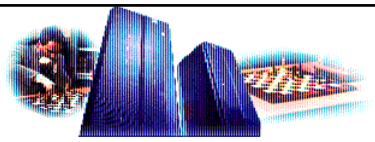
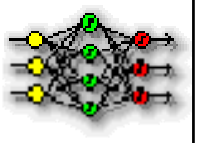




Sistemi Intelligenti
Learning and Clustering

Alberto Borghese and Iuri Frosio

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

Riassunto

- I tipi di apprendimento
- Il clustering

I vari tipi di apprendimento

$x(t+1) = f[x(t), a(t)]$ Ambiente
$a(t) = \mathbf{g}[x(t)]$ Agente

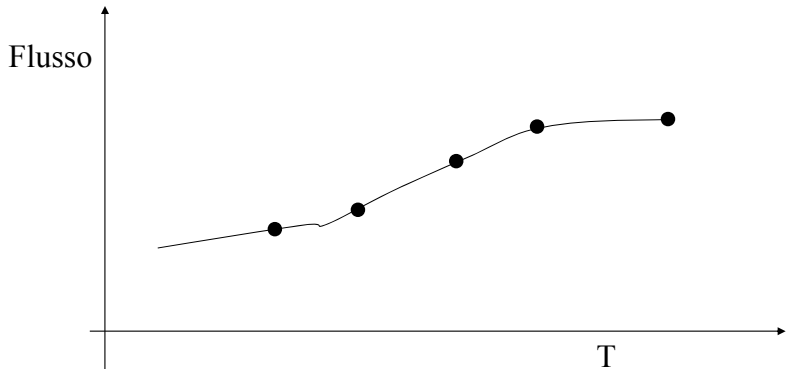
Supervisionato (learning with a teacher). Viene specificato per ogni pattern di input, il pattern desiderato in output.

> **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.

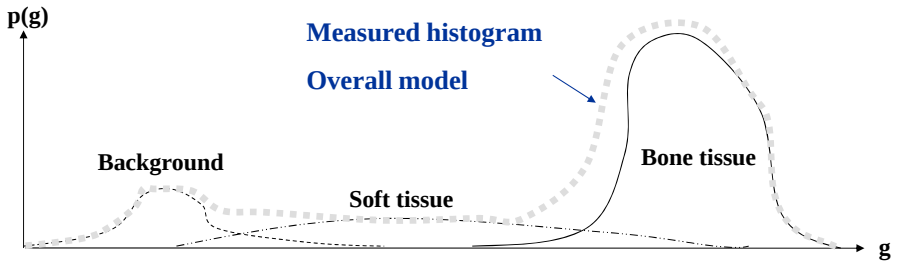
Apprendimento con rinforzo (reinforcement learning, learning with a critic). L'ambiente fornisce un'informazione puntuale, di tipo qualitativo, ad esempio success or fail.

Apprendimento supervisionato: regressione = predictive learning



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

I modelli parametrici



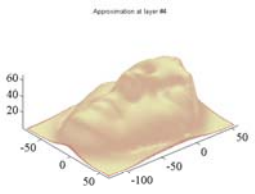
$$p(g)=\sum_{j=1}^{M} P(j)\cdot p(g\mid j)=\sum_{j=1}^{M} w_j \cdot p_j(g)$$

La probabilità di avere un livello di grigio g è la somma pesata delle tre probabilità di avere background, $p_1(g)$, tessuto molle, $p_2(g)$ o tessuto osseo, $p_3(g)$.

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

Da calcolare

Funzione di base (fissate)

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.

$$z(p(x,y)) = f(x) = \sum_i w_i x$$

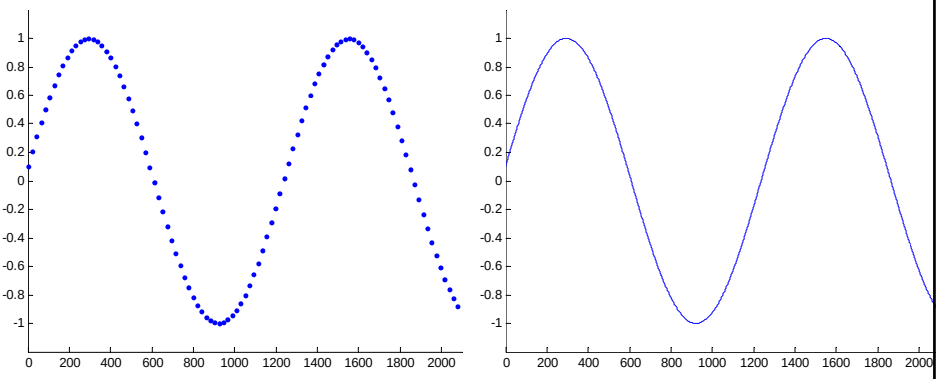
f(.) $= w_i$ è funzione lineare

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

f(.) è funzione non lineare

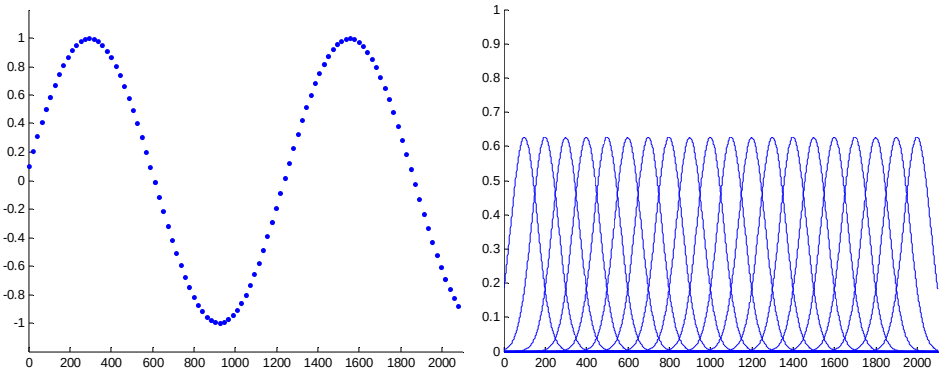
e.g. $f(.) = e^{G(.)}$

Funzionamento di un modello parametrico (non-lineare)



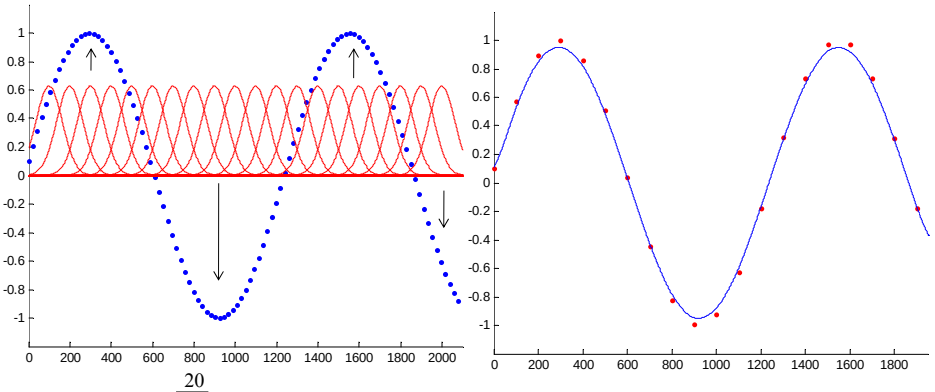
I punti vengono fittati perfettamente da una sinusoide: y = **A** sin(**ω**x + **ϕ**). Devo determinare i parametri della sinusoide (non lineare), i cui valori ottimali sono: ω = 1/200, ϕ = 0.1.

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?

Funzionamento di un modello semi-parametrico (lineare)

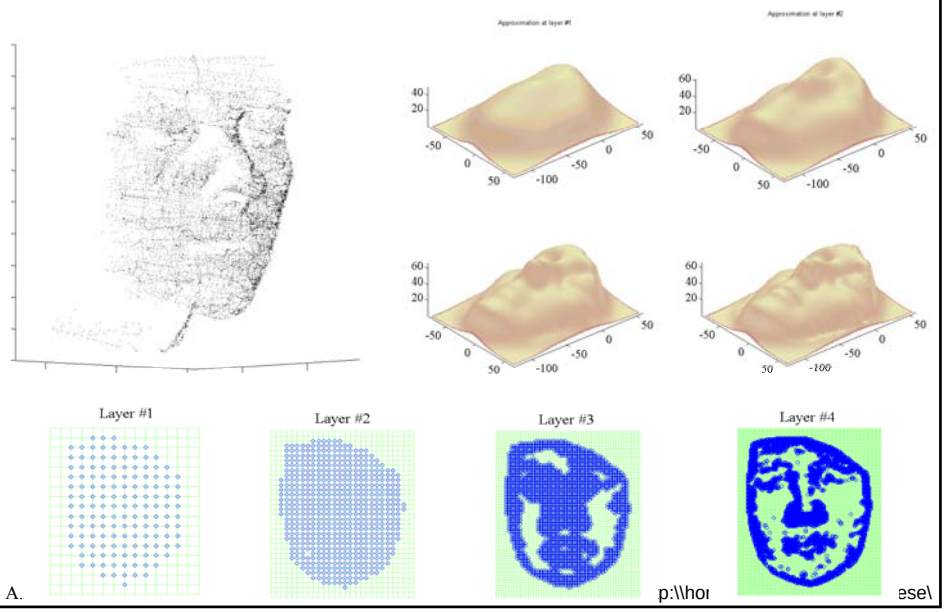


$$y(x) = \sum_{i=1}^{20} w_i G(x - x_{o_i}; 90)$$

Devo definire, i $\{w_i\}$

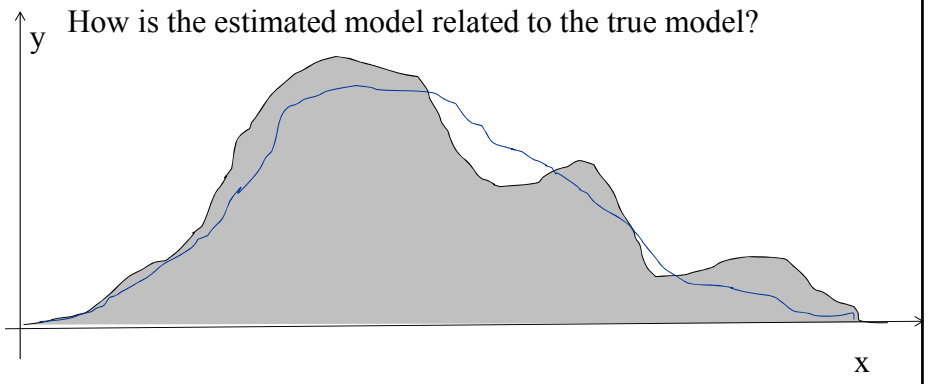
I σ sono tutti uguali ed uguali a 90, le Gaussiane sono equispaziate.
Le Gaussiane sono note tutte a priori, devono essere definiti i pesi.

Applicazione della regressione



How to classify the error introduced by a model?

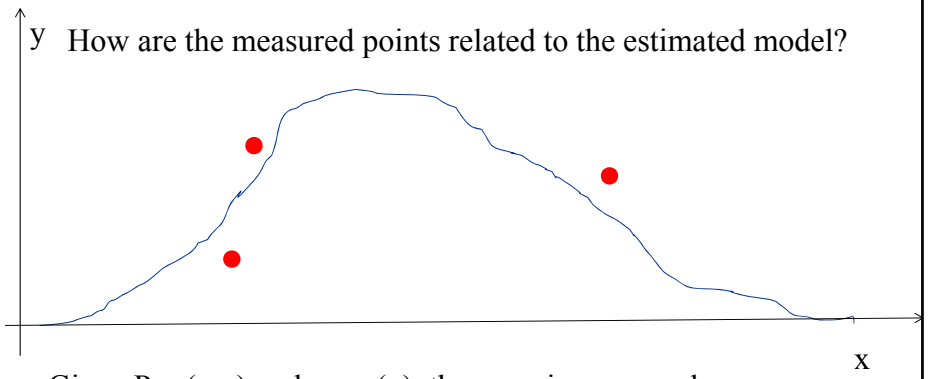
How is the estimated model related to the true model?



Bias and variability trade-off

Bias is the distance of the model curve from the true unknown curve. It is associated to model error.

Variability



How are the measured points related to the estimated model?

Given $P_{mes}(x,y)$ and $y = y(x)$, the error is measured as:

- $\min \| P - y(x) \|$
- $\| y_{mes}(x_{mes}) - y(x_{mes}) \|$

It is associated to measurement error.

If variability goes to zero, bias increases and overfitting arises.

Problemi

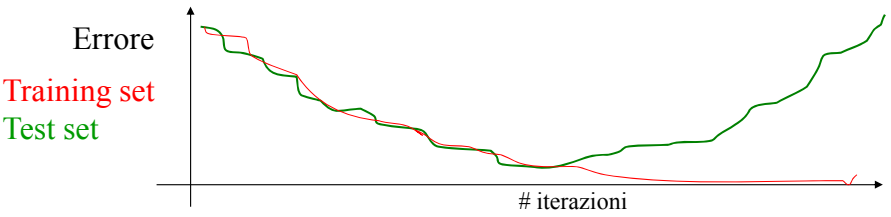
Quando si termina l'algoritmo di apprendimento?

Bootstrap – Vengono estratti pattern con ripetizioni.

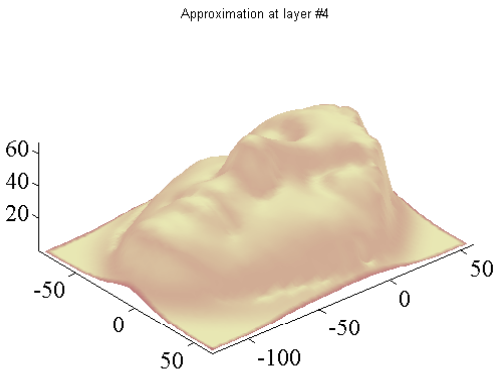
Cross-Validation - Errore sull'insieme di training = Errore sull'insieme di test.

Utilizzare lo "structural risk" invece dell'"empirical risk".

Si vuole evitare che la rete si specializzi troppo sui pattern di training e non sia in grado di interpolare.



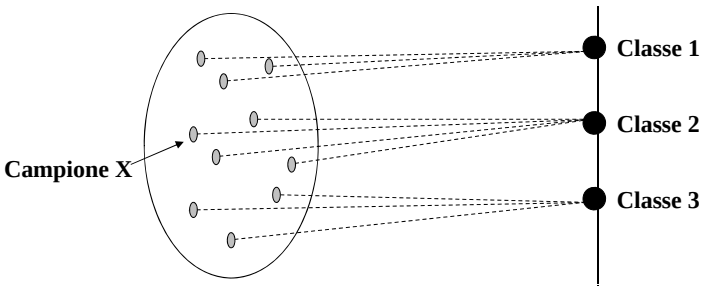
Problema dell'overfitting dovuto a sovraparametrizzazione



Quante unità?

Classificazione

Un'interpretazione geometrica:
Mappatura dello spazio dei campioni nello spazio delle classi.



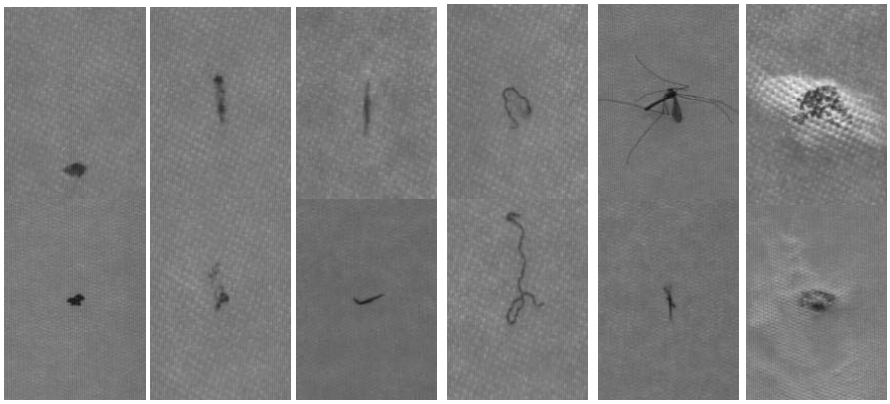
SPAZIO DEI CAMPIONI / DELLE FEATURES (CARATTERISTICHE)

SPAZIO DELLE CLASSI

Apprendimento Supervisionato: Classificazione

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

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



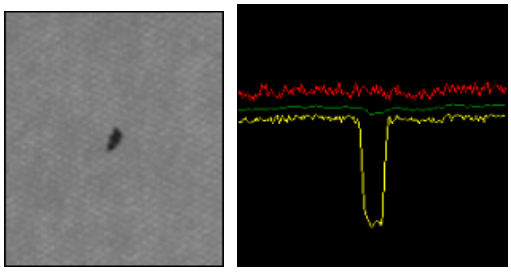
regolari | irregolari | allungati | fili | insetti | macchie su denso

Difetti – Classificazione real-time e apprendimento mediante **booosting.**
Committee (linear combinaion) of weak (binary) classifiers.

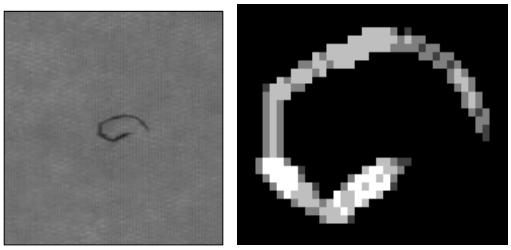
Features

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

Macchie dense



Fili

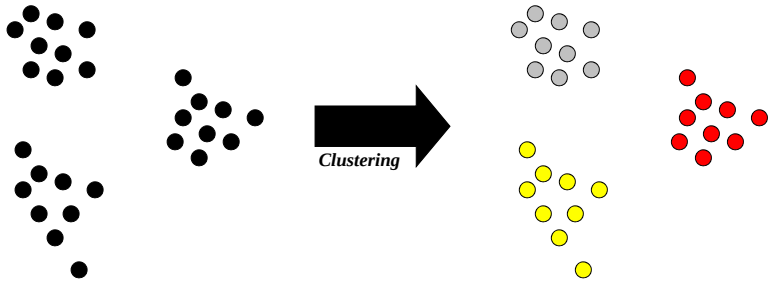


Riassunto

- I tipi di apprendimento
- Il clustering

Clustering

- Clustering: raggruppamento degli "oggetti" in classi omogenee tra loro.
  - Raggruppamento per colore
  - Raggruppamento per forme
  - Raggruppamento per tipi
  - …..



Novel name: **data mining**

Esempio di clustering

Clustering -> Indicizzazione

Ricerca immagini su WEB.

Il clustering per…

… Confermare ipotesi sui dati (es. “E’ possibile identificare tre diversi tipi di clima in Italia: mediterraneo, continentale, alpino…”);

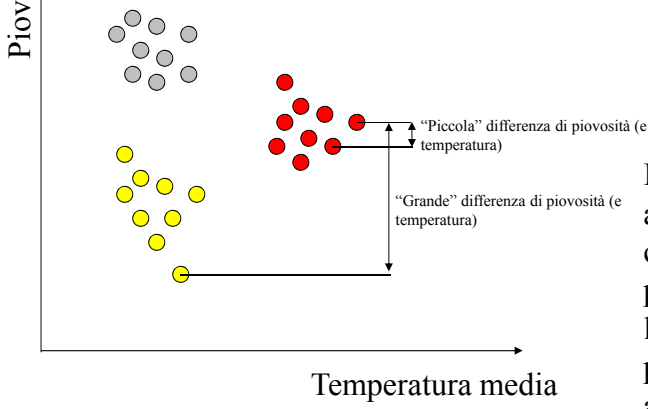
… Esplorare lo spazio dei dati (es. “Quanti tipi diversi di clima sono presenti in Italia?”);

… 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

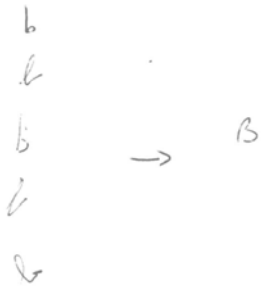
Processo attraverso il quale i dati (pattern, vettori) vengono organizzati in cluster, basata sulla similarità. L’elaborazione verrà poi effettuata sui prototipi che rappresentano ciascun cluster.



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

Clustering: definizioni

- **Pattern**: un singolo dato **p** = [p_1, p_2, … p_n]. Il dato appartiene quindi ad uno spazio multi-dimensionale, solitamente eterogeneo.
- **Feature**: le caratteristiche dei dati significative per il clustering, possono costituire anch'esso un vettore, il vettore delle feature: x_1, x_2, … x_d. Questo vettore costituisce l'input agli algoritmi di clustering.

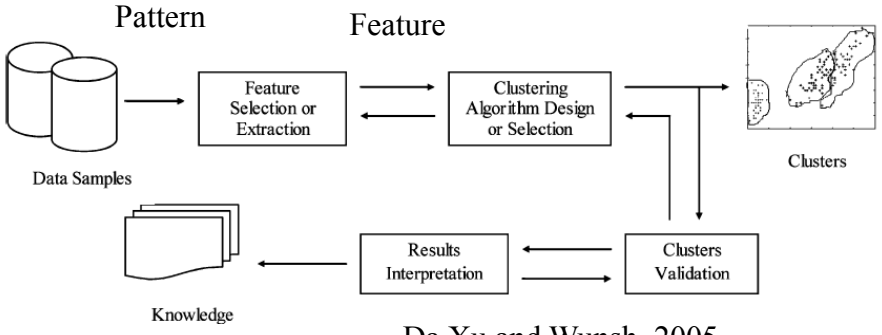


Inclinazione, occhielli, lunghezza, linee orizzontali, archi di cerchio ...

Clustering: definizioni

- **n**: dimensione dello spazio dei pattern;
- **d**: dimnesione 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).

Analisi mediante clustering



Da Xu and Wunsh, 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 (gerachico 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).

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 scegliere i pattern?
  - Vicinanza ai bordi di ciascun cluster (Support Vector Machines)
  - Tutti i pattern

- Come valutare le feature?
  - 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 inverianti.

  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]

Similarità tra feature

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

  Esempio:
  Distanza euclidea…
  dist (Roma, Milano) = …
  dist ([17°; 500mm], [13°; 900mm]) = …
  = … Distanza euclidea? = …
  = $((17-13)^2+(500-900)^2)^{½} = 400.02 \sim 400$

Ha senso?

Normalizzazione feature

Att.ne!
dist (Roma, Milano) = …
dist ([17°; 500mm], [13°; 900mm]) = …
= … Distanza euclidea? = …
= $((17-13)^2+(500-900)^2)^{½} = 400.02 \sim 400$

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 le feature**!

T_{Max}= 20° T_{Min}= 5° → $T_{Norm} = (T - T_{Min})/(T_{Max} - T_{Min})$
P_{Max}= 1000mm P_{Min}= 0mm → $P_{Norm} = (P - P_{Min})/(P_{Max} - P_{Min})$
$Roma_{Norm}$ = [0.8 0.5]
$Milano_{Norm}$ = [0.53 0.9]
$dist(Roma_{Norm}, Milano_{Norm}) = ((0.8-0.53)^2+(0.5-0.9)^2)^{½} = 0.4826$

Altre funzioni di distanza

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

Tassonomia (sintetica) degli algoritmi di clustering

- Algoritmi gerarchici (agglomerativi, divisivi), e.g. **Hirarchical 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: **mixture models**.

Riassunto

- I tipi di apprendimento
- Il clustering