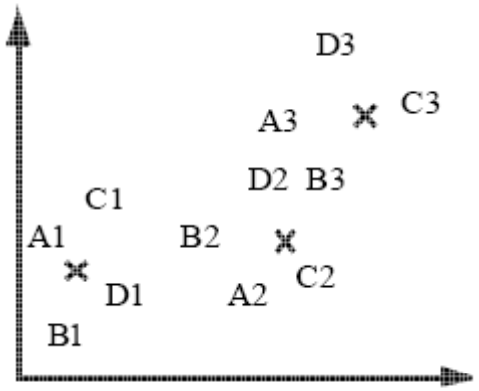
# Scelta dei punti iniziali

- [Fayyad, Reina and Bradley 1998]
  - costruisci *m campioni* differenti da D
  - Clusterizza ogni campione, fino ad ottenere m stime per i k rappresentanti
    - $A = (A_1, A_2, \ldots, A_k)$, $B = (B_1, B_2, \ldots, B_k)$, ..., $M = (M_1, M_2, \ldots, M_k)$
  - Clusterizza $DS = A \cup B \cup \ldots \cup M$ *m* volte
    - utilizzando gli insiemi *A*, *B*, ..., *M* come centri iniziali
- Utilizza il risultato migliore come inizializzazione sull'intero dataset

# Inizializzazione



D




Centri su 4 campioni

× Centri effettivi

# Scelta del parametro k

- **Iterazione**
  - Applica l'algoritmo per valori variabili di k
    - K=1, 2, …
  - Scegli il clustering "migliore"
- **Come scegliere il clustering migliore?**
  - La funzione cost(S) è strettamente decrescente
    - $K_1 > k_2 \Rightarrow cost(S_1) < cost(S_2)$
- **Indici di bontà di un cluster**

# Indici di bontà

- In generale, si valuta intra-similarità e intersimilarità

$$\gamma \sum_{C_i} \sum_{x \in C_i} \sum_{y \in C_i} dist(x,y) + (1-\gamma) \sum_{\substack{x \in C_i \\ i \neq j}} \sum_{y \in C_j} dist(x,y)$$

- a(x): distanza media di x dagli oggetti del cluster A cui x appartiene
- b(x): distanza media di x dagli oggetti del secondo cluster B più prossimo a x
- Coefficiente su x

$$s_x = \frac{a(x) - b(b)}{\max\{a(x), b(x)\}}$$

  - $s_x$=-1: x è più prossimo agli elementi di B
  - $s_x$ = 0: non c'è differenza tra A e B
  - $s_x$ = 1: x è stato clusterizzato bene
- Generalizzazione $s_c$
  - Media dell'indice su tutti gli oggetti
  - Valori bassi: clustering debole
  - Valori alti: clustering robusto

# Minimum description Length

- **Il principio del Minimum Description Length**
  - **Rasoio di Occam: le ipotesi più semplici sono più probabili**

$$\Pr(M, D) = \Pr(D \mid M)\Pr(M)$$

$$\Pr(M \mid D) = \frac{\Pr(D \mid M)\Pr(M)}{\Pr(D)}$$

$$\Pr(M \mid D) \approx \Pr(D \mid M)\Pr(M)$$

  - **(Il termine Pr(D) non dipende da M e può essere ignorato)**
- **Più alto è Pr(M|D), migliore è il modello**

# Minimum Description Length, Numero ottimale di clusters

- Fissiamo un range possibile di valori per k
  - k = 1...K (con K abbastanza alto)
- Calcoliamo $Pr(M_k|D)$ per ogni k
  - Il valore k che esibisce il valore più alto è il migliore

- Problema: Come calcoliamo $Pr(M_k|D)$?

# Bayesian Information Criterion e Minimum Description Length

- **Problema: calcolare**

$$\Pr(D \mid M)\Pr(M)$$

- **Idea: adottiamo il trick del logaritmo**

$$\log\left[\Pr(D \mid M)\Pr(M)\right] = \log\left[\Pr(D \mid M)\right] + \log\left[\Pr(M)\right]$$

- **Cosa rappresentano i due termini?**
  - $\log\left[\Pr(D \mid M)\right]$ **= accuratezza del modello**
  - $\log\left[\Pr(M)\right]$ **= complessità del modello**

# Bayesian Information Criterion

- $\log[\Pr(D \mid M)]$ = accuratezza del modello

- **Assunzione: algoritmo K-Means, distanza euclidea**

$$\text{og}[\Pr(D \mid M)] = \log\left[\prod_i p(x_i \mid M)\right] = \sum_i \log[p(x_i \mid M)] \approx -Cost(D \mid M)$$

- **La giustificazione è data dall'interpretazione "gaussiana" del clustering K-Means**

$$p(x_i \mid M) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(\frac{-d(x_i, \mu_k)^2}{{\sigma_k}^2}\right) \text{ t.c. } x_i \in C_k$$

$${\sigma_k}^2 = \sum_{x \in C_k} d(x_i, \mu_k)^2$$

$$\log[p(x_i \mid M)] \approx -\sum_k \sum_{x \in C_k} d(x_i, \mu_k)$$

# Bayesian Information Criterion

- $\log[\Pr(M)]$ **= complessità del modello**
  - **Intuitivamente, quanti più cluster ci sono, più il clustering è complesso**
    - **E quindi, meno è probabile**
  - **Adottando un criterio basato sull'encoding (in bit) dell'informazione relativa i clusters, otteniamo**

$$\log[\Pr(M)] \approx n \log k$$

    - **Per ogni tuple nel dataset, codifichiamo a quale cluster appartiene**

# Bayesian Information Criterion

$$BIC(M) = Cost(D \mid M) - \alpha n \log k$$

- **Codifica il costo di un clustering M fatto di k clusters**
  - **Direttamente proporzionale al costo del clustering dei dati**
    - **Più i cluster sono omogenei, meno costoso è il clustering**
  - **Inversamente proporzionale al costo del modello**
    - **Quanti meno clusters ci sono, meno costoso è il clustering**
    - **α è una costante di normalizzazione**
      - (serve a confrontare valori che potrebbero avere scale diverse)