

Sistemi Intelligenti Learning and Clustering

Alberto Borghese and Iuri Frosio

Università degli Studi di Milano
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)
Dipartimento di Scienze dell'Informazione
borghese@dsi.unimi.it



A.A. 2009-2010

1/50

<http://homes.dsi.unimi.it/~borghese/>



Riassunto



- I tipi di apprendimento
- Il clustering
- K means
- Quad-tree decomposition

A.A. 2009-2010

2/50

<http://homes.dsi.unimi.it/~borghese/>



I vari tipi di apprendimento

$$x(t+1) = f[x(t), a(t)] \quad \text{Ambiente}$$

$$a(t) = g[x(t)] \quad \text{Agente}$$

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.

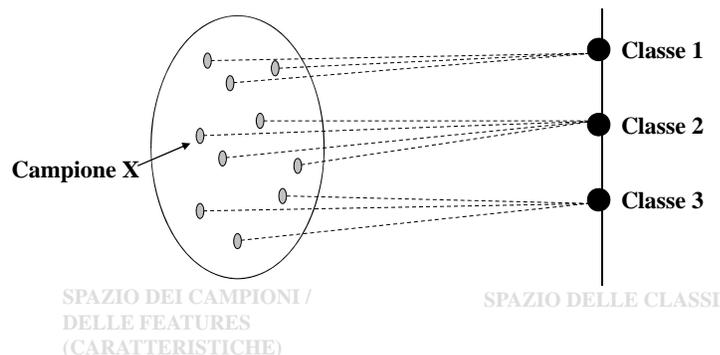
Apprendimento con rinforzo (reinforcement learning, learning with a critic). L'ambiente fornisce un'informazione puntuale, di tipo qualitativo, ad esempio success or fail.



Classificazione

Un'interpretazione geometrica:

Mappatura dello spazio dei campioni nello spazio delle classi.





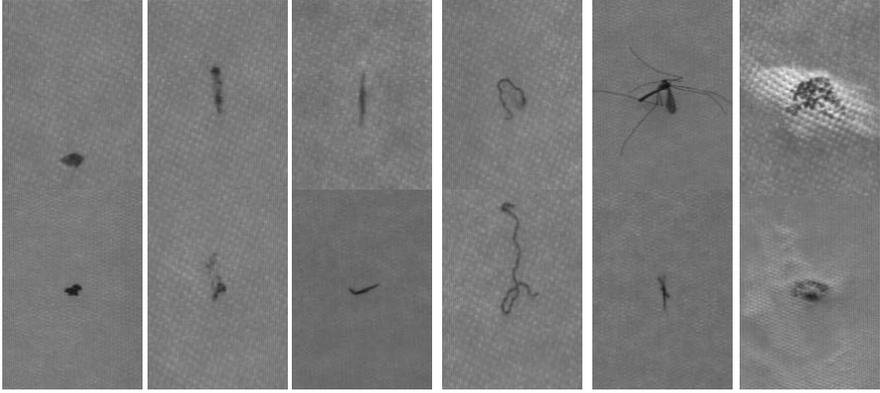

Apprendimento Supervisionato: Classificazione

Task di classificazione
Uscita intera (etichetta o label della classe)

A.A. 2009-20 <http://homes.dsi.unimi.it/~borghese/>




CLASSIFICAZIONE: Riconoscimento difetti in linee di produzione
(progetto finanziato da Electronic Systems: 2006-2007)



regolari irregolari allungati fili insetti macchie su denso

Difetti – Classificazione real-time e apprendimento.

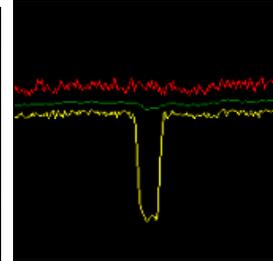
A.A. 2009-2010 6/50 <http://homes.dsi.unimi.it/~borghese/>



Features

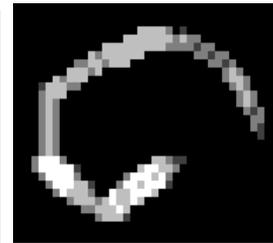


Macchie
dense



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

Fili



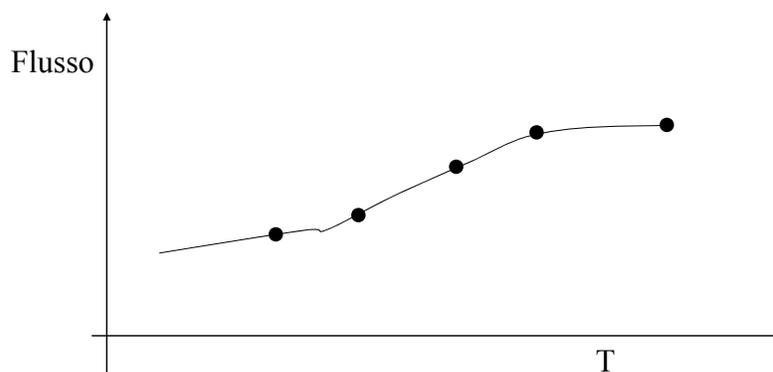
A.A. 2009-2010

7/50

<http://homes.dsi.unimi.it/~borghese/>



Apprendimento supervisionato: **regressione**

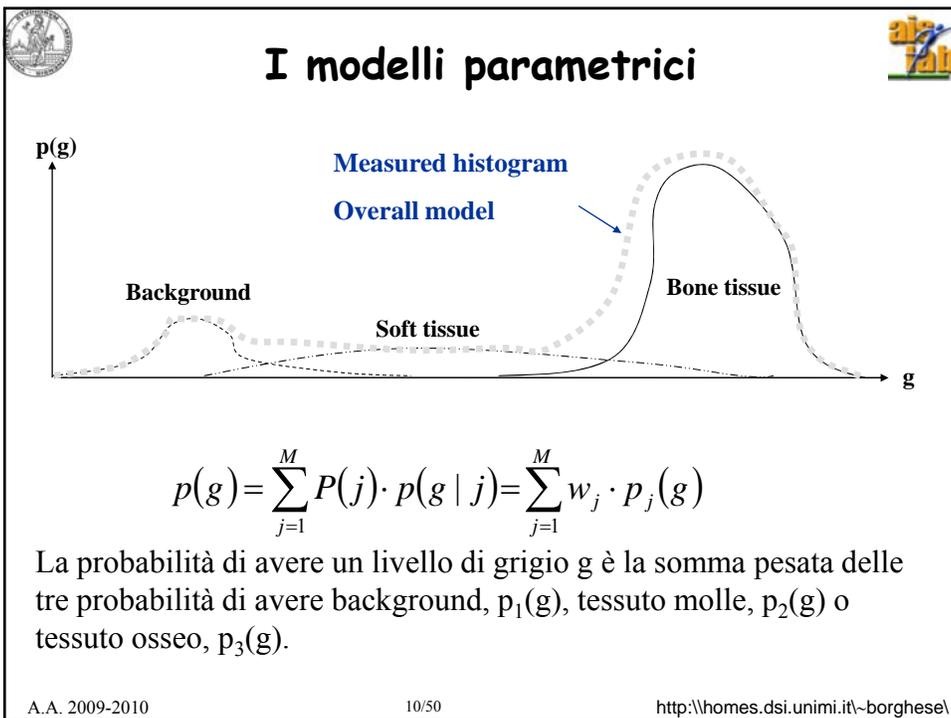
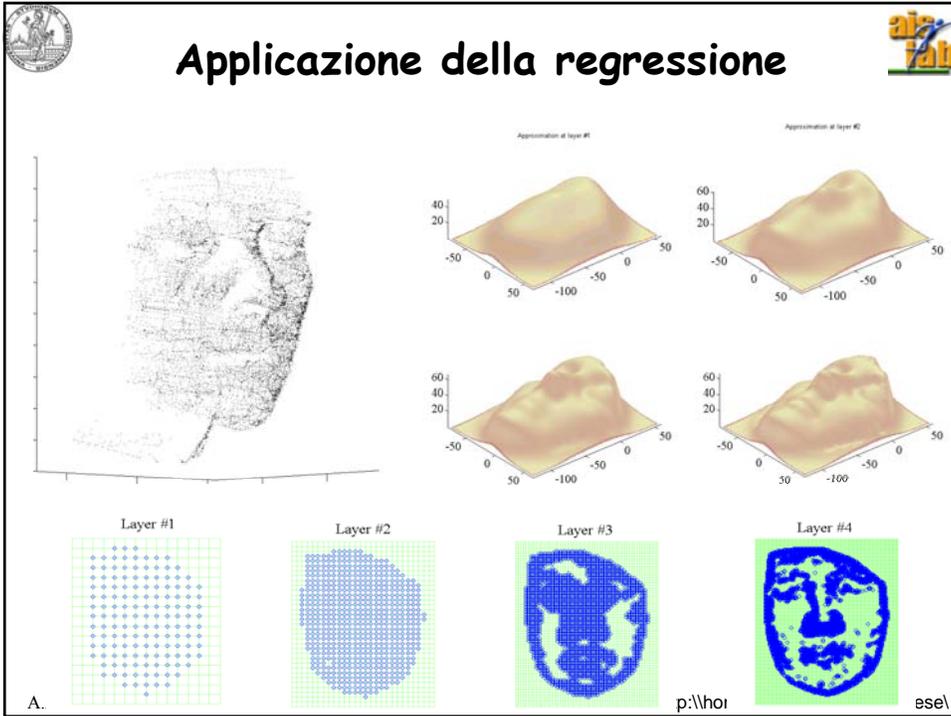


Controllo della portata di un condizionatore in funzione della temperatura. “Imparo” una funzione continua a partire da alcuni campioni: devo imparare ad **interpolare**.

A.A. 2009-2010

8/50

<http://homes.dsi.unimi.it/~borghese/>



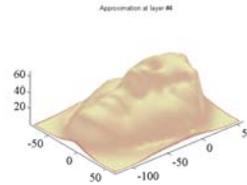


I modelli semi-parametrici



- L'approssimazione è ottenuta mediante funzioni “generiche”, dette di **base**, soluzione molto utilizzata nelle NN e in Machine learning.
- (Il concetto di Base in matematica è definito mediante certe proprietà di approssimazione che qui non consideriamo, consideriamo solo l'idea intuitiva).

$$z(p(x, y)) = \sum_i w_i G(p, p_i; \sigma)$$



Combinazione
lineare di funzioni
di base

Funzione di base



Modelli lineari e non lineari



Classificazione alternativa dei modelli. Vengono utilizzate classi molto diversi di algoritmi per stimare i parametri di questi due tipi di modelli.

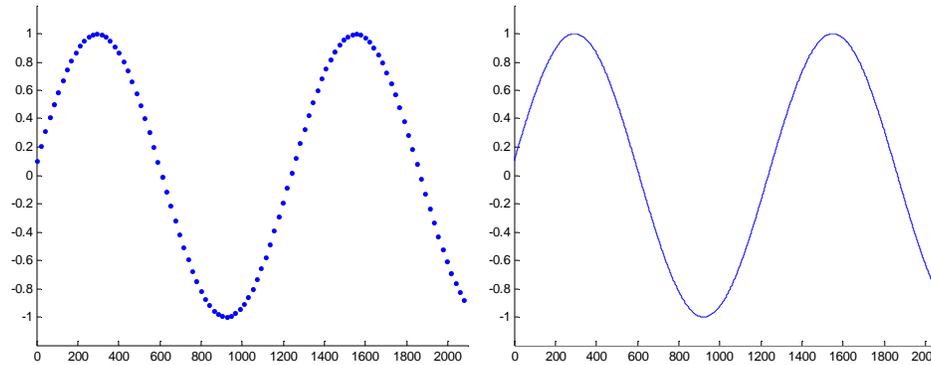
$$z(p(x, y)) = \sum_i w_i G(p, p_i; \sigma)$$

$$z(p(x, y)) = \sum_i f_i(G(p, p_i; \sigma))$$

$f(\cdot)$ è funzione non lineare



Funzionamento di un modello parametrico (non-lineare)



I punti vengono fittati perfettamente da una sinusoide: $y = \sin(\omega x + \phi)$,
a patto di definire: $\omega = 1/200$, $\phi = 0.1$.

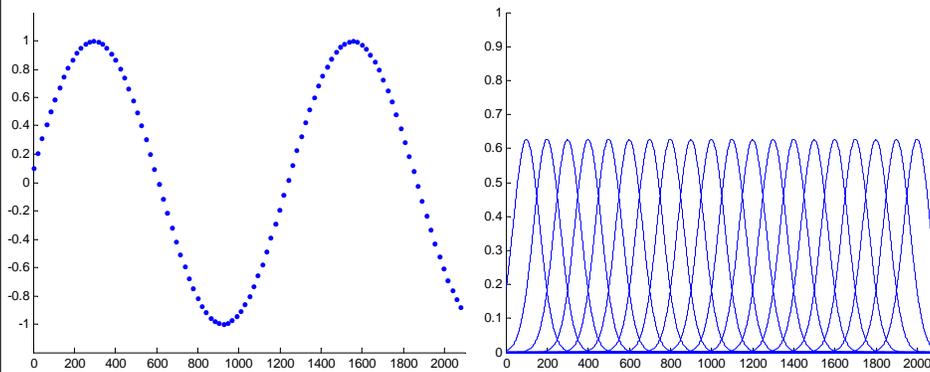
A.A. 2009-2010

13/50

<http://homes.dsi.unimi.it/~borghese/>



Approssimazione mediante un modello semi-parametrico (lineare)



Vogliamo fittare i punti con l'insieme di Gaussiane riportate sulla dx. In questo caso hanno tutte $\sigma = 90$. Come le utilizzo?

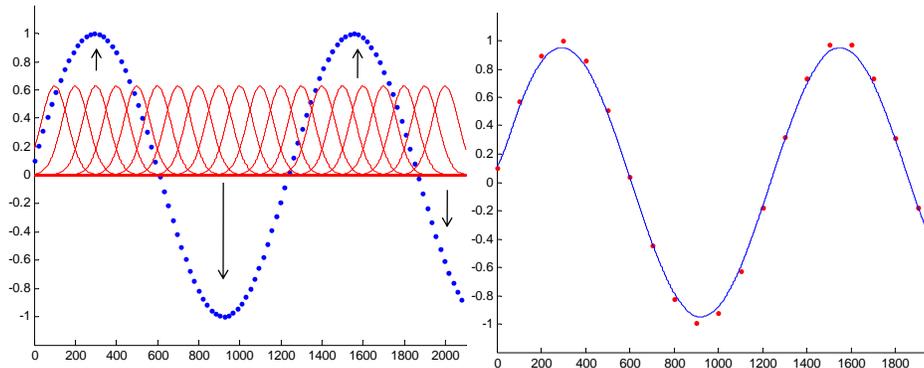
A.A. 2009-2010

14/50

<http://homes.dsi.unimi.it/~borghese/>



Funzionamento di un modello semi-parametrico (lineare)



$$y(x) = \sum_{i=1}^{20} w_i G(x - x_{o_i}, \sigma)$$

Devo definire, i $\{w_i\}$

I σ sono tutti uguali ed uguali a 90, le Gaussiane sono equispaziate.

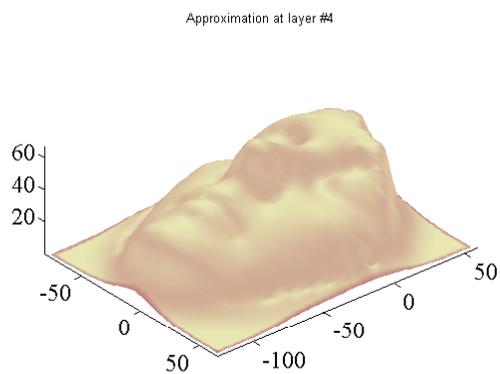
A.A. 2009-2010

15/50

<http://homes.dsi.unimi.it/~borghese/>



Quante unità semi-parametriche? (rischio di overfitting)



A.A. 2009-2010

16/50

<http://homes.dsi.unimi.it/~borghese/>



Riassunto



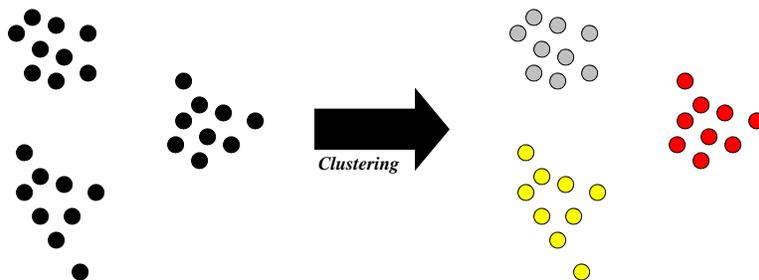
- I tipi di apprendimento
- **Il clustering**
- K means
- Quad-tree decomposition



Apprendimento non-supervisionato



- Clustering: raggruppamento degli “oggetti” in classi omogenee tra loro.
 - ◆ Raggruppamento per colore
 - ◆ Raggruppamento per forme
 - ◆ Raggruppamento per tipi
 - ◆





Esempio di clustering



Ricerca immagini su WEB.



Clustering -> Indicizzazione



Il clustering per...



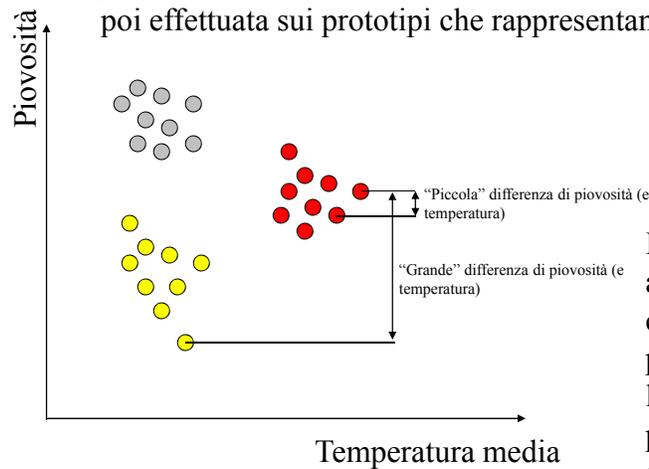
- ... Confermare ipotesi sui dati (es. “E’ possibile identificare tre diversi tipi di clima in Italia: mediterraneo, continentale, alpino...”);
- ... Esplorare lo spazio dei dati (es. “Quanti tipi diversi di clima sono presenti in Italia?”);
- ... Semplificare l’interpretazione dei dati (“Il clima di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



Clustering



Processo attraverso il quale i dati (pattern, vettori) vengono organizzati in cluster, basata sulla similarità. 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.

A.A. 2009-2010

21/50



Clustering: definizioni



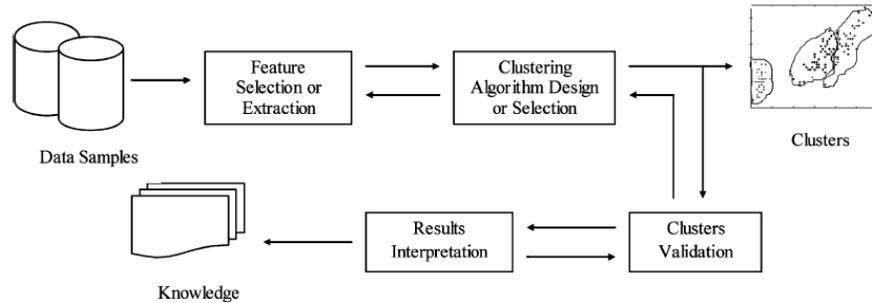
- **Pattern:** un singolo dato $x = [x_1, x_2, \dots, x_d]$. Il dato appartiene quindi ad uno spazio multi-dimensionale, solitamente eterogeneo.
- **Feature:** ogni componente del vettore dei pattern x_1, x_2, \dots, x_d ;
- **d:** dimensione dello spazio dei pattern;
- **Classe:** in generale, un processo che governa la generazione dei pattern di un determinato cluster (in particolare, la ddp dei dati di un cluster);
- **Funzione di 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).

A.A. 2009-2010

22/50



Analisi mediante clustering



Da Xu and Wunsch, 2005

I cluster ottenuti sono significativi?
Il clustering ha operato con successo?



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;
Definizione di una misura di prossimità dei pattern;
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 network di Fritzke).



Rappresentazione dei pattern



- Feature selection: identificazione delle feature più significative per la descrizione dei pattern.

Esempio: descrizione del clima e della città di Roma.

Roma: [17°; 500mm; 1.500.000 ab.]

- Come valutare le feature?
 - ◆ Vicinanza ai bordi di ciascun cluster (Support Vector Machines)
 - ◆ Analisi statistica del potere discriminante: correlazione tra feature e loro significatività.



Feature & feature



- Feature extraction: trasformazione delle feature per creare nuove, significative feature;
- Elaborazione di primo livello, per ottenere informazioni caratteristiche del fenomeno che, ad esempio, siano invariante.

Esempio: descrizione di oggetti circolari.

Posso misurare l'area e il perimetro, ma il loro rapporto è più significativo.

Esempio: descrizione del clima.

Milano: [13°; 900mm; 265 giorni sole; 100 giorni pioggia]

oppure

Milano: [13°; 900mm / 100 giorni pioggia; 265 giorni sole]



Prossimità di pattern eterogenei



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

esempio:

Distanza euclidea...

$\text{dist}(\text{Roma}, \text{Milano}) = \dots$

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

$= \dots$ Distanza euclidea? $= \dots$

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

Ha senso?



Normalizzazione feature



Att.ne!

$\text{dist}(\text{Roma}, \text{Milano}) = \dots$

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

$= \dots$ Distanza euclidea? $= \dots$

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

La distanza tra le due città in termini di gradi è insignificante nel nostro calcolo... **E' necessario trovare una metrica corretta per la rappresentazione dei dati. Ad esempio, normalizzare i dati!**

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

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

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

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

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



Altre funzioni di distanza



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



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: fuzzy c-mean, neural-gas, enhanced vector quantization, **mappe di Kohonen**.
- Algoritmi statistici: fitting di **mixture models**.



Riassunto



- I tipi di apprendimento
- Il clustering
- **K means**
- Quad-tree decomposition



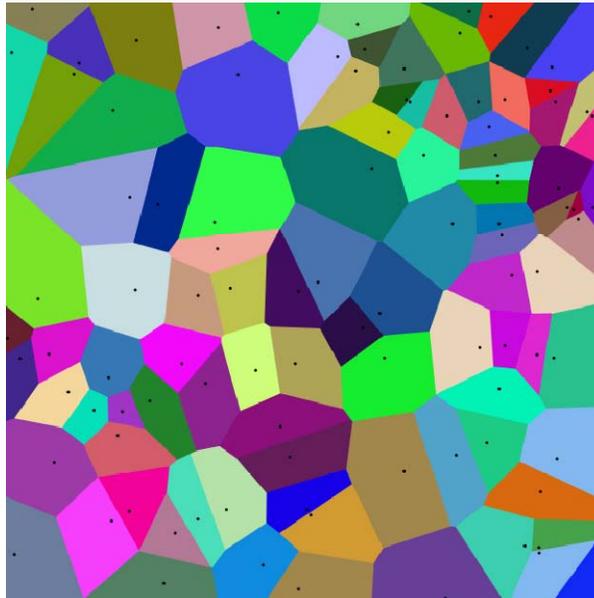
K-means (partitional): framework



- Siano $\mathbf{X}_1, \dots, \mathbf{X}_D$ i dati di addestramento (per semplicità, definiti in \mathbb{R}^2);
- Siano $\mathbf{C}_1, \dots, \mathbf{C}_K$ i *prototipi* di K classi, definiti anch'essi in \mathbb{R}^2 ; ogni *prototipo* identifica il baricentro della classe corrispondente;
- Lo schema di classificazione adottato sia il seguente: “ \mathbf{X}_i appartiene a \mathbf{C}_j se e solo se \mathbf{C}_j è il *prototipo* più vicino a \mathbf{X}_i (distanza euclidea)”;
- L'algoritmo di addestramento permette di determinare le posizioni dei *prototipi* \mathbf{C}_j mediante successive approssimazioni.



Risultato del clustering è un diagramma di Voronoi



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



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

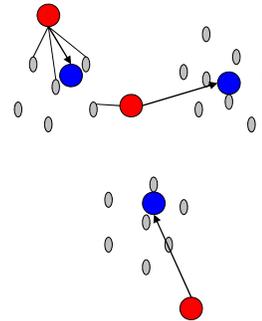
Aggiornamento C_j

I prototipi C_j si sono spostati significativamente?

NO

Termine addestramento

SI



Aggiornamento C_j : baricentro degli X_i classificati da C_j .

A.A. 2009-2010

35/50



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:

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

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

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

- La funzione obiettivo viene definita come:
$$\sum_{i=1}^K \sum_{j=1}^N (x_j - C_k)^2$$

A.A. 2009-2010

36/50

<http://homes.dsi.unimi.it/~borgnese/>



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.



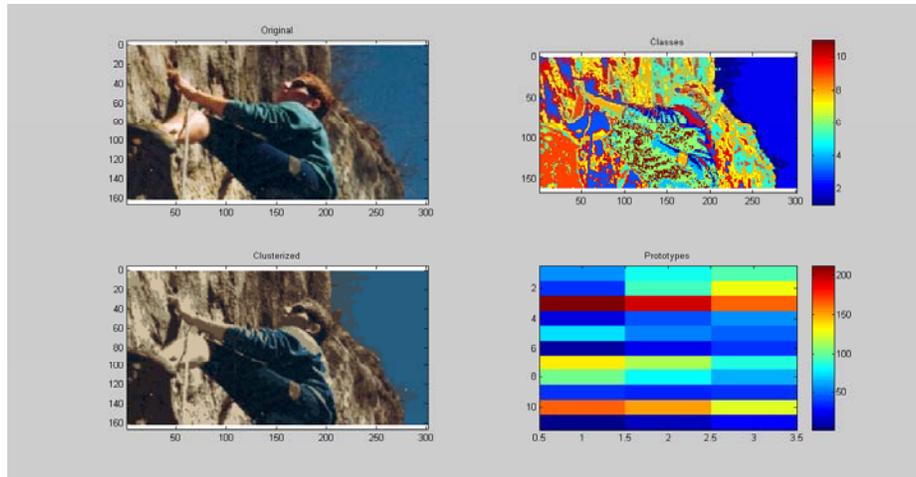
K-means::limiti



- Partitional, hard, deterministic;
- Veloce, semplice da implementare;
- Trova un minimo locale della funzione $f = \sum_j \sum_i [\text{dist}(x_i, \text{prot}_j)] / N_j$;
- Il risultato dipende dall'inizializzazione!
- Possono essere usati altri metodi (es. GA) per inizializzare K-means... es. GA per la minimizzazione di f , effettuano una ricerca globale, ma sono lenti!



K-Means per immagine RGB



Da 255 colori a 33 colori

A.A. 2009-2010

39/50



Riassunto

- I tipi di apprendimento
- Il clustering
- K means
- **Quad-tree decomposition**

A.A. 2009-2010

40/50

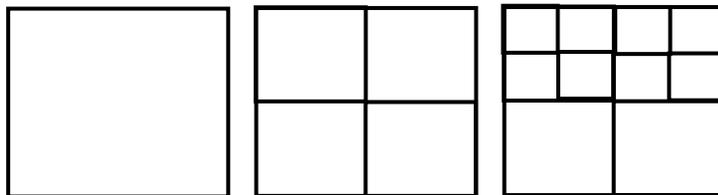
<http://homes.dsi.unimi.it/~borghese/>



Algoritmi gerarchici: QTD



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



A.A. 2009-2010

41/50



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. 2009-2010

42/50



Algoritmi gerarchici: QTD



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

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

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

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

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

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

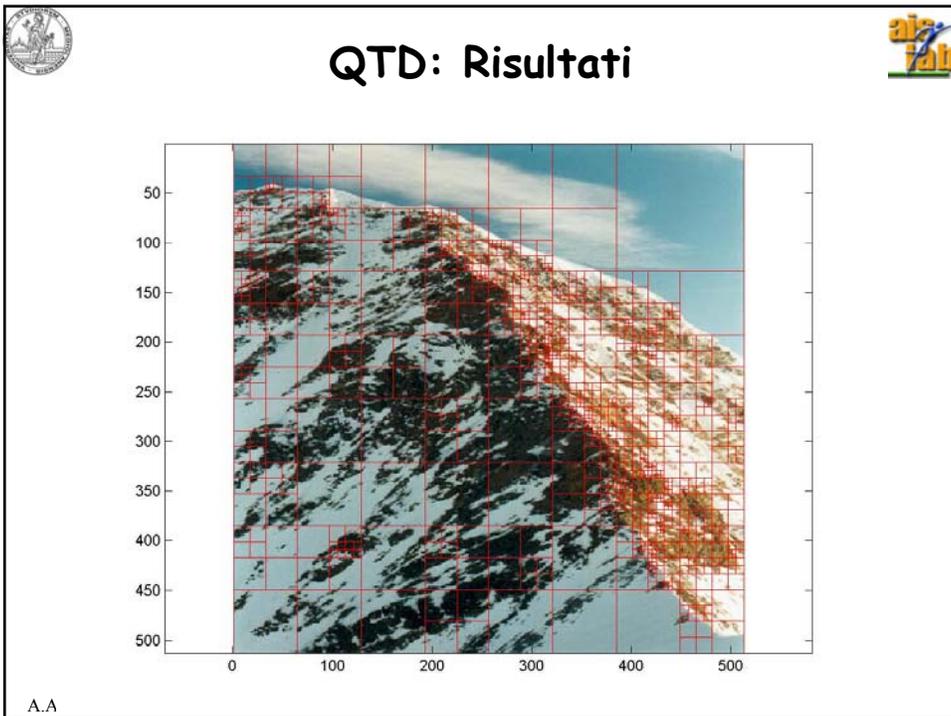
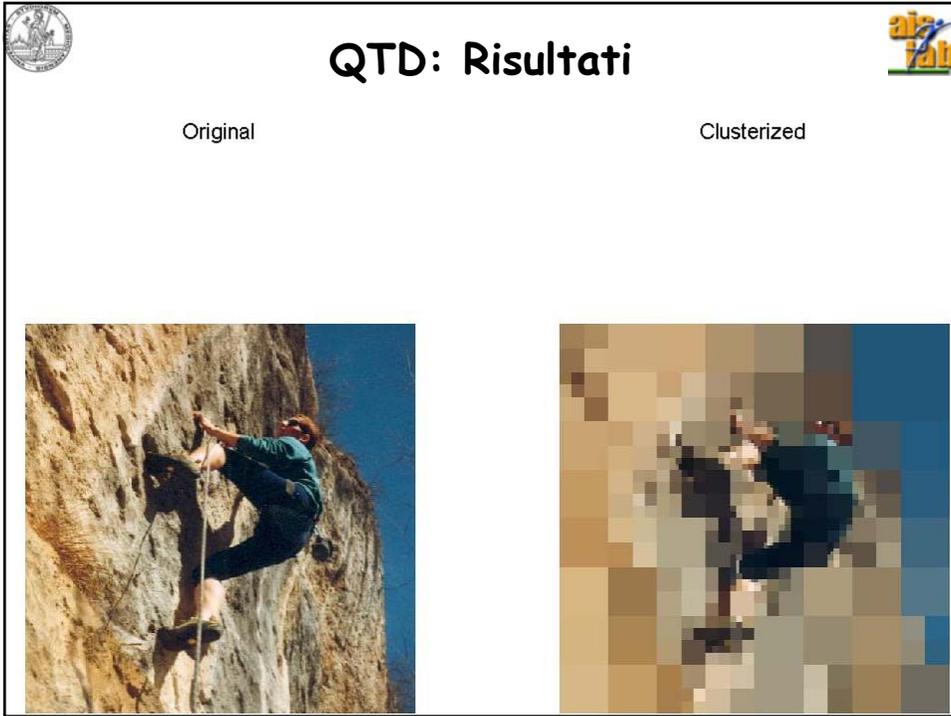


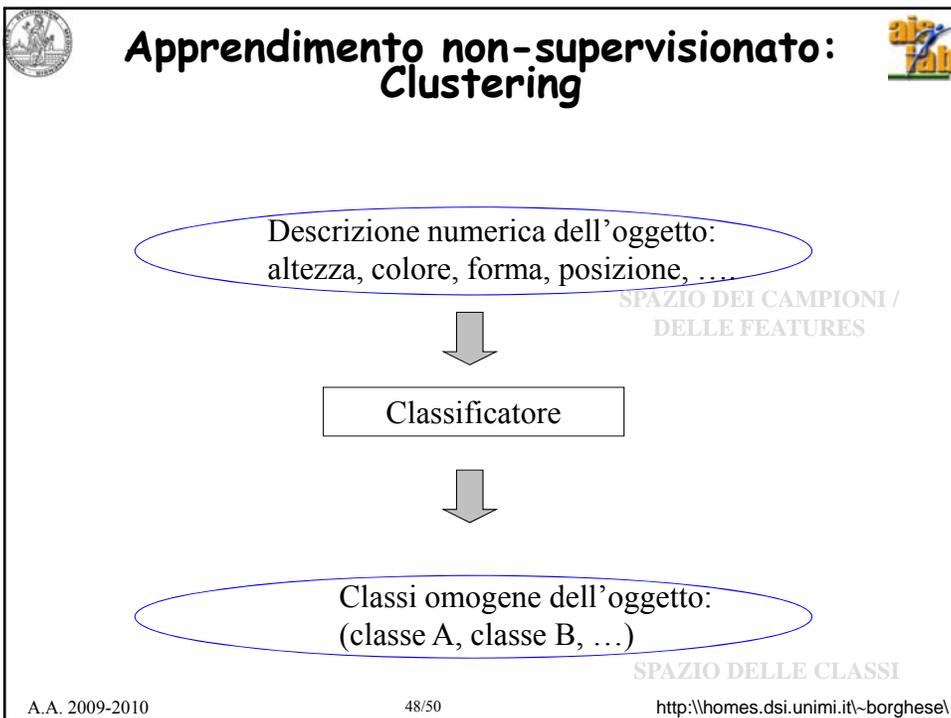
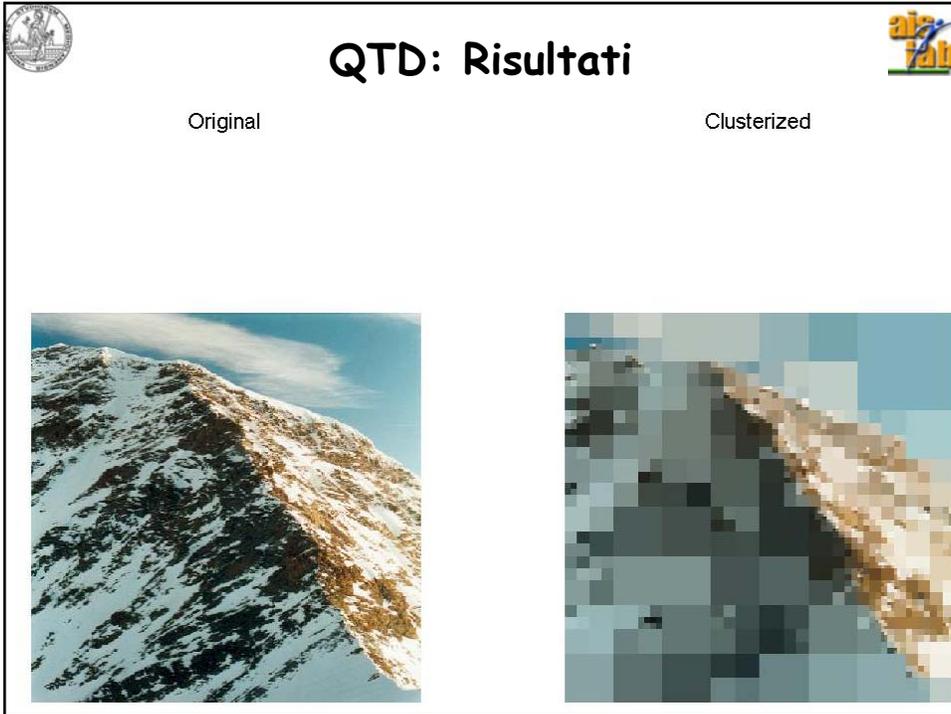
Algoritmi gerarchici: QTD



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...



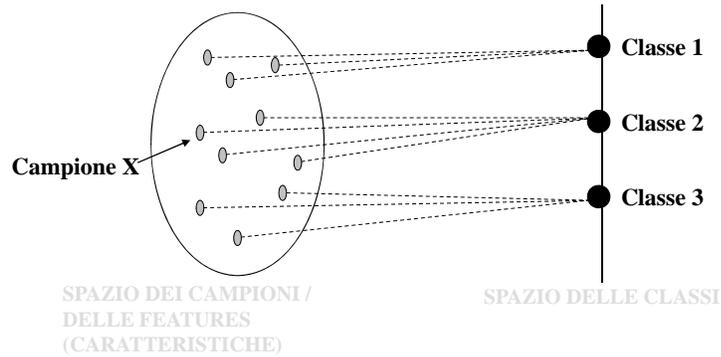




Classificazione e clustering

Un'interpretazione geometrica:

Mappatura dello spazio dei campioni nello spazio delle classi.



*Che differenza c'è rispetto al clustering?
Cos'è un concetto?*

A.A. 2009-2010

49/50

<http://homes.dsi.unimi.it/~borghese/>



Riassunto

- I tipi di apprendimento
- Il clustering
- K means
- Quad-tree decomposition

A.A. 2009-2010

50/50

<http://homes.dsi.unimi.it/~borghese/>