

# Learning and Clustering

Alberto Borghese

Università degli Studi di Milano  
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)  
Dipartimento di Informatica  
[alberto.borghese@unimi.it](mailto:alberto.borghese@unimi.it)



A.A. 2015-2016

1/48

<http://borghese.di.unimi.it>



## Riassunto



- **I tipi di apprendimento**
- Il clustering e le feature
- Clustering gerarchico
- Clustering partitivo: K-means

A.A. 2015-2016

2/48

<http://borghese.di.unimi.it>



## I vari tipi di apprendimento



$$\begin{array}{ll} x(t+1) = f[x(t), a(t)] & \text{Ambiente} \\ a(t) = g[x(t)] & \text{Agente} \end{array}$$

**Supervisionato** (learning with a teacher). Viene specificato per ogni pattern di input, il pattern desiderato in output.

**Semi-Supervisionato**. Viene specificato solamente per **alcuni** pattern di input, il pattern desiderato in output.

**Non-supervisionato** (learning without a teacher). Estrazione di similitudine statistiche tra pattern di input. Clustering. Mappe neurali.

**Apprendimento con rinforzo** (reinforcement learning, learning with a critic). L'ambiente fornisce un'informazione puntuale, di tipo qualitativo, ad esempio success or fail.



## I gruppi di algoritmi



Clustering (data mining)

Classification

Predictive regression



## Riassunto



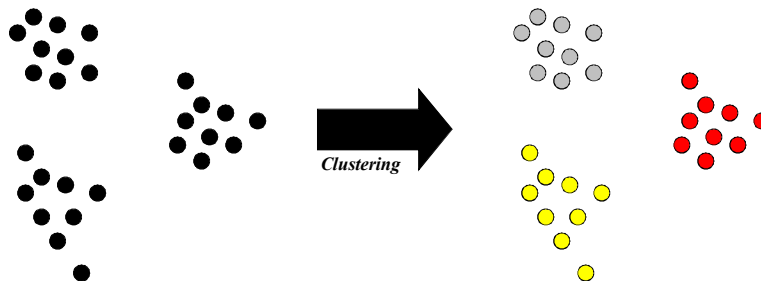
- I tipi di apprendimento
- **Il clustering e le feature**
- Clustering gerarchico
- Clustering partitivo: K-means



## Clustering



- Clustering: raggruppamento degli “oggetti” in cluster omogenee tra loro. Gli oggetti di un cluster sono più “simili” tra loro che a quelli degli altri cluster.
  - ◆ Raggruppamento per colore
  - ◆ Raggruppamento per forme
  - ◆ Raggruppamento per tipi
  - ◆ .....

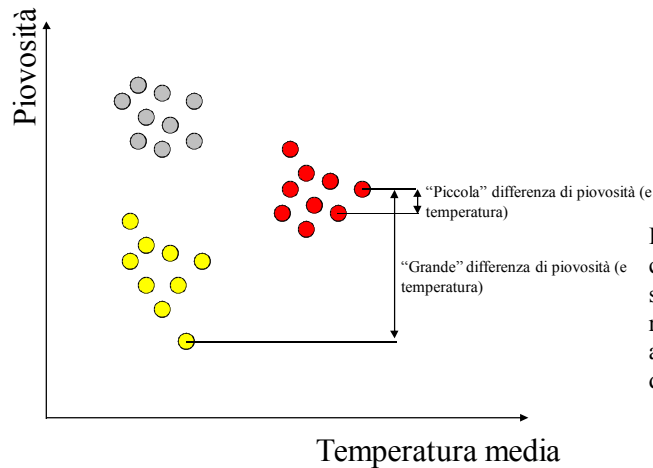


Novel name: **data mining**



## Clustering

L'elaborazione verrà poi effettuata sui prototipi che rappresentano ciascun cluster.



I pattern appartenenti ad un cluster valido sono più simili l'uno con l'altro rispetto ai pattern appartenenti ad un cluster differente.



## Il clustering per...

- ... Confermare ipotesi sui dati (es. “E’ possibile identificare tre diversi tipi di clima in Italia: mediterraneo, continentale, alpino...”);
- ... Esplorare lo spazio dei dati (es. “Quanti tipi diversi di clima sono presenti in Italia? Quante sfere sono presenti in un’immagine?”);
- ... Semplificare l’interpretazione dei dati (“Il clima di ogni città d’Italia è approssimativamente mediterraneo, continentale o alpino.”).
- ... “Ragionare” sui dati o elaborare i dati in modo stereotipato.



## Esempio di clustering



Ricerca immagini su WEB.



Clustering -> Indicizzazione

A.A. 2015-2016

9/48

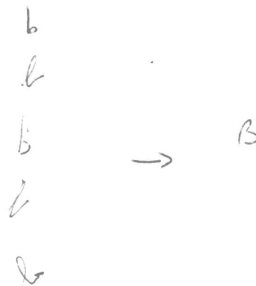
<http://borghese.di.unimi.it>



## Clustering: definizioni



- **Pattern:** un singolo dato  $\mathbf{X} = [x_1, x_2, \dots, x_D]$ . Il dato appartiene quindi ad uno spazio multi-dimensionale ( $D$  dimensionale), solitamente eterogeneo.
- **Feature:** le caratteristiche dei dati significative per il clustering, possono costituire anch'esso un vettore, il vettore delle feature:  $f_1, f_2, \dots, f_M$ . Questo vettore costituisce l'input agli algoritmi di clustering.



Inclinazione, occhielli,  
lunghezza, linee  
orizzontali, archi di cerchio  
...

A.A. 2015-2016

<http://borghese.di.unimi.it>



## Clustering: definizioni

- **D**: dimensione dello spazio dei pattern;
- **M**: dimensione dello spazio delle feature;
- **Cluster**: in generale, insieme che raggruppa dati simili tra loro, valutati in base alle feature;
- **Funzione di similarità o distanza**: una metrica (o quasi metrica) nello spazio delle feature, usata per quantificare la similarità tra due pattern.
- **Algoritmo**: scelta di come effettuare il clustering (motore di clustering).



## Clustering

- Dati,  $\{X_1 \dots X_N\} \in R^D$
- Cluster  $\{C_1 \dots C_M\} \rightarrow \{P_1 \dots P_M\} \in R^D$

$P_j$  is the prototype of cluster  $j$  and it represents the set of data inside its cluster.

To cluster the data:

- The set of data inside each cluster has to be determined (the boundary of a cluster defined)
- The cluster boundaries are determined considering features associated to the data.



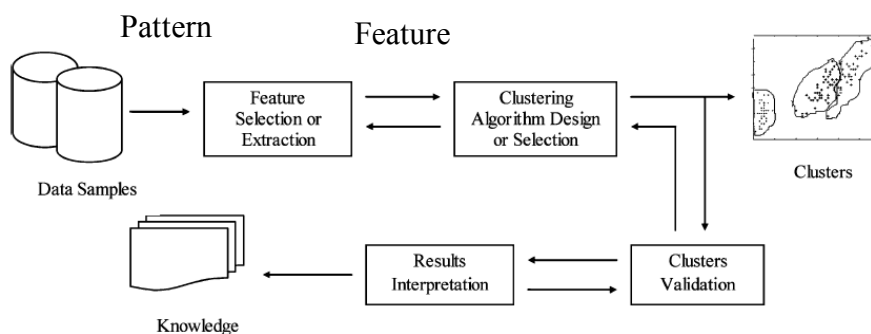
## Tassonomia (sintetica) degli algoritmi di clustering



- Algoritmi gerarchici (agglomerativi, divisivi), e.g. **Hierarchical clustering**.
- Algoritmi partizionali, hard: **K-means, quad-tree decomposition**.
- Algoritmi partizionali, **soft-clustering**: fuzzy c-mean, neural-gas, enhanced vector quantization, **mappe di Kohonen**.
- Algoritmi statistici: **mixture models**.



## Analisi mediante clustering



Da Xu and Wunsch, 2005

I cluster ottenuti sono significativi?

Il clustering ha operato con successo?

NB i cammini all'indietro consentono di fare la sintonizzazione dei diversi passi.



## Il clustering

Per una buona review: Xu and Wunsch, IEEE Transactions on Neural Networks, vol. 16, no. 3, 2005.

Il clustering non è di per sé un problema ben posto. Ci sono diversi gradi di libertà da fissare su come effettuare un clustering.

Rappresentazione dei pattern;

Calcolo delle feature;

Definizione di una misura di prossimità dei pattern attraverso le feature;

Tipo di algoritmo di clustering (gerarchico o partizionale)

Validazione dell'output (se necessario) -> Testing.



Problema a cui non risponderemo: **quanti cluster?** Soluzione teorica (criterio di Akaike), soluzione empirica (growing networks di Fritzke).



## Features


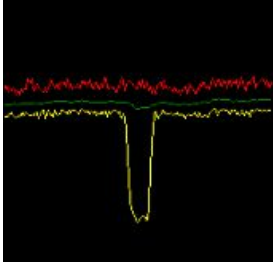

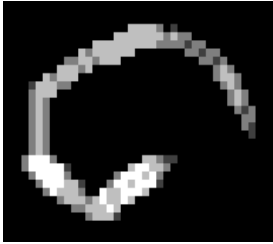
- Globali: livello di luminosità medio, varianza, contenuto in frequenza.....
- Feature locali



## Features


- *Località.*
- *Significatività.*
- *Rinoscibilità.*

Macchie dense		
Fili		

A.A. 2015-2016

17/48

<http://borghese.di.unimi.it>

## Rappresentazione dei dati

- La similarità tra dati viene valutata attraverso le feature.
- Feature selection: identificazione delle feature più significative per la descrizione dei pattern.

Esempio: descrizione del clima e della città di Roma.  
Roma è caratterizzata da: [17°; 500mm; 1.500.000 ab., 300 chiese]

- Quali feature scegliere?
- Come valutare le feature?
  - ◆ Analisi statistica del potere discriminante: correlazione tra feature e loro significatività.

A.A. 2015-2016

18/48

<http://borghese.di.unimi.it>



## Similarità tra feature



- Definizione di una **misura di distanza tra due features**;

Esempio:

Distanza euclidea...

$\text{dist}(\text{Roma}, \text{Milano}) = \text{dist}([17^\circ; 500\text{mm}], [13^\circ; 900\text{mm}]) = \dots$

$= \dots \text{Distanza euclidea?} = ((17-13)^2 + (500-900)^2)^{1/2} = 400.02 \sim 400$

Ha senso?



## Normalizzazione feature



**E' necessario trovare una metrica corretta per la rappresentazione dei dati. Ad esempio, normalizzare le feature!**

$$T_{\text{Max}} = 20^\circ \quad T_{\text{Min}} = 5^\circ \rightarrow T_{\text{Norm}} = (T - T_{\text{Min}}) / (T_{\text{Max}} - T_{\text{Min}})$$

$$P_{\text{Max}} = 1000\text{mm} \quad P_{\text{Min}} = 0\text{mm} \rightarrow P_{\text{Norm}} = (P - P_{\text{Min}}) / (P_{\text{Max}} - P_{\text{Min}})$$

$$\text{Roma}_{\text{Norm}} = [0.8 \ 0.5]$$

$$\text{Milano}_{\text{Norm}} = [0.53 \ 0.9]$$

$$\text{dist}(\text{Roma}_{\text{Norm}}, \text{Milano}_{\text{Norm}}) = ((0.8-0.53)^2 + (0.5-0.9)^2)^{1/2} = 0.4826$$

E' una buona scelta?



## Altre funzioni di distanza



- Mahalanobis:  
 $\text{dist}(x,y) = (x_k - y_k) S^{-1} (x_k - y_k)$ , con S matrice di covarianza.  
(Normalizzazione mediante covarianza)

Altre metriche:

- Distanza euclidea:  
 $\text{dist}(x,y) = [\sum_{k=1..d} (x_k - y_k)^2]^{1/2}$
- Minkowski:  
 $\text{dist}(x,y) = [\sum_{k=1..d} (x_k - y_k)^p]^{1/p}$
- Context dependent:  
 $\text{dist}(x,y) = f(x, y, \text{context})$



## Riassunto



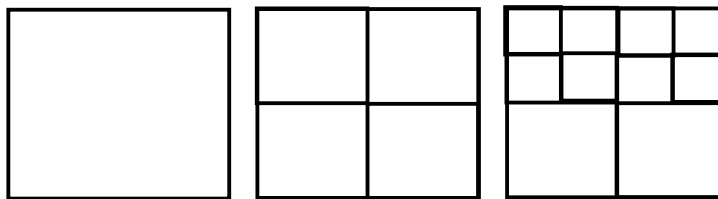
- I tipi di apprendimento
- Il clustering e le feature
- **Clustering gerarchico**
- Clustering partitivo: K-means



## Algoritmi gerarchici divisivi: QTD



- Quad Tree Decomposition;
- Suddivisione gerarchica dello spazio delle feature, mediante splitting dei cluster;
- Criterio di splitting ( $\sim$ distanza tra cluster).



A.A. 2015-2016

23/48

<http://borghese.di.unimi.it>



## Algoritmi gerarchici: QTD



- Clusterizzazione immagini RGB, 512x512;
- Pattern: pixel (x,y);
- Feature: canali R, G, B.
- Distanza tra due pattern (non euclidea):  
 $\text{dist}(p_1, p_2) =$   
 $\text{dist}([R_1 \ G_1 \ B_1], [R_2 \ G_2 \ B_2]) =$   
 $\max(|R_1 - R_2|, |G_1 - G_2|, |B_1 - B_2|).$

A.A. 2015-2016

24/48

<http://borghese.di.unimi.it>



## Algoritmi gerarchici: QTD



$$p1 = [0 \ 100 \ 250]$$

$$p2 = [50 \ 100 \ 200]$$

$$p3 = [255 \ 150 \ 50]$$

$$\text{dist}(p1, p2) = \text{dist}([R1 \ G1 \ B1], [R2 \ G2 \ B2]) = \\ \max(|R1-R2|, |G1-G2|, |B1-B2|) = \max([50 \ 0 \ 50]) = 50.$$

$$\text{dist}(p2, p3) = 205.$$

$$\text{dist}(p3, p1) = 255.$$

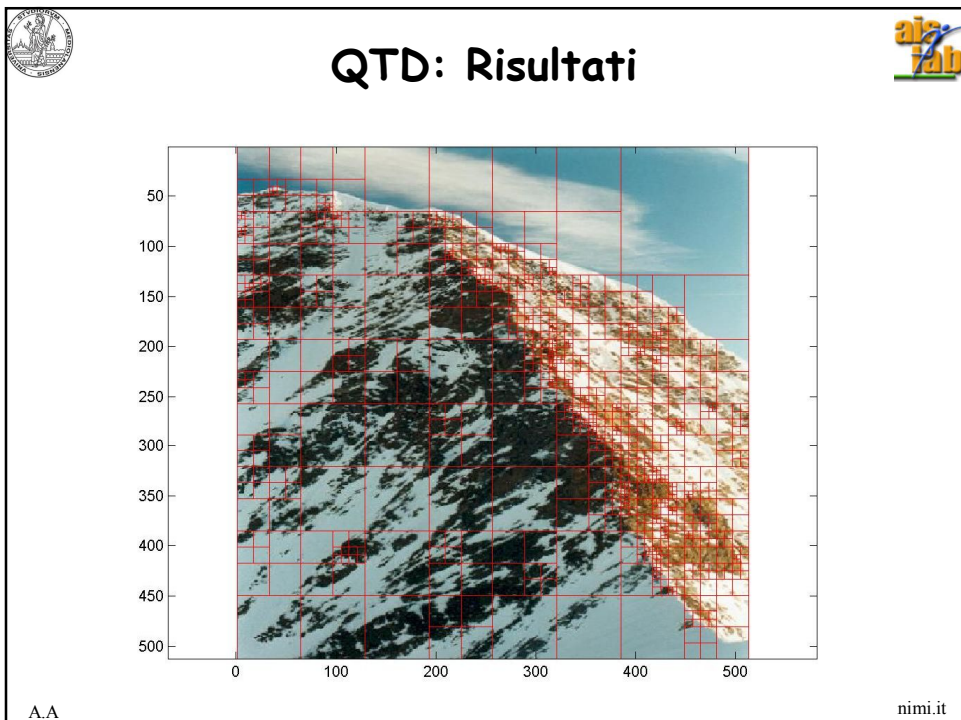
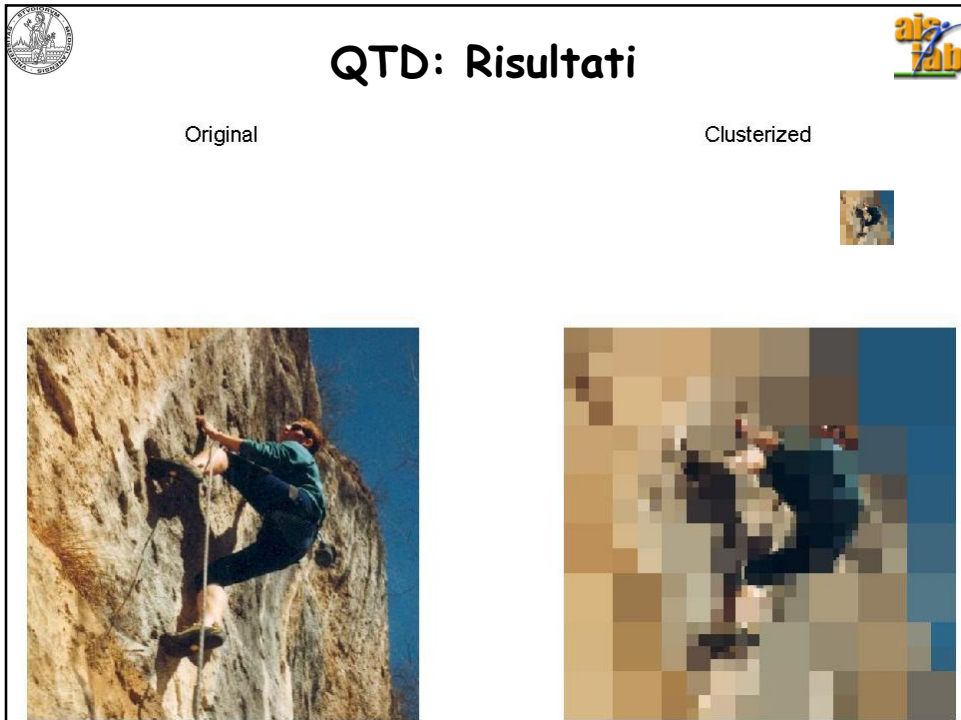


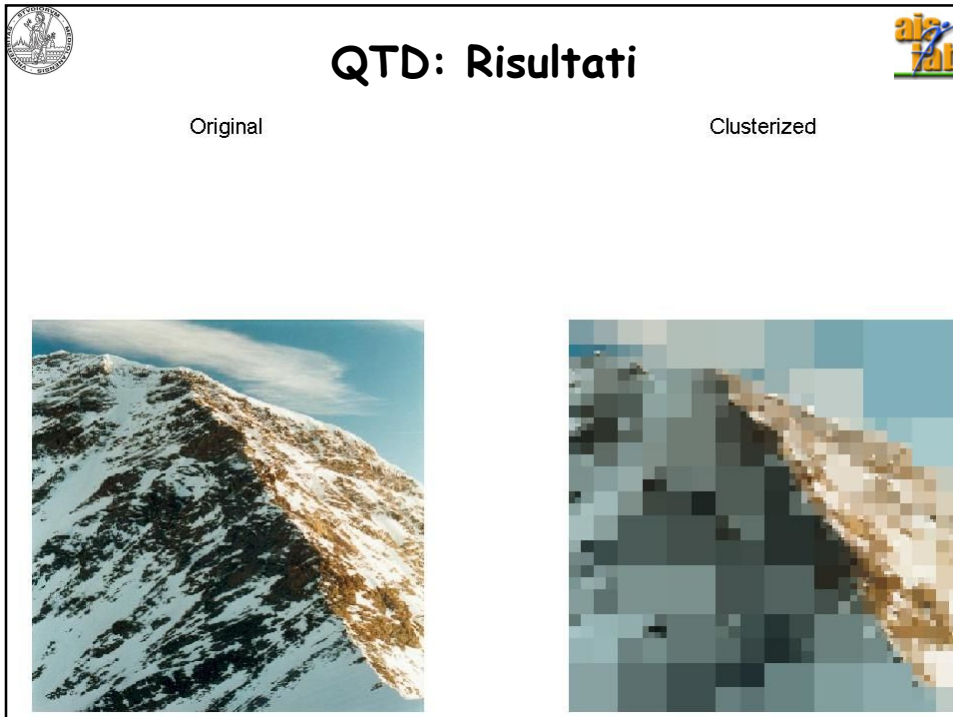
## Algoritmi gerarchici: QTD



Criterio di splitting: se due pixel all'interno dello stesso cluster distano più di una determinata soglia, il cluster viene diviso in 4 cluster.

Esempio applicazione: segmentazione immagini, compressione immagini, analisi locale frequenze immagini...





## Hierarchical Clustering

- In brief, HC algorithms build a whole hierarchy of clustering solutions
  - ◆ Solution at level  $k$  is a *refinement* of solution at level  $k-1$
- Two main classes of HC approaches:
  - ◆ Agglomerative: solution at level  $k$  is obtained from solution at level  $k-1$  by merging two clusters
  - ◆ Divisive: solution at level  $k$  is obtained from solution at level  $k-1$  by splitting a cluster into two parts
    - ⇒ Less used because of computational load

A.A. 2015-2016 30/48 <http://borghese.di.unimi.it>



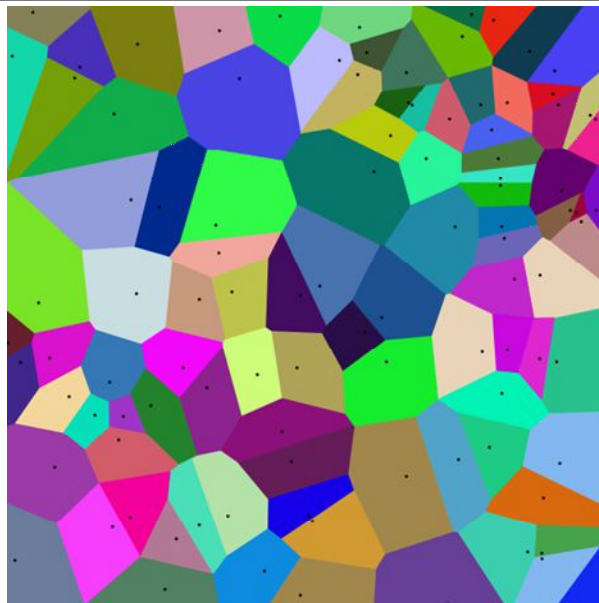
## Riassunto



- I tipi di apprendimento
- Il clustering e le feature
- Clustering gerarchico
- **Clustering partitivo: K-means**



**Risultato del clustering è  
un diagramma di Voronoj**



I poligoni azzurri rappresentano i diversi cluster ottenuti. Ogni punto marcato all'interno del cluster (cluster center) è rappresentativo di tutti i punti del cluster





## K-means (partitional): framework



- Siano  $X_1, \dots, X_D$  i dati di addestramento, features (per semplicità, definiti in  $R^2$ );
- Siano  $C_1, \dots, C_K$  i *prototipi* di  $K$  classi, definiti anch'essi in  $R^2$ ; ogni *prototipo* identifica il baricentro della classe corrispondente;
- Lo schema di classificazione adottato sia il seguente: “ $X_i$  appartiene a  $C_j$  se e solo se  $C_j$  è il *prototipo* più vicino a  $X_i$  (distanza euclidea)”;
- L'algoritmo di addestramento permette di determinare le posizioni dei *prototipi*  $C_j$  mediante successive approssimazioni.



## Algoritmo K-means



L'obiettivo che l'algoritmo si prepone è di minimizzare la varianza totale intra-cluster. Ogni cluster viene identificato mediante un centroide o punto medio. L'algoritmo segue una procedura iterativa. Inizialmente crea  $K$  partizioni e assegna ad ogni partizione i punti d'ingresso o casualmente o usando alcune informazioni euristiche. Quindi calcola il centroide di ogni gruppo. Costruisce quindi una nuova partizione associando ogni punto d'ingresso al cluster il cui centroide è più vicino ad esso. Quindi vengono ricalcolati i centroidi per i nuovi cluster e così via, finché l'algoritmo non converge (Wikipedia).



## K-means: addestramento

Inizializzazione  $C_j$

Classificazione  $X_i$

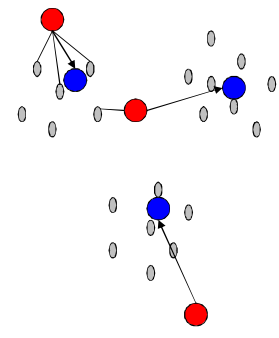
Aggiornamento  $C_j$

I prototipi  $C_j$  si sono spostati significativamente?

NO

Termine addestramento

SI



Aggiornamento  $C_j$ : baricentro degli  $X_i$  classificati da  $C_j$ .



## Algoritmo K-means::formalizzazione

- Dati  $N$  pattern in ingresso  $\{x_j\}$  e  $C_k$  prototipi che vogliamo diventino i centri dei cluster,  $x_j$  e  $C_k \in \mathbb{R}^N$ . Ciascun cluster identifica una regione nello spazio,  $P_k$ .
- Valgono le seguenti proprietà:

$$\bigcup_{k=1}^K P_k = Q \supseteq \mathbb{R}^D \quad \text{I cluster coprono lo spazio delle feature}$$

$$\bigcap_{k=1}^K P_k = \emptyset \quad \text{I cluster sono disgiunti.}$$

$$x_j \in C_k \quad \text{Se: } \left( |x_j - C_k| \right)^2 \leq \left( |x_j - C_l| \right)^2 \quad l \neq k$$

$$\text{La funzione obiettivo viene definita come: } \sum_{i=1}^K \sum_{j=1}^N \left( |x_{j^{(k)}} - C_k| \right)^2$$



## Algoritmo K-means: dettaglio dei passi



- Inizializzazione.
  - ◆ Posiziono in modo arbitrario o guidato i K centri dei cluster.
- Iterazioni
  - ◆ Assegno ciascun pattern al cluster il cui centro è più vicino, formando così un certo numero di cluster ( $\leq K$ ).
  - ◆ Calcolo la posizione dei cluster,  $C_k$ , come baricentro dei pattern assegnati ad ogni cluster, spostando quindi la posizione dei centri dei cluster.
- Condizione di uscita
  - I centri dei cluster non si spostano più.

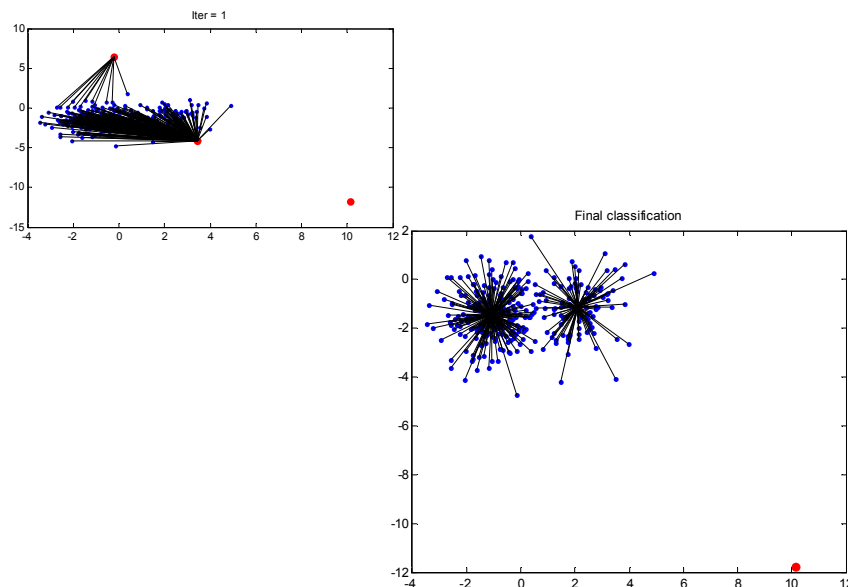
A.A. 2015-2016

37/47

<http://borghese.di.unimi.it/>



## Bad initialization



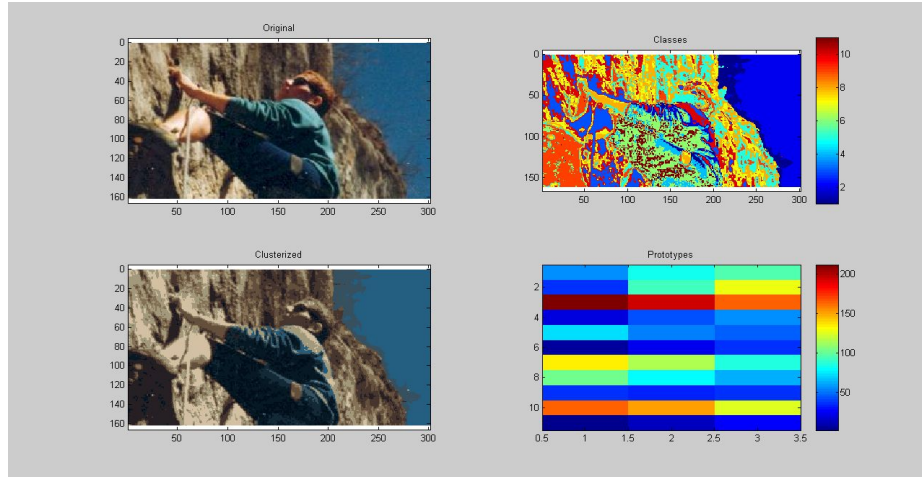
A.A. 2015-2016

38/47

<http://borghese.di.unimi.it/>



## K-Means per immagine RGB



Da 255 colori a 33 colori



## Riassunto

- I tipi di apprendimento
- Il clustering e le feature
- Clustering gerarchico
- Clustering partitivo: K-means