Intelligenza Artificiale

Reti Neurali

Roberto Marmo
Laboratorio di Visione Artificiale, Pavia
marmo@vision.unipv.it

Obiettivi

Apprendere un paradigma di calcolo che deriva dal cervello umano e capire quali problemi può risolvere

- cosa è una rete neurale
- come si differenzia da altri paradigmi di calcolo
- come si costruisce
- quando usarla

Sommario

Parte 1. Il cervello umano

Parte 2. La rete neurale

Parte 3. L'apprendimento della rete

Parte 4. Creare una rete per classificare

Parte 5. Esempi di applicazione e commento di articolo scientifico in lingua inglese.

Parte 1

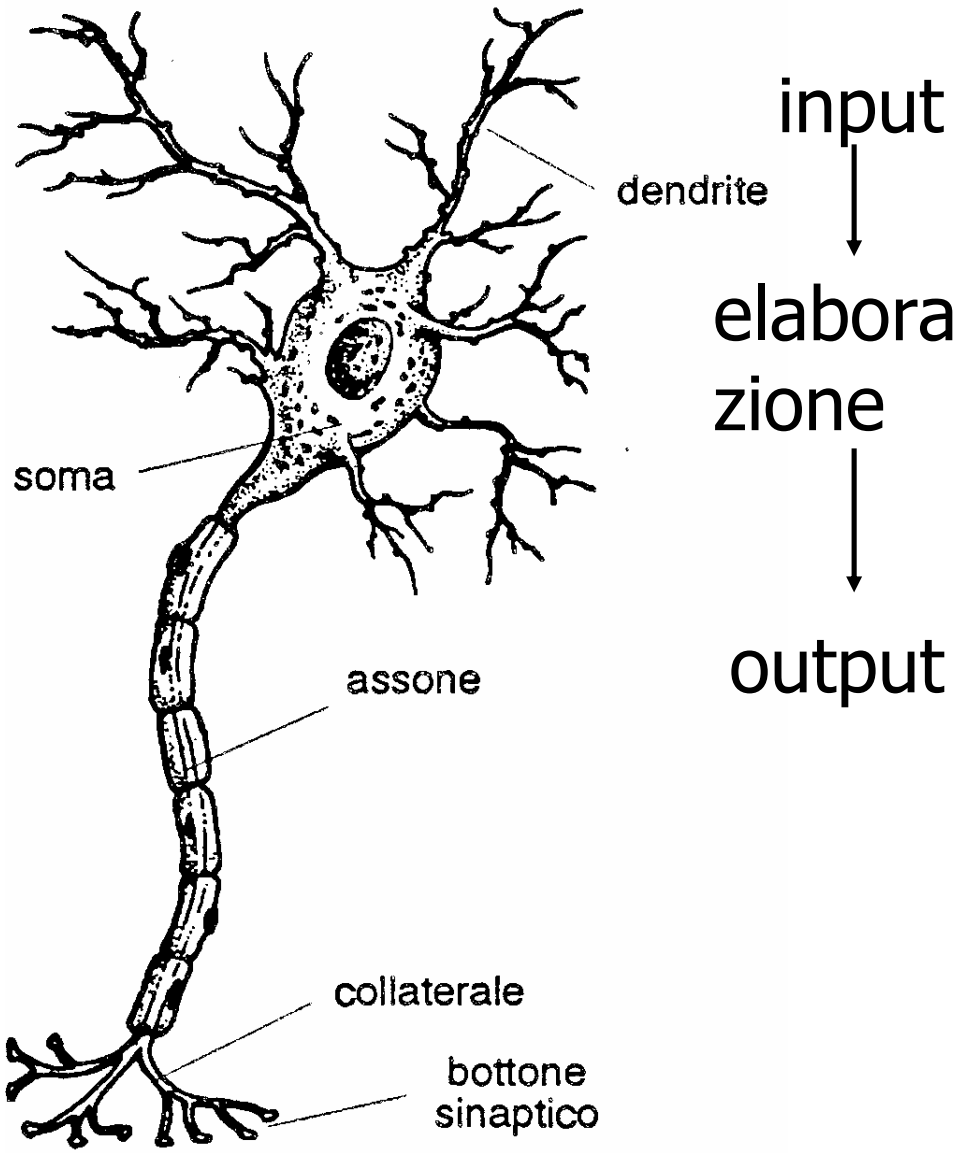
Il cervello umano

John von Neuman: non c'è alcun modello del cervello più semplice del cervello stesso.

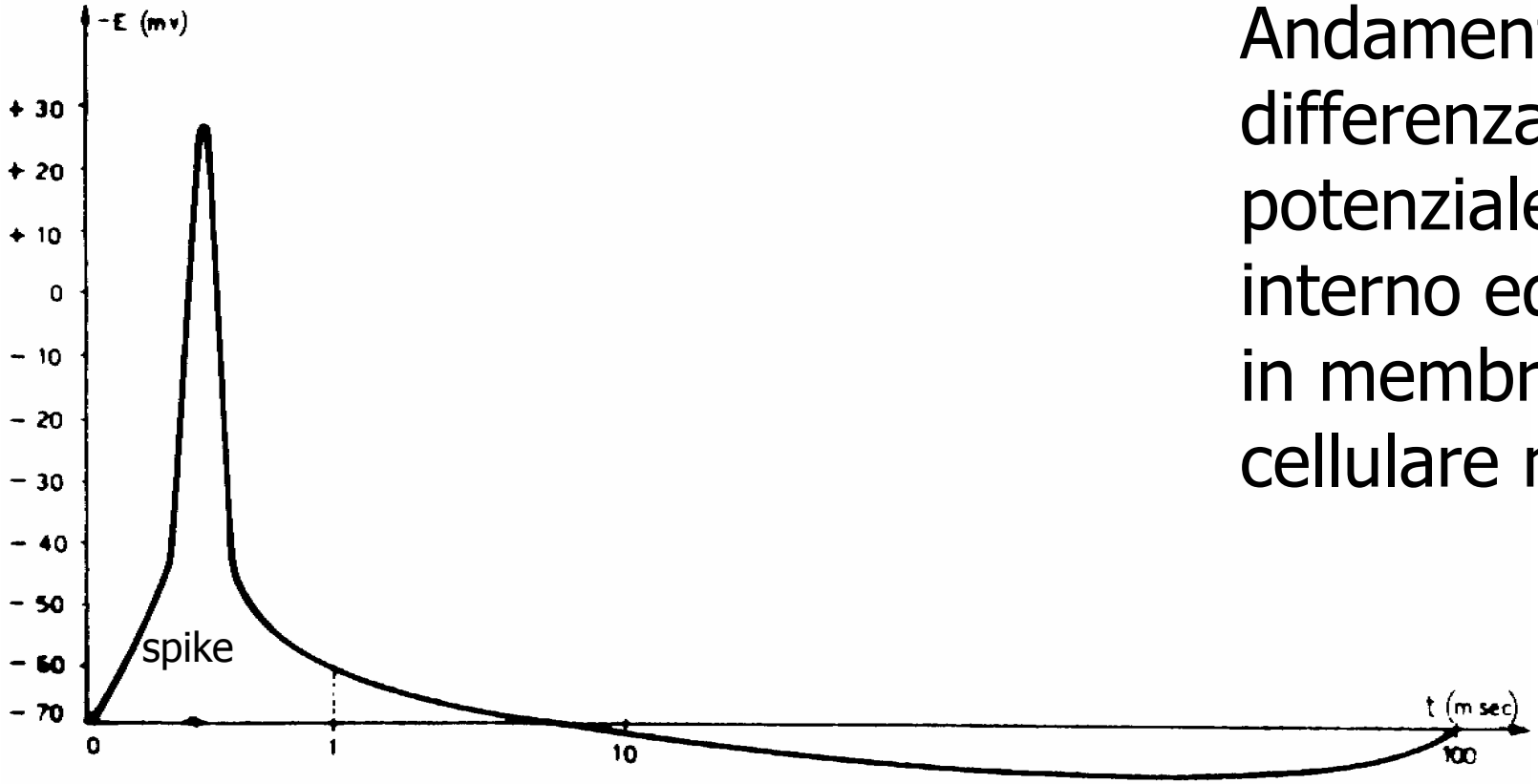
Per simulare alcuni comportamenti della struttura cerebrale degli esseri viventi attraverso un insieme di regole di calcolo, occorre conoscere la struttura del cervello umano e cercare di riprodurlo con un modello matematico.

La cellula neuronale

Unità fondamentale del cervello umano. Ogni neurone riceve come input i segnali elettrici da tutti i dendriti, e se la somma pesata supera il valore di attivazione emette un impulso elettrico in uscita verso l'assone.



Funzione di sparo del neurone

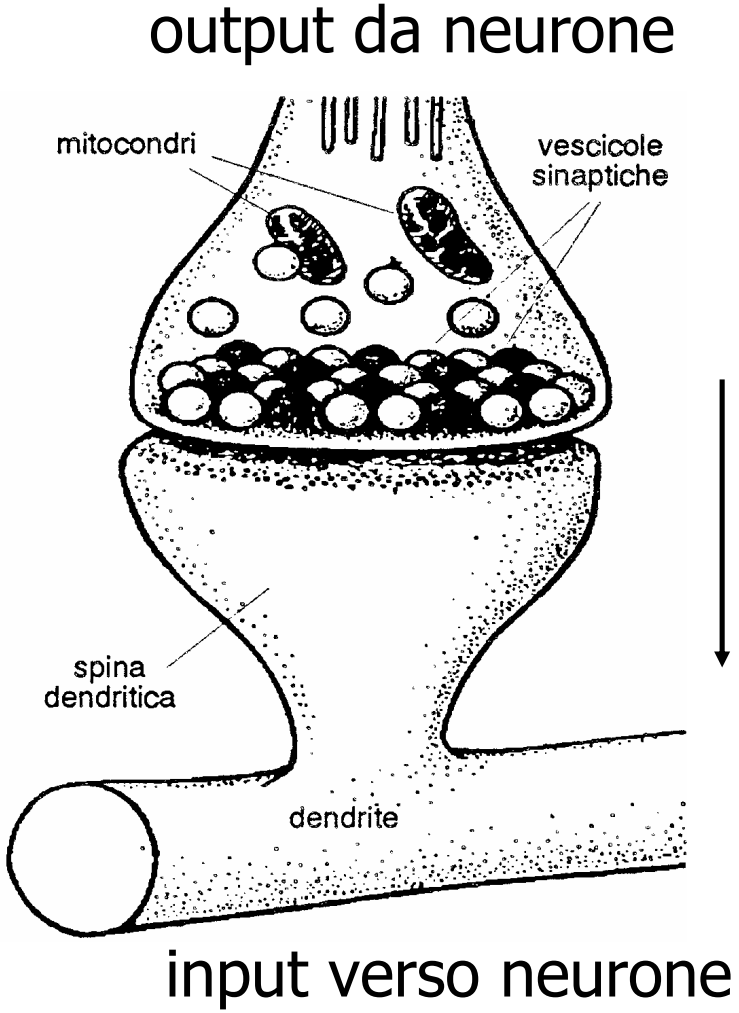


Andamento di –E, differenza di potenziale elettrico interno ed esterno in membrana cellulare neurone

Dopo stimolo di adeguata intensità, il neurone risponde con spike (sparo) o no: non c'è risposta intermedia.

Contatto sinaptico

Punto di contatto tra due neuroni. Le sinapsi aumentano o diminuiscono nel tempo. La parte superiore trasforma il segnale elettrico in sostanza chimica (neurotrasmettitore) che passa dall'altra parte e viene riconvertito in segnale elettrico. La sinapsi ha azione eccitatoria o inibitoria del collegamento variando i neurotrasmettitori.



Il cervello umano

- nome scientifico: encefalo
- esistono vari tipi di neuroni
- 10 miliardi di neuroni, 80000 neuroni per mm^2
- tessuto gelatinoso con peso di 1300-1500 grammi
- ogni neurone interagisce con 1000-10000 neuroni
- l'impulso elettrico viaggia alla velocità di 130 metri/sec
- sinapsi tra neuroni anche molto distanti

Il cervello umano

- importa il numero di sinapsi non di neuroni
- elaborazione d'informazione in parallelo tra migliaia di neuroni porta all'emergere dei processi cognitivi
- elaborazione in due emisferi separati da un corpo calloso che sono poi suddivisi in regioni localizzate e con compiti specifici
- struttura cerebrale sempre in evoluzione: ogni giorno centinaia di neuroni muoiono, ma non c'è declino mentale perché aumentano le sinapsi per compensare le perdite

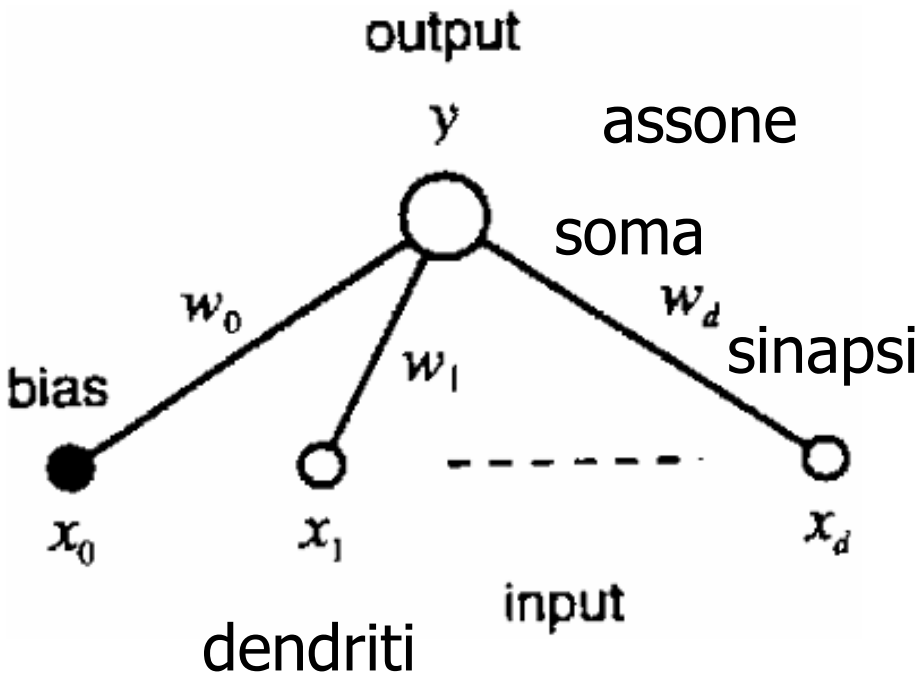
Parte 2

La rete neurale

Il neurone artificiale

Modello matematico molto semplificato del neurone biologico.

Ad ogni input x_i è associato un peso w_i con valore positivo o negativo per eccitare o inibire il neurone. Il bias varia secondo la propensione del neurone ad attivarsi, per variare la soglia di attivazione del neurone.



Algoritmo del neurone

1. caricare: input x_i e pesi w_i i=1..d, bias x_0 e w_0
2. calcolare la somma dei valori input pesata con i relativi pesi
3. calcolare il valore della funzione di attivazione g con il risultato della somma pesata
4. l'output del neurone y è il risultato della funzione di attivazione

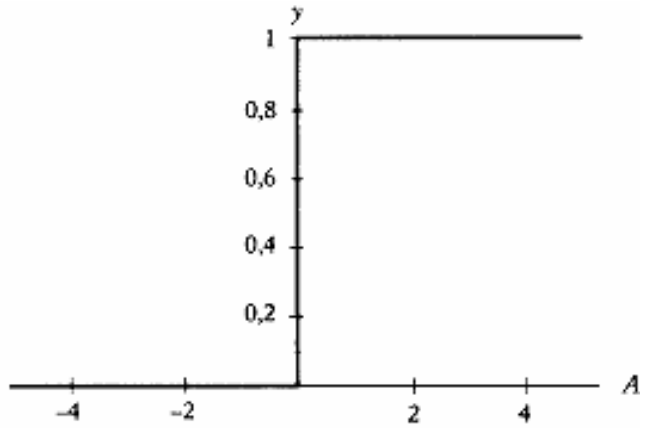
$$y(x) = g\left(\sum_{i=1}^{d} w_i x_i + w_0 x_0\right) = \overline{w}^T \overline{x}$$

Funzioni di attivazione

Determina la risposta del neurone.

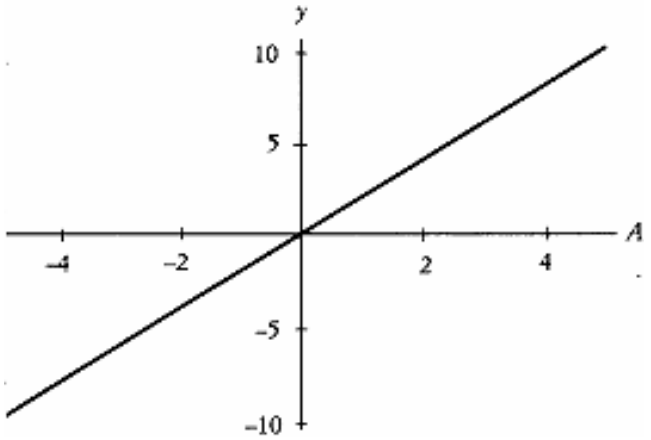
A gradino

$$g(A) = \begin{cases} 1 \ se \ A > 0 \\ 0 \ altrimenti \end{cases}$$



Lineare continua

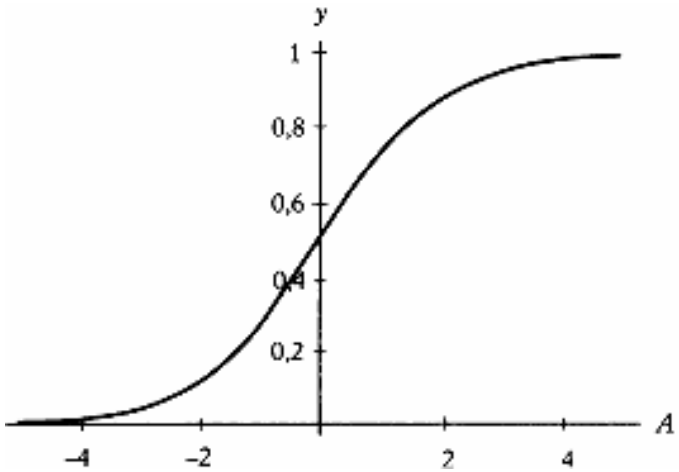
$$g(A) = kA$$



Funzioni di attivazione

La sigmoide o logistica ha valori positivi, è continua e derivabile

$$g(A) = \frac{1}{1+e^{-A}}$$



La softmax crea ogni output in [0,1] e somma di tutti gli output pari a 1, per interpretare la risposta della rete come stime di probabilità

$$g(a_i) = \frac{e^{a_i}}{\sum_{i=1}^{n} e^{a_i}}$$

La rete neurale

Sistema dinamico avente la topologia di un grafo orientato con nodi, i neuroni artificiali, ed archi, i pesi sinaptici. Il termine rete è riferito alla topologia dei collegamenti tra i neuroni.

La rete neurale è un modello matematico che usa neuroni matematici, la rete neuronale usa neuroni biologici.

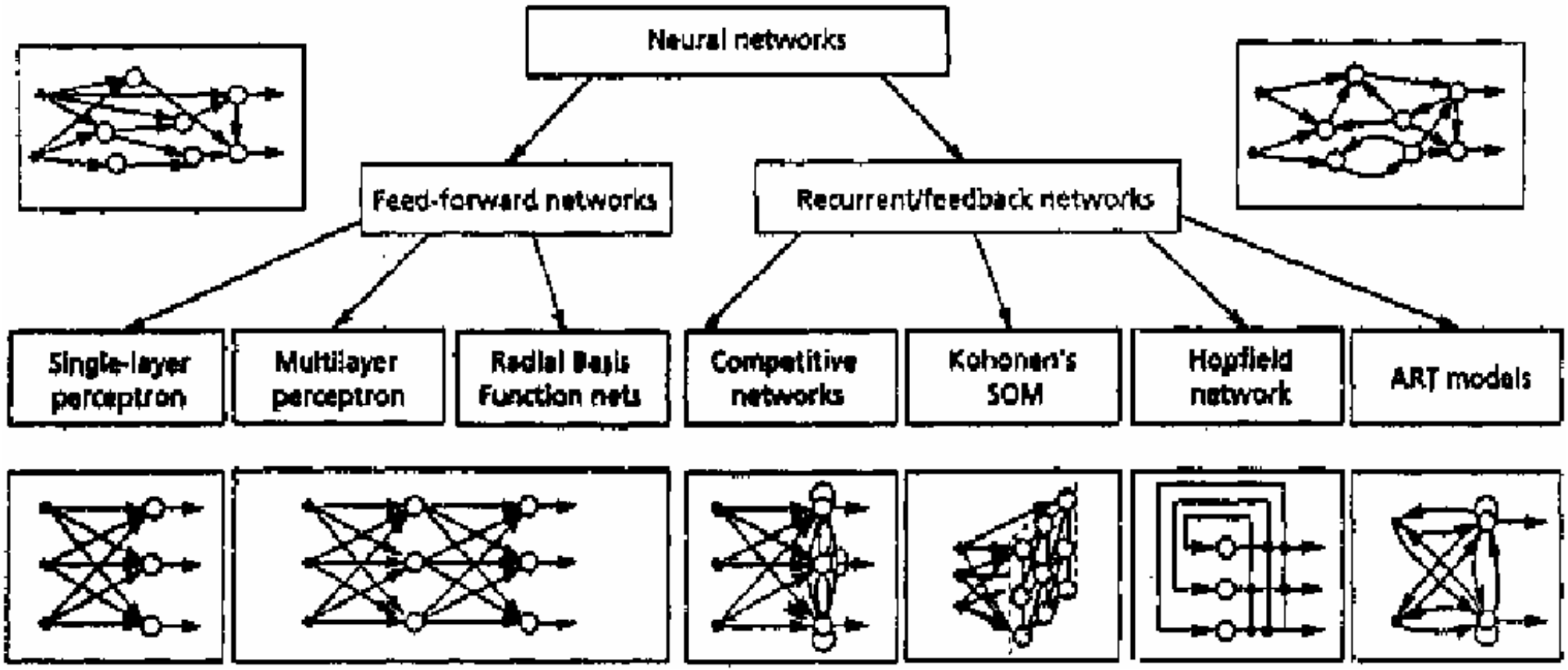
Viene studiata con un approccio statistico.

La rete neurale

Altre definizioni:

- scatola nera, cioè si può ignorare il funzionamento, che associa un input a un output e le associazioni possono essere create con l'apprendimento
- modello matematico che calcola la funzione output=f(input,pesi) al variare dei pesi e senza specificare la forma della funzione f
- algoritmo non lineare per ottenere una soluzione approssimata a problemi di cui non esiste un modello, tramite l'uso di esempi di calcolo

Esempi di reti neurali



Vantaggi

- adatte per problemi che non chiedono risposte accurate, ma risposte approssimate con un grado di errore o di variazione
- risolvono problemi complicati in cui non è facile descrivere la soluzione, esempio riconoscimento di facce o caratteri scritti a mano
- facili da implementare, basta definire il neurone e poi crearne delle copie e creare i collegamenti tra i neuroni

Vantaggi

- funzionamento veloce perché parallelo; ogni neurone usa solo il suo input
- stabilità dell'output rispetto a valori di input: incompleti, con rumore, non ben noti, che accettano un grado di errore o di variazione
- determinano il risultato tenendo conto contemporaneamente di tutti gli input

Svantaggi

- incapacità di rendere conto dell'elaborazione: non si può capire perché ha dato quel risultato specifico in quanto non si può descrivere e localizzare la conoscenza che viene memorizzata su tutta la rete
- carenza di hardware con cui implementare, si usano su computer classici
- tecniche di addestramento sofisticate che richiedono molto tempo di calcolo

Svantaggi

- non sempre esiste una rete che risolve il problema, perché non sempre esiste un algoritmo di apprendimento che converge dando un output della rete con basso errore
- i valori di output non sono precisi, ma hanno un margine in cui possono variare
- serve una casistica di esempi molto ampia per ottenere un buon apprendimento e un basso errore di output

Campi di applicazione

1. pattern classification e clustering
2. approssimazione di funzioni
3. predizioni in serie temporali
4. ottimizzazione
5. memorie associative
6. controllo di apparati
7. elaborazione di segnali ed immagini
8. internet e sicurezza computer

Rapporti con algoritmi genetici

Gli algoritmi genetici possono essere applicati alle reti neurali per risolvere problemi come:

- scegliere la struttura della rete
- scegliere i valori dei pesi per ridurre l'errore di output

Rapporti con logica fuzzy

Le reti neurali non calcolano con le variabili linguistiche e non usano regole qualitative. Esistono le reti neuro-fuzzy, strumento matematico molto efficiente e complesso che unisce i vantaggi delle due tecniche.

Rapporti con i sistemi esperti

Le reti neurali si differenziano dai sistemi esperti:

- non usano conoscenze esplicite ma implicite contenute in un ampio insieme di esempi formato da vettori di numeri
- usano numeri invece di simboli, regole, IF…THEN..
- tutti i neuroni collaborano, se si toglie un neurone spesso i risultati non cambiano molto mentre se si toglie una regola i risultati possono essere molto diversi
- non vengono programmate, vengono addestrate con l'apprendimento da esempi

Rapporti con i sistemi esperti

- un sistema esperto può fare ragionamenti come "Socrate è un uomo, gli uomini sono mortali, Socrate è mortale", la rete neurale no ma può facilmente riconoscere una faccia in una immagine

- un sistema esperto non tiene conto del supporto materiale che ragiona, si dedica alla funzione del ragionamento e cerca di imitare la mente umana. Invece il connessionismo è l'approccio allo studio dell'intelligenza con reti neurali; considera fondamentale la struttura del supporto materiale che ragiona, cerca di imitare cervello umano e sistema nervoso

Rapporti con i sistemi esperti

- la rete neurale accetta dati di input parziali e con rumore senza modificare troppo l'output

- il sistema esperto può spiegare perché e come ha ottenuto una conclusione, può indicare quali regole ha usato; la rete neurale non può giustificare perché e come ha ottenuto dei valori specifici in uscita, è difficile capire quale neurone o peso è importante per avere l'output

Parte 3

L'apprendimento della rete

La programmazione

Esempio: per riconoscere un computer bisogna scrivere un programma con struttura IF…THEN che include tutti i casi possibili:

IF ha microprocessore THEN

IF ha memoria THEN

IF ha tastiera THEN

……

PRINT “è un computer”

Apprendimento da esempi

Apprendere significa migliorare la capacità di esecuzione di un certo compito attraverso l'esperienza.

Esempio: all'inizio la rete neurale non conosce il concetto di computer. Alla rete neurale vengono inseriti come input tanti esempi di immagini di computer con associata come output l'etichetta "computer". La rete neurale impara e si crea una esperienza sul riconoscimento dei computer; la prossima volta che ha in input una immagine di un computer, diversa da quella negli esempi di apprendimento, riconosce e fornisce l'output giusto.

Apprendimento da esempi

E’ importante fornire alla rete un insieme significativo di esempi: deve coprire tutte le possibili combinazioni. Servono almeno 500 esempi.

Esempio: riconoscimento dei volti. Bisogna preparare un insieme che comprende facce di uomini e donne, di tutte le età e con tutte le espressioni facciali possibili.

Non programmare ma apprendere

La programmazione serve solo per creare il software che crea la rete e l'algoritmo di apprendimento. Per insegnare alla rete a risolvere un problema, occorre un periodo di apprendimento in cui insegnare alla rete come comportarsi con l'input che riceve, perché all'inizio la rete non ha nessuna forma di conoscenza.

L'operatore crea la struttura della rete e quindi i pesi sono gli unici parametri che possono essere modificati. Infatti la conoscenza è memorizzata sui pesi e la rete apprende usando tecniche di ottimizzazione per variare i valori dei pesi, cercando di minimizzare una funzione di errore.

Tipi di apprendimento

Ogni modello di rete neurale ha il suo specifico tipo di apprendimento:

- supervisionato
- non supervisionato
- hebbiano

Apprendimento supervisionato

Alla rete viene presentato un training set preparato da un supervisore esterno, e composto da molte coppie significative di valori (input,output atteso):

- la rete riceve l'input e calcola il suo corrispondente output

- per un certo input, l'errore è dato dalla differenza tra l'output della rete e l'output atteso; serve a supervisionare l'apprendimento per far capire alla rete quanto si sbaglia nel calcolare quell'output

- la rete modifica i pesi in base all'errore cercando di minimizzarlo e commetterà sempre meno errori.

Apprendimento supervisionato

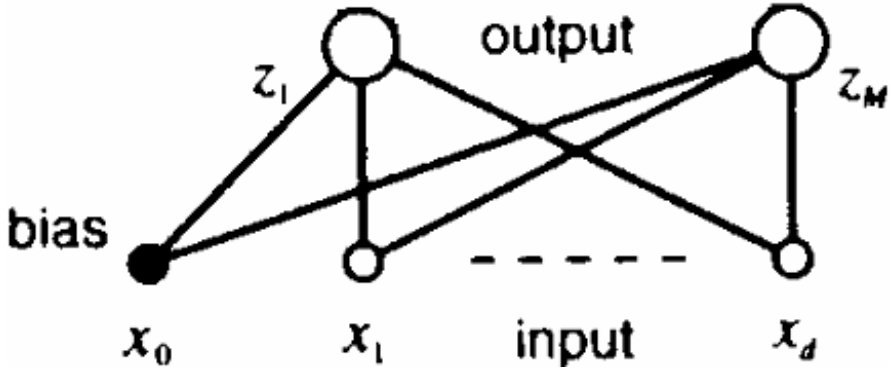
Occorre preparare alcuni esempi di funzionamento studiati appositamente, e la rete impara da questi esempi.

Modelli che usano questo apprendimento:

- perceptron
- multi layer perceptron MLP
- radial basis function RBF

Il perceptron

Modello creato da Rosenblatt nel 1957 per riconoscere immagini simulando la percezione umana. Ogni neurone ha funzione di attivazione a gradino ed uscita con valori binari: 1 indica la presenza di un oggetto, 0 l'assenza. Composto da neuroni affiancati.

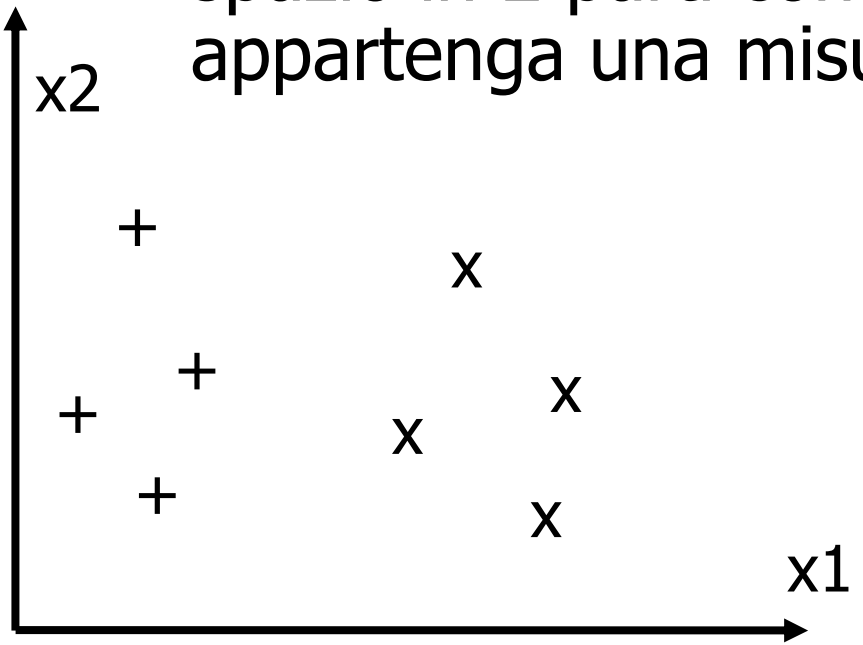


Spazio delle feature

- Le feature di un oggetto sono N numeri che descrivono le proprietà caratteristiche, significative dell'oggetto. Sono unite in un vettore a N dimensioni.
- Lo spazio delle feature è un iperspazio a N dimensioni in cui rappresentare e classificare le feature.

Spazio delle feature

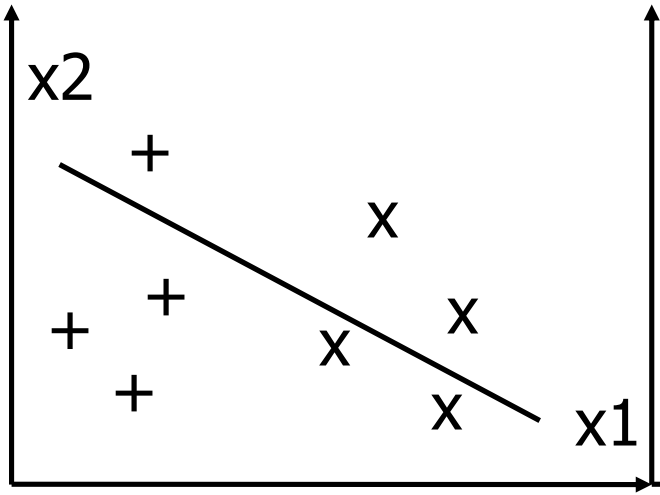
Esempio: un palazzo ed una villa hanno entrambi altezza x2 e larghezza x1. Sullo spazio delle feature vanno le misure di alcuni oggetti indicate con x (villa) e + (palazzo). Trovare una regola per dividere lo spazio in 2 parti con cui decidere a quale classe appartenga una misura di un nuovo oggetto.



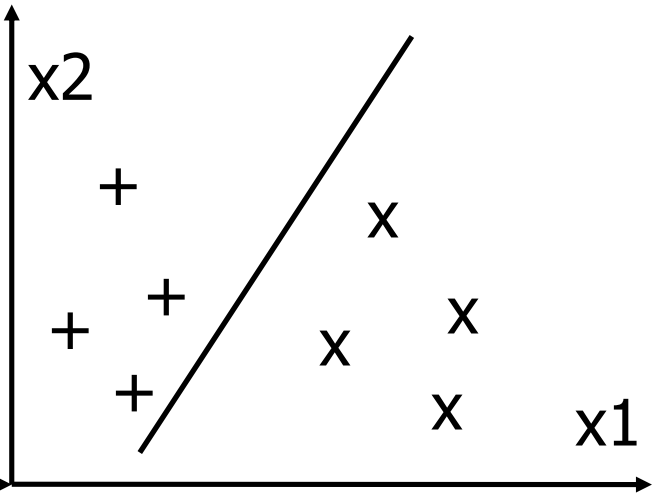
Il perceptron

Spazio di 2 valori input linearmente separabile e perceptron con apprendimento trova sempre una retta di separazione del piano nei due semipiani con i gruppi di oggetti

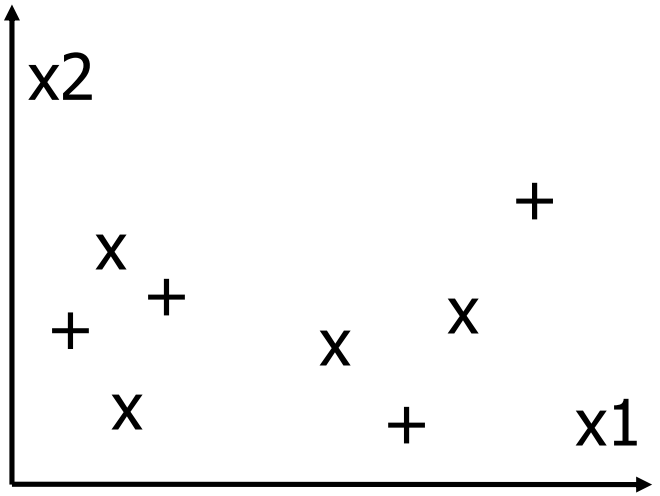
prima



dopo

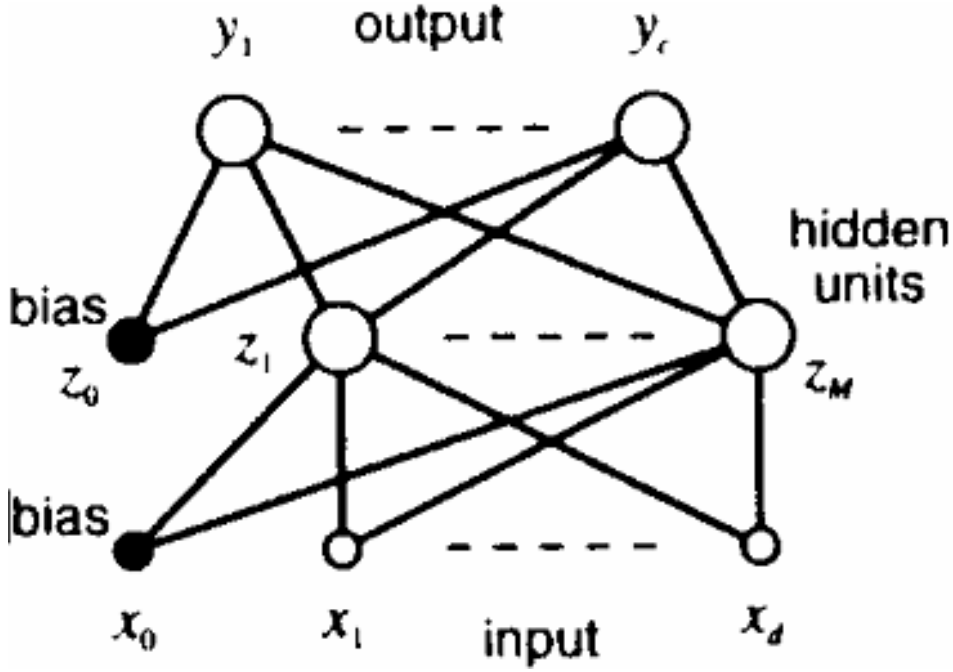


Spazio di 2 valori input non linearmente separabile e perceptron non trova una retta di separazione



Multi Layer Perceptron (Mlp)

Formato dalla sovrapposizione di vari perceptron. Il livello di input non contiene neuroni, nella figura ci sono 2 livelli. Ogni neurone è collegato con tutti i neuroni dello strato precedente e successivo, i neuroni sullo stesso strato non sono collegati, non esistono cicli dallo strato di output verso strato di input.



Struttura di Mlp

la conoscenza viene elaborata dal livello input verso il livello output: si calcola l'output di tutti i nodi di un livello che diventano poi gli input dei nodi del livello successivo

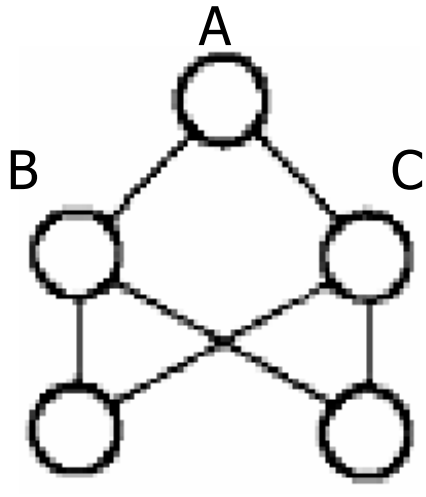
Mlp con c valori input, M neuroni hidden, K neuroni output,

$$y_k(x) = g\left(\sum_{j=0}^{M} w_{kj}^{(2)}\, g\left(\sum_{i=0}^{c} w_{ji}^{(1)} x_i\right)\right)$$

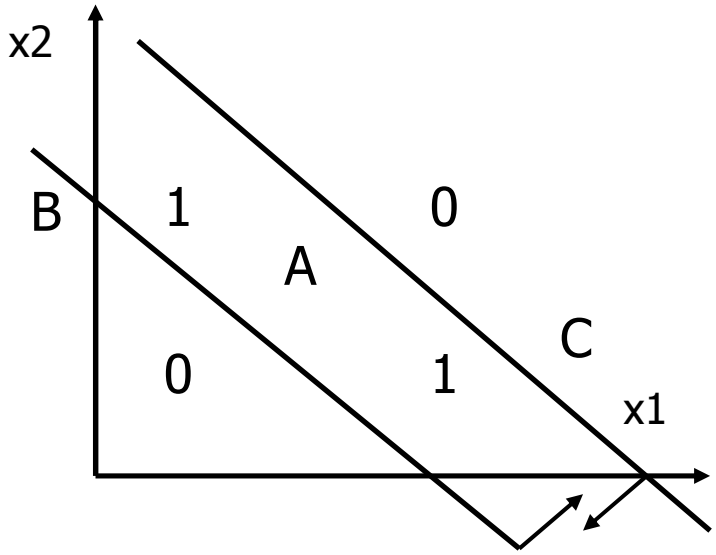
g funzione di attivazione dei neuroni, $w^{(2)}_{kj}$ il peso dal neurone k al neurone j al livello 2 della rete, y_k valore del neurone di output numero k

i neuroni allo stesso livello hanno la stessa funzione di attivazione

Regioni delimitate da Mlp

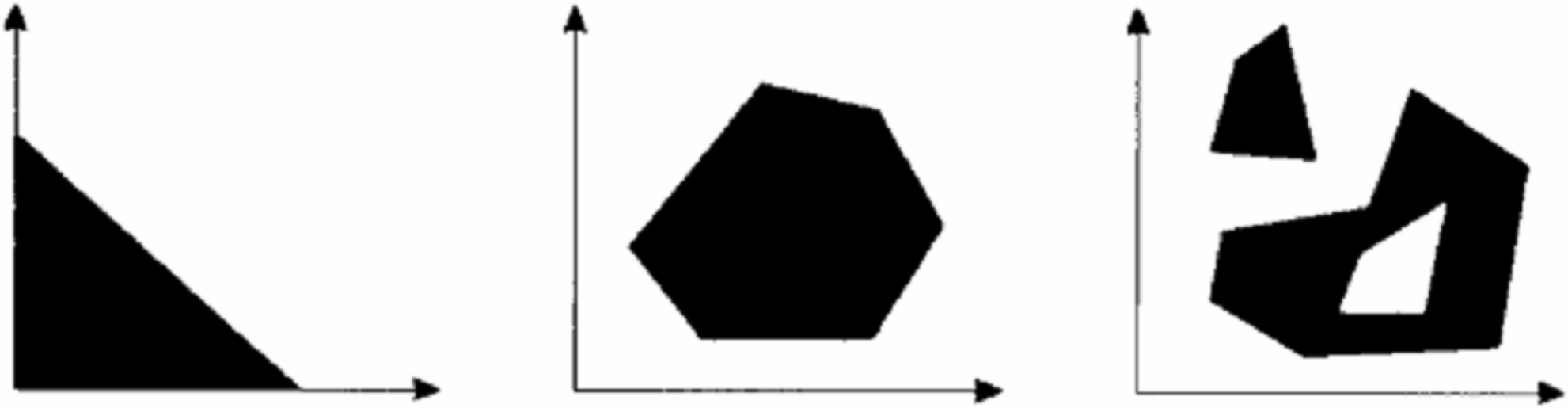


Supera i limiti del perceptron usando strati di neuroni hidden (interni) che realizzano una rappresentazione interna dell'input più complessa, perché individua regioni arbitrarie intersecando iperpiani nello iperspazio dei valori input.



Nell'esempio i perceptron B e C creano ciascuno un semipiano individuato dalle rette e il perceptron A interseca i due semipiani individuando la parte di piano tra le rette.

Regioni delimitate da Mlp



Con funzione di attivazione a gradino, a sinistra la regione delimitata da perceptron, al centro la regione delimitata da Mlp a 1 livello hidden, a destra la regione delimitata da Mlp con 2 livelli hidden che delimita aree arbitrarie.

Funzione di errore

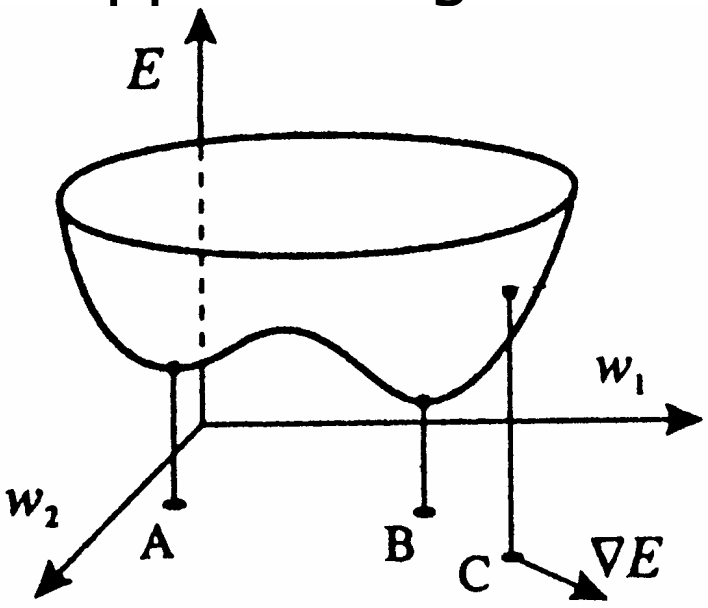
Esprime la differenza fra l'output della rete y e l'output desiderato y' nell'apprendimento. Si scrive E(w) perché la conoscenza risiede sui pesi w e l'output è ottenuto in base al valore dei pesi; quindi devo modificare i pesi tramite la E derivabile nei pesi w ed occorre trovare il vettore w che rende minimo l'errore. Esistono molte formulazioni di E.

L'uso della funzione di attivazione sigmoide rende facile il calcolo delle derivate. In generale

$$E(w) = \sum_{i=1}^{c} \left(\left(y_i - y_i^{'} \right)^2 \right)$$

Funzione di errore

E è non lineare: gli algoritmi cercano un minimo nella sua superficie con modifiche di w in base al gradiente di E. A sinistra c'è una superficie di E quadratica nei pesi w_1 e w_2; A e B sono minimi e C punto di calcolo del gradiente locale della superficie, si va nella direzione opposta al gradiente. Può non convergere verso il minimo assoluto B ma verso un minimo locale A; in alcuni punti il gradiente è nullo.



Il gradiente ∇E indica la direzione di crescita di E per cui andiamo nella direzione opposta per trovare il minimo di E.

Formule del back-propagation

Ogni neurone computa $a_j = \sum_i w_{ji} z_i$ con z_i l'attivazione di unità che manda connessione a unità j e w_{ji} il peso tra unità j e i, la sommatoria è su tutte le unità connesse all'unità j. Usando funzione attivazione g si ha l'attivazione z_j dell'unità j $z_j = g(a_j)$ Se le z_j indicano unità input allora sono $z_j=x_j$, se le unità j sono di output l'attivazione è denotata $y_k=z_j$.

$E = \sum_n E^n$ è l'errore per n pattern input.

Formule del back-propagation

La variazione dell'errore dipende dalla $\frac{\vartheta E^n}{\vartheta w_{ji}} = \frac{\vartheta E^n}{\vartheta a_j}\frac{\vartheta a_j}{\vartheta w_{ji}}$

variazione dei valori dei pesi, in cui

il primo fattore riflette la variazione dell'errore in funzione della variazione dell'input all'unità e l'altro riflette la variazione di un certo peso sull'input

posto $\delta_j = \frac{\vartheta E^n}{\vartheta a_j}$ e $\frac{\vartheta a_j}{\vartheta w_{ji}} = z_i$ si ha $\frac{\vartheta E^n}{\vartheta w_{ji}} = \delta_j z_i$ serve

calcolare δ_j per ogni unità

Formule del back-propagation

Calcolo δ_k per unità output

$$y_k = g(a_k) \qquad \frac{\vartheta y_k}{\vartheta a_k} = g'(a_k) \qquad \text{e} \qquad \frac{1}{\vartheta a_k} = \frac{g'(a_k)}{\vartheta y_k}$$

si ha

$$\delta_k = \frac{\vartheta E^n}{\vartheta a_k} = g'(a_k)\frac{\vartheta E^n}{\vartheta y_k}$$

con z_k denotato da y_k occorre g' e la derivata che si calcola subito perché al livello output si ha output desiderato e output da rete, quindi si calcola subito l'errore e il gradiente

Formule del backpropagation

Calcolo δ_j per unità hidden

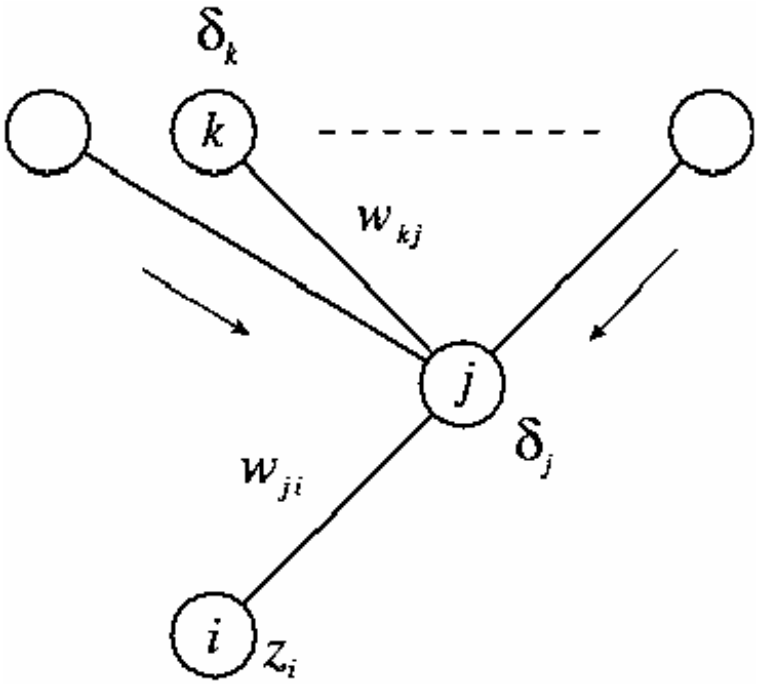
serve errore da unità di livello superiore per calcolare la funzione di errore locale con somma su tutte unità k di livello superiore cui unità j manda connessioni

$$\delta_j = \frac{\vartheta E^n}{\vartheta a_j} = \sum_k \frac{\vartheta E^n}{\vartheta a_k} \frac{\vartheta a_k}{\vartheta a_j} = \sum_k \delta_k \frac{\vartheta a_k}{\vartheta a_j}$$

cioè la sommatoria dell'errore retropropagato per la quantità di errore dato dall'influenza dell'unità j su unità k

Formule del back-propagation

δ_j dipende solo da valori input e da pesi di connessioni di j cioè dai neuroni vicini per cui il calcolo è locale, quindi:



- se errore è piccolo e peso ha valore grande allora la connessione non porta molto errore

- se errore è grande e peso ha valore grande allora la connessione porta molto errore e il peso va cambiato

Formule del back-propagation

Sostituendo la definizione di δ e $a_k = \sum_j w_{kj} z_j = \sum_j w_{kj} g(a_j)$

$\frac{\vartheta a_k}{\vartheta a_j} = w_{kj} g'(a_j)$ sostituendo $\delta_j = g'(a_j) \sum_k w_{kj} \delta_k$

per cui i valori di un δ di una unità hidden dipendono dai δ delle unità a livello più alto; poiché i δ delle unità output si calcolano subito, gli altri si ottengono ricorsivamente.

Formule del backpropagation

Variazione del peso $w_{ji}^{t+1} = w_{ji}^{t} - \eta \frac{\vartheta E^n}{\vartheta w_{ji}} = w_{ji}^{t} - \eta \delta_j z_i$ in

cui si addiziona ad ogni peso un incremento, positivo o negativo determinato dalla sua influenza nella formazione dell'errore, e η è il learning rate (coefficiente di apprendimento) tra 0 e 1, in generale

$$\Delta w_{ji} = w_{ji}^{t+1} - w_{ji}^{t} = -\eta \frac{\vartheta E^n}{\vartheta w_{ji}} = -\eta \delta_j z_i$$

valori di pesi in [-1,1]

Algoritmo per il backpropagation

Si esegue un'epoca di apprendimento, cioè si attuano per tutte le coppie (input, output) significative del training set i passi seguenti:

Algoritmo back propagation

1. prende una coppia e calcola la risposta della rete per quell'input; il calcolo procede dal livello input verso il livello output calcolando l'attivazione di tutte le unità, quindi propaga in avanti l'errore
2. calcola l'errore E tra l'output della rete e output della coppia e calcola i δ_k delle unità output
3. propaga all'indietro l'errore verso il livello di input, calcolando i δ_i per ogni unità hidden
4. variazione dei pesi
5. ripete dal passo 1 fino a terminare le coppie
6. calcola l'errore globale e se è ancora alto si ripete l'epoca di apprendimento.

Problemi del back propagation

- algoritmo lento che può finire intrappolato in un minimo di E credendo di aver trovato il valore ottimo dei pesi che rende minima la E
- dipendenza dal valore iniziale dei pesi w; può capitare di cominciare da un punto della superficie di errore già molto vicino a un minimo locale
- scelta critica di η: troppo piccolo crea apprendimento lento, troppo grande crea oscillazioni

Problemi del backpropagation

Esistono tecniche che aumentano la velocità di convergenza verso il minimo assoluto di E: smorzano le oscillazioni aggiungendo alla formula di variazione del peso un altro parametro α detto momentum che tiene conto delle oscillazioni precedenti.

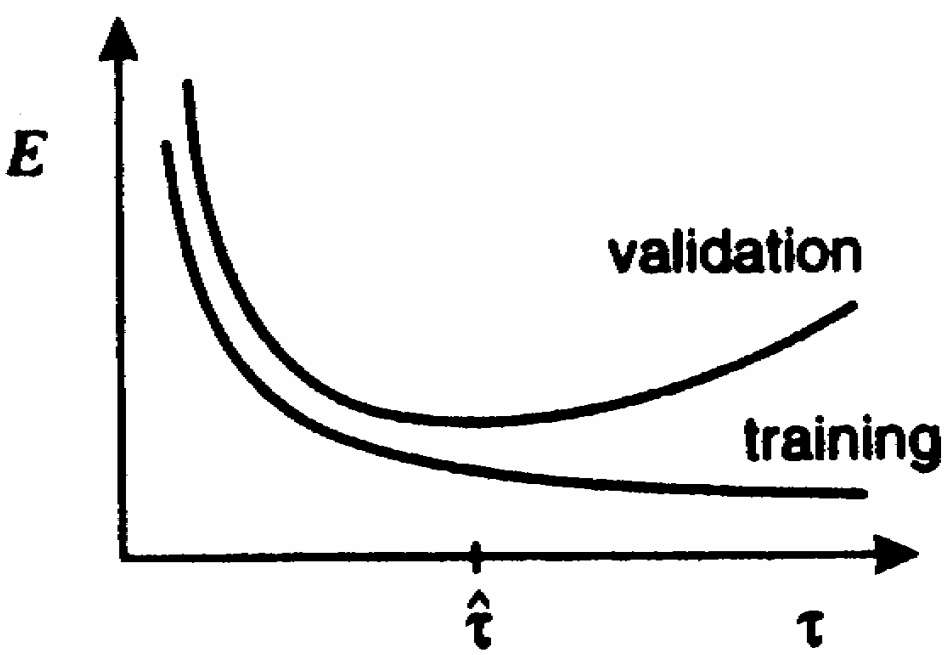
$$\Delta w^t = -\eta \delta_j z_i + \alpha \Delta w^{t-1}$$

Overfitting

La rete deve comprendere il modello statistico dei dati, non deve memorizzare i soli dati del training set creando il fenomeno di overfitting. Così può generalizzare, cioè rispondere esattamente a input non in training set.

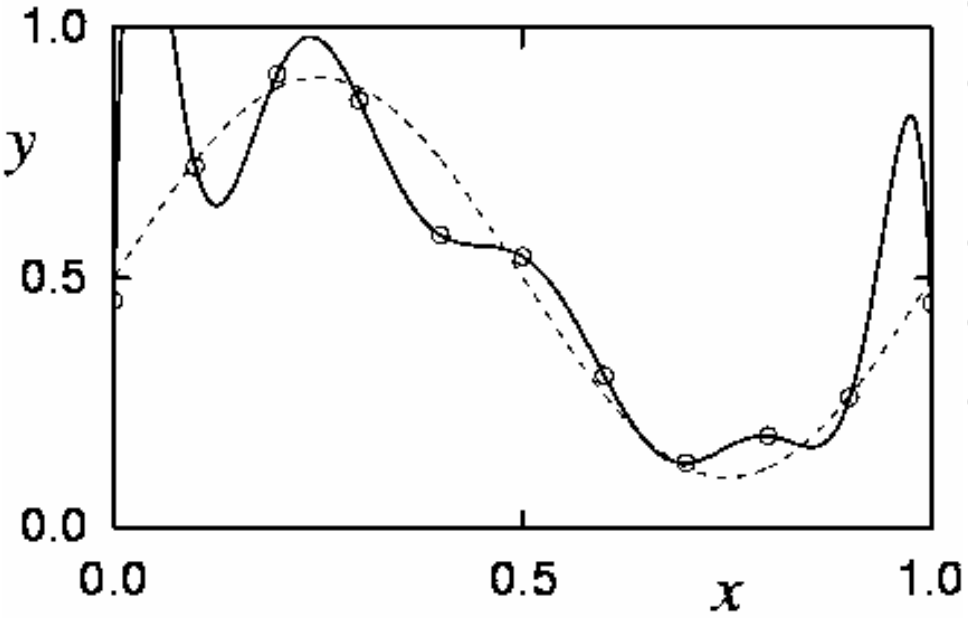
Early stopping

Aumenta la capacità di generalizzare. Durante lo apprendimento l'errore E tende a 0 perché la rete sta imparando meglio. Con un validation set di coppie non usate nel training set, si misura la E e si crea la seconda curva che tende a un minimo e poi ricresce: da quel punto la rete sta imparando il training set e non il modello statistico. Si ferma l'apprendimento al valore cioè $\hat{\tau}$ il minimo di E rispetto al validation set dove τ indica le epoche.



Overfitting

I cerchi sulla curva tratteggiata indicano i punti del training set, la curva in tratteggio è la funzione da apprendere, la curva continua è la funzione appresa dalla rete neurale con errore nullo di training senza uso di early stopping. Evidente la differenza tra le curve ed i valori nei cerchi che non appartengono alla curva tratteggiata:la rete neurale ha imparato esattamente solo i valori nei cerchi del training set e fa grossi errori negli altri punti.



Valutare prestazioni

Per misurare le prestazioni di una rete neurale dopo l'apprendimento, si crea il test set formato da coppie non usate per i training e validation set. In genere il test set è un terzo della grandezza del training set ed è composto da input critici su cui la risposta della rete deve essere buona, altrimenti si butta via la rete.

Per ogni coppia del test set:

- calcolare la risposta della rete all'input
- calcolare l'errore dato dalla differenza tra output rete e output atteso di coppia

Valutare prestazioni

L'errore totale è dato dalla somma degli errori avuti per ogni coppia.

Secondo il tipo di problema, si usano varie tecniche statistiche per decidere se usare o meno la rete creata; in genere si accetta una rete se sul test set ha mostrato un errore inferiore al 20-25% delle coppie.

Radial basis function (Rbf)

Disposizione dei neuroni uguale alla Mlp.

Un solo livello hidden di M neuroni che realizza funzioni di base $\phi_j(x)$ circolari centrate sui punti dello spazio di input; esistono varie forme delle funzioni, in genere si usano gaussiane. Ideale per interpolazioni di funzioni e di predizione in serie temporali, in cui stimare il valore successivo in una serie di numeri.

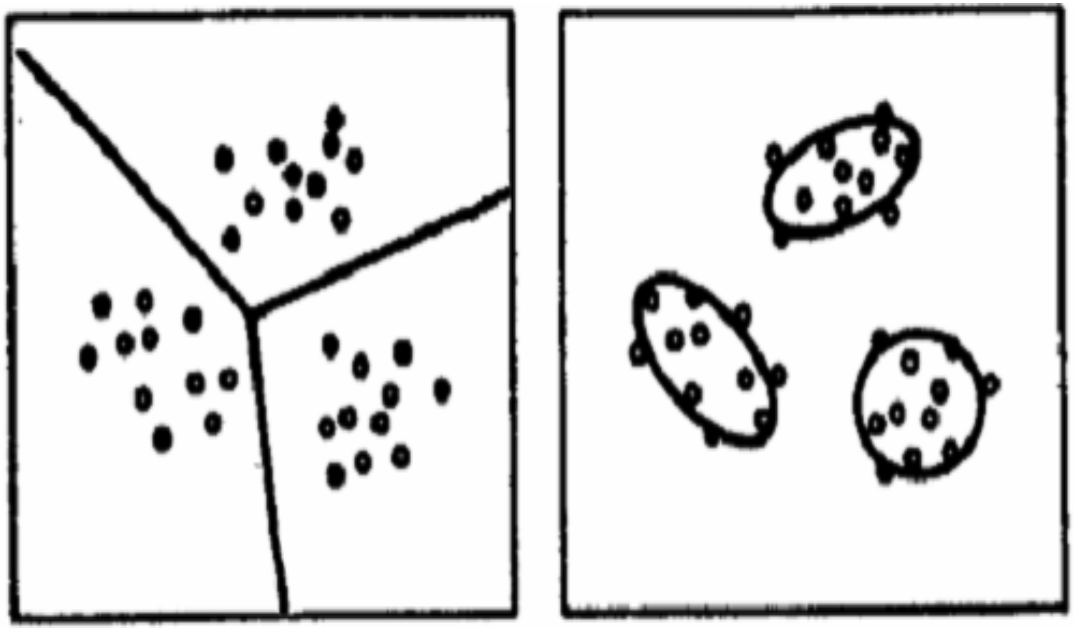
$$y_k(x) = \sum_{i=1}^{M} w_{kj}\phi_j(x) + w_{k0}$$

Equazione del neurone output k su input x, w_{kj} è il peso tra neuroni k e j

Radial basis function (Rbf)

Apprendimento: i parametri delle funzioni di base devono essere appresi, altrimenti ho stesse capacità di Mlp, poi si calcolano i pesi delle connessioni verso il nodo di output.

Divisione dello iperspazio dei valori input: a sinistra Mlp che usa intersezione di iperpiani, a destra Rbf con cluster per raggruppare insieme solo i punti simili.



La funzione di base deve specificare la forma ed il centro del cluster.

Apprendimento non supervisionato

Alla rete vengono presentati solo i valori di input e la rete li divide autonomamente in gruppi usando misure di similarità, senza usare confronti con output noti, e cercando di mettere input simili nello stesso gruppo. E' un apprendimento autonomo e non c'è controllo esterno sull'errore. Adatto per ottimizzare risorse e se non si conoscono a priori i gruppi in cui dividere gli input.

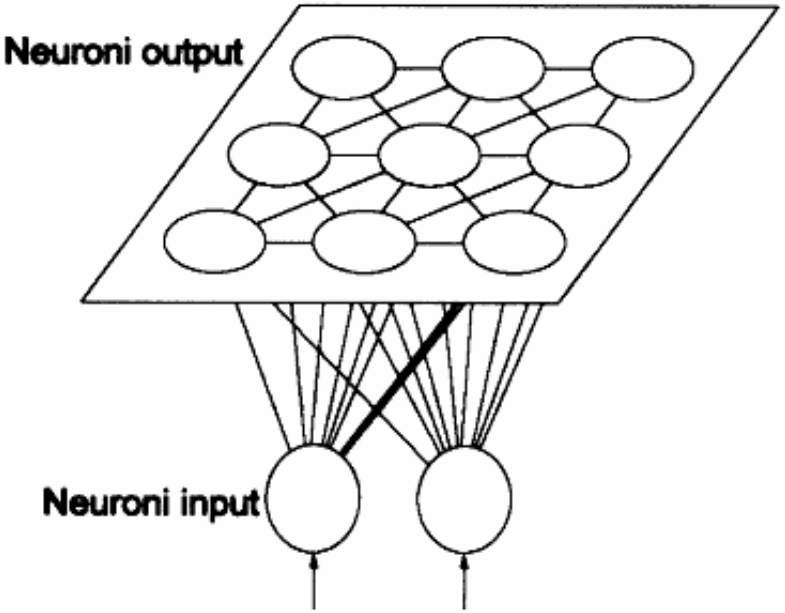
Modelli che usano questo apprendimento:

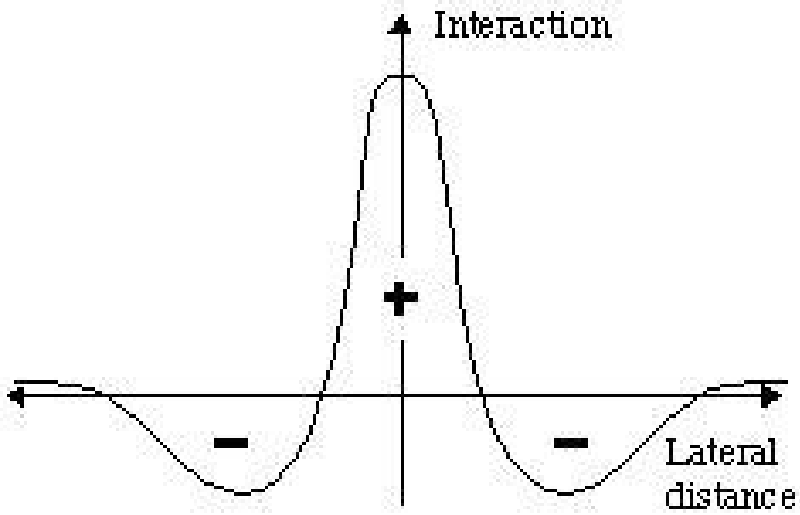
-Kohonen

-Hopfield

Self Organizing Maps (SOM)

Ideate da Tuevo Kohonen nel 1982 ispirandosi alla topologia della corteccia del cervello dove neuroni vicini sono attivati da stimoli simili. Tiene conto delle connessioni tra neuroni e dell'influenza che può avere un neurone sui suoi vicini: i neuroni vicini a neuroni attivi rinforzano i legami, a quelli che si trovano ad una data distanza vengono indeboliti i legami.



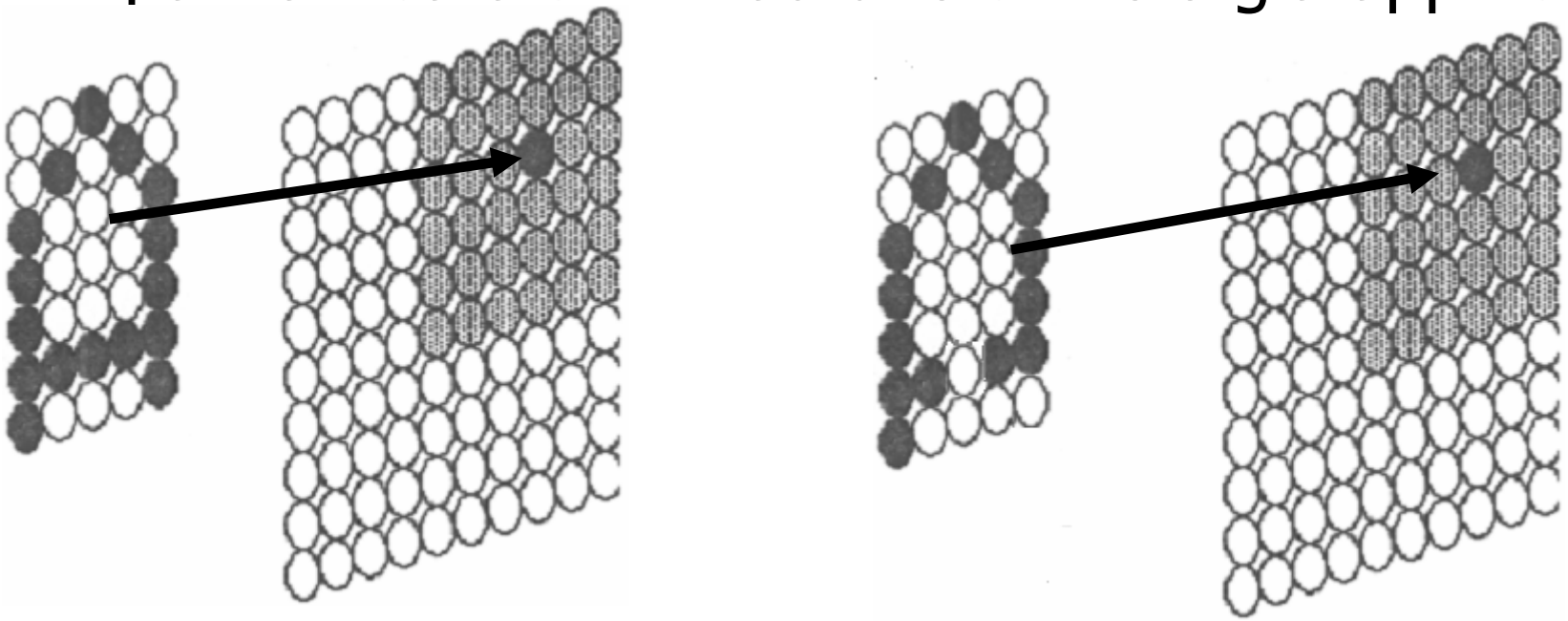


Self Organizing Maps (SOM)

- Una rete SOM è composta dal livello di input e livello di neuroni che realizza l'output e la competizione che sono localizzati su griglia ad 1 o 2 dimensioni e con diverse forme della griglia.
- Ciascun neurone di input è connesso a tutti i neuroni della griglia; ogni neurone di output ha il vettore dei pesi con le stesse dimensioni del vettore di input.
- La configurazione finale dei pesi dei singoli neuroni permette di suddividere gli elementi forniti in ingresso in cluster (raggruppamenti di oggetti simili) che di quegli elementi rappresentano una classificazione.
- Occorre localizzare sulla mappa i neuroni attivi ed associarli con gli input presentati.

Interpretazione della mappa

Alla fine dell'addestramento, ogni lettera attiverà una differente combinazione di neuroni contenente il neurone vincitore per quella lettera. Il neurone vincente sarà attivato anche da sottomissione della lettera parziale o con rumore perché lo stimolo parziale sarà simile ad uno stimolo già appreso.



Apprendimento in Som

N valori in input creano un punto $\alpha=(\alpha_1, .., \alpha_N)$ in spazio a N dimensioni. Le unità di output O_i sono disposte come in esempio e sono connesse completamente a unità input con pesi w_{ij}. La regola di apprendimento competitivo seleziona come vincitore i* l'unità output avente il vettore dei pesi più vicino all'input α secondo $\left|w_{i*} - \alpha\right| \leq \left|w_i - \alpha\right| \forall i$ quindi la regola è

$\Delta w_{ij} = \eta \Lambda(i, i*)(\alpha_j - w_{ij}) \forall i, j$ dove $\Lambda(i, i*) = 1$ se $i = i*$

e il valore diminuisce con l'aumento di distanza tra le unità i e i* nel vettore di output.

Apprendimento in Som

L'unità vincitrice i* e quelle vicine a lei hanno forti modifiche dei pesi; le unità lontane, per cui $\Lambda(i, i^*)$ assume un piccolo valore, hanno piccole variazioni dei pesi.

Apprendimento in Som

La Λ è la funzione di vicinato che contiene le informazioni topologiche dello strato di output della rete.

Trascina verso α il vettore dei pesi dell'unità vincitrice e i vettori dei pesi delle unità vicine, così la rete diventa elastica e si avvicina agli input.

Per avere veloce convergenza la Λ ha all'inizio un ampio raggio di azione e un η elevato, poi vengono gradualmente diminuiti.

Creare un training set con molti vettori di input significativi; stavolta non ci sono output con cui paragonare la risposta della rete.

Apprendimento in Som

Riepilogando per ogni vettore di input:

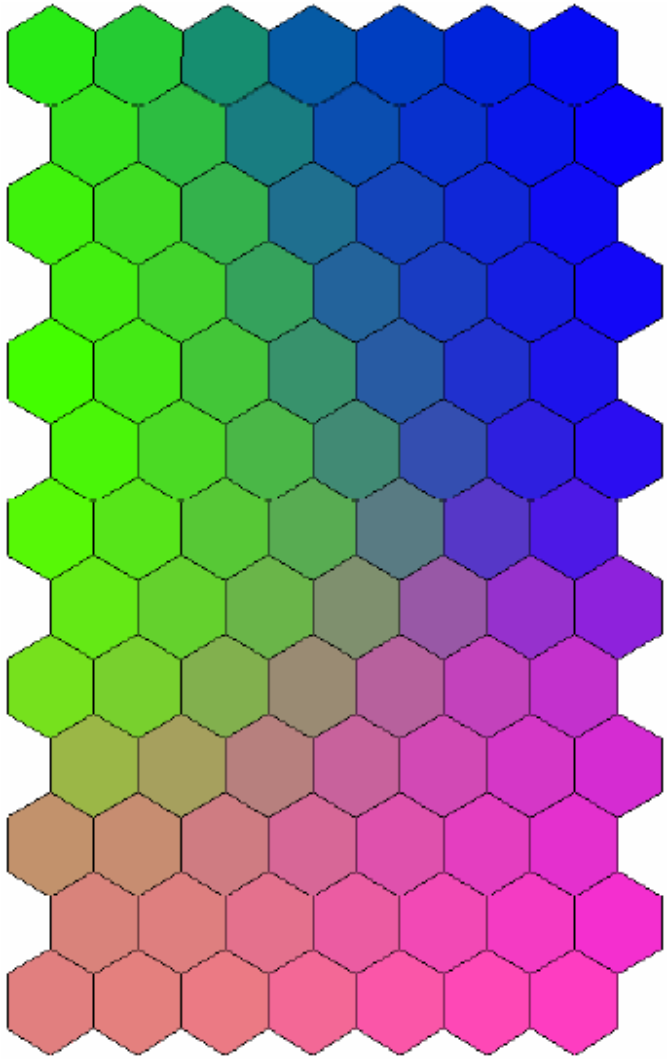
1. ogni neurone di output riceve le componenti del vettore input e ne calcola la distanza euclidea dal suo vettore di pesi
2. il neurone di output, che ha la minima distanza euclidea dall'input, si attiva e dà una risposta maggiore, modifica poi i suoi pesi e quelli dei neuroni vicini per avvicinarli all'input.

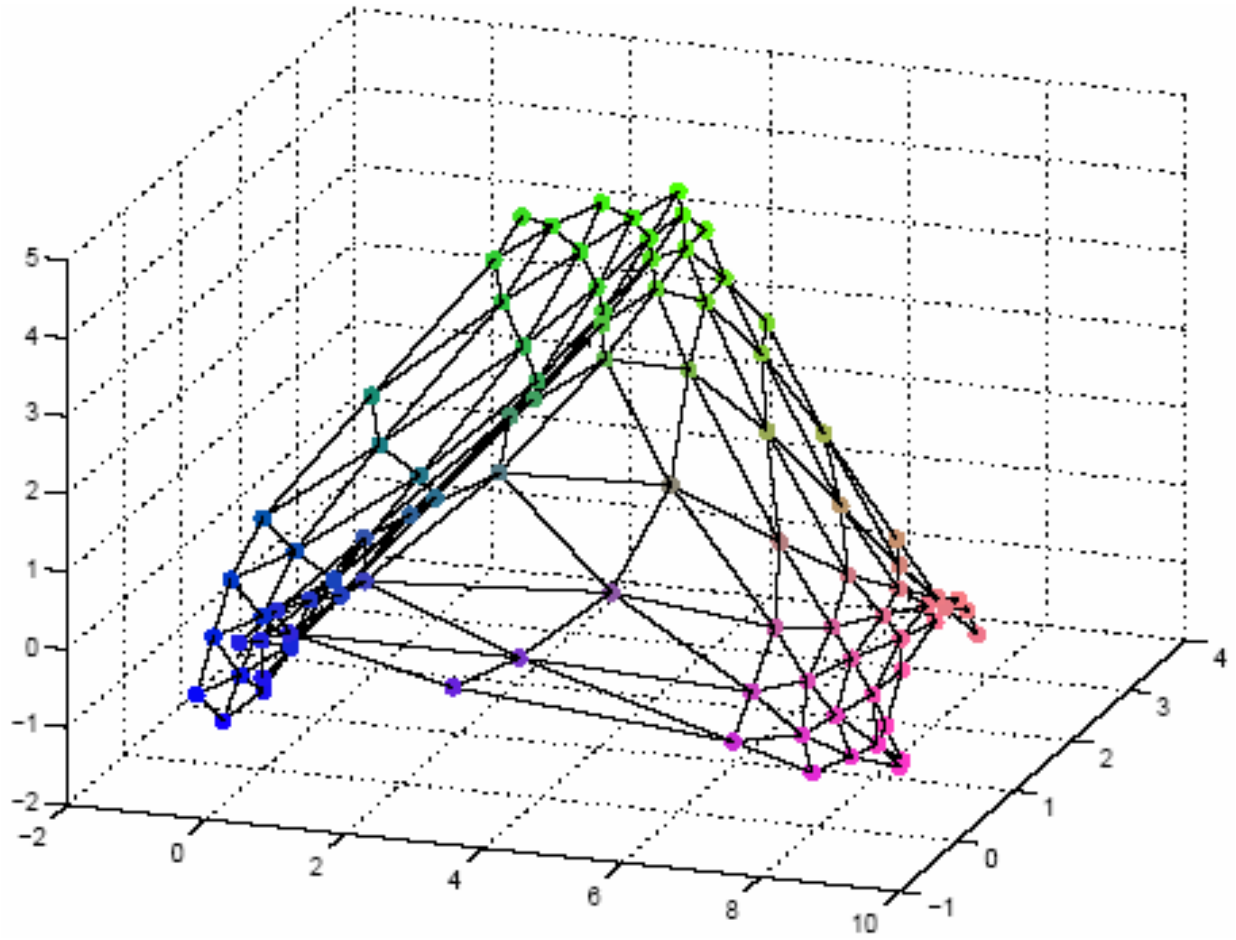
Così input simili attiveranno neuroni vicini.

Occorre stabilire quali sono i neuroni vicini e come modificare i loro pesi.

Visualizzazione numeri neuroni output SOM

- ad ogni numero corrisponde uno specifico colore secondo la mappa dei colori e neuroni con colori simili formano un cluster di neuroni

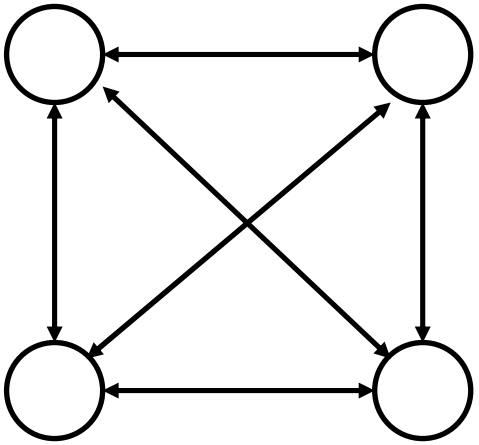




Reti di Hopfield

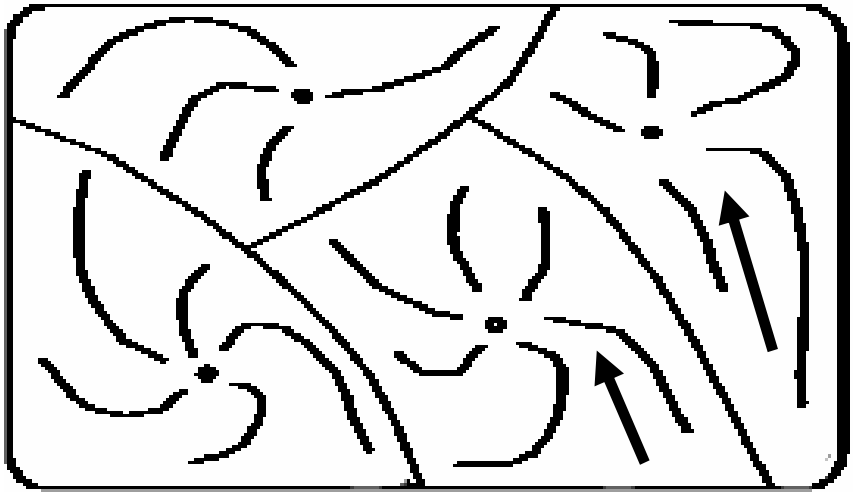
Hopfield nel 1982 propone una rete per memorizzare informazioni. Composta da neuroni completamente connessi, funzione di attivazione a gradino, ogni neurone è nodo di ingresso e di uscita e ha stato attivo o disattivo.

$$a_i = \sum_j w_{ij} x_j \qquad g(a_i) = \begin{cases} +1 \ se \ a_i > 0 \\ -1 \ altrimenti \end{cases}$$



Reti di Hopfield

Si associa una funzione energia da minimizzare durante l'apprendimento realizzato con una successione di stati, fino a raggiungere uno stato finale stabile corrispondente al minimo della funzione energia. La funzione energia viene rappresentata con una superficie con delle buche che costituiscono i punti di minimo.



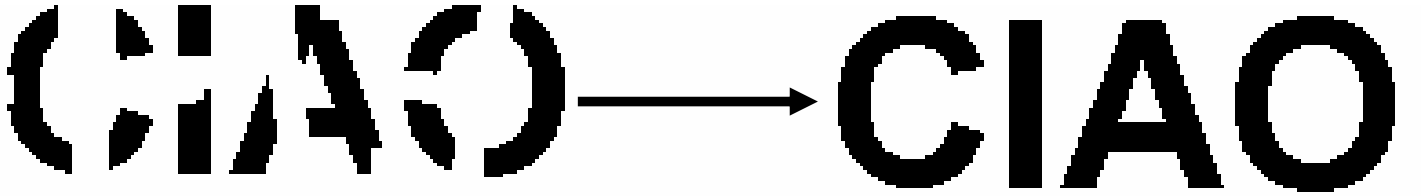
Reti di Hopfield

Proprietà:

- rappresentazione distribuita su tutti i neuroni
- adatta per ottimizzazione vincolata, quando il problema può essere espresso in termini di una funzione obiettivo da trasformare in una funzione energia
- recupero di informazioni anche con danni alla rete
- adatta per creare memoria indirizzabile per contenuto, cioè che recupera un'informazione non tramite l'indirizzo in memoria ma basta produrne una versione parziale del contenuto, invece la perdita di un bit nell'indirizzo in memoria non permette più il recupero del contenuto

Reti di Hopfield

- viene inserito l'input a tutti i neuroni, durante la ricerca la rete esplora la superficie di energia cercando il punto di minimo (la buca o attrattore) più vicina, raggiunge la stabilità e dall'output di tutti i neuroni si risale al contenuto associato al punto di minimo



Apprendimento hebbiano

Ideato dallo psicologo Donald Hebb.

L'attivazione simultanea di due neuroni connessi determina la modifica dei pesi delle connessioni che li uniscono, in modo da aumentare la probabilità che uno dei due neuroni si attivi quando l'altro si attiva. Così si rafforza la connessione tra due neuroni molto usati e si indebolisce la connessione tra due neuroni poco attivati contemporaneamente.

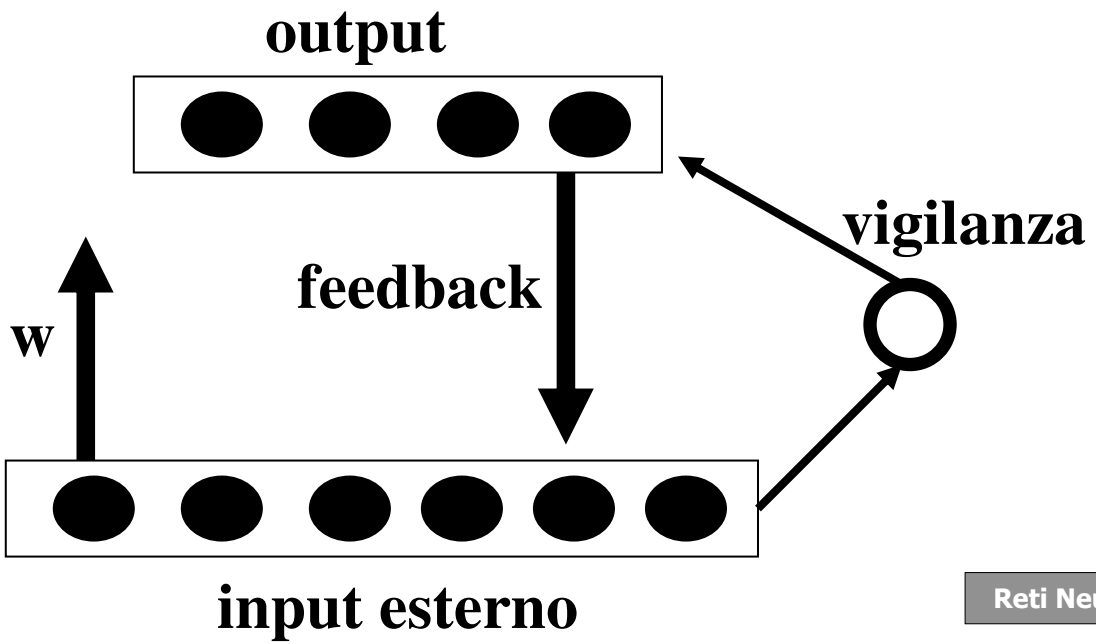
Teoria della risonanza adattiva ART

- Dopo aver creato una rete neurale il numero di neuroni output, che individua una delle possibili classi in cui classificare i valori di input, non è modificabile.
- L'analisi di nuovi dati non usati nei tre set suggerisce la creazione di una altra classe per migliorare la classificazione dell'input, ma ormai la rete è bloccata (stabilità) e non c'è tempo per creare una altra rete e rifare l'addestramento (plasticità).
- Il modello ART di S. Grossberg supera il dilemma della stabilità-plasticità quindi

- può apprendere nuove classificazioni e aggiungere altri neuroni di output in base all'analisi dei nuovi dati
- non dimentica quanto appreso in precedenza e non bisogna rifare l'addestramento su tutti i dati

Teoria della risonanza adattiva ART

- Una funzione di vigilanza decide quando i nuovi dati sono molto diversi dai precedenti ed occorre creare un nuovo neurone di output.
- Modello auto-organizzante per spiegare come il cervello utilizza tutte le categorie esistenti nel codificare nuove esperienze con apprendimento continuo on-line e senza supervisione.



Parte 4

Creare una rete per classificare

La classificazione

Classificare significa dividere un insieme di oggetti in insiemi disgiunti secondo un criterio stabilito a priori; in genere si assegna una etichetta ad ogni insieme creato.

Il pattern recognition (riconoscimento di configurazioni) è la tecnica che consente di creare classificatori numerici e automatici.

Ogni oggetto deve essere rappresentato con un vettore di numeri per essere classificato da una rete neurale, per cui ad ogni oggetto si associa un pattern, un vettore di feature che contraddistingue univocamente l'oggetto.

Il classificatore numerico

Un classificatore numerico si può così definire: date N classi di appartenenza tra cui discriminare, il vettore di input x a L dimensioni delle feature da classificare, il vettore di uscita y che individua la classe formato da N valori, un classificatore riceve in input il vettore x e restituisce in uscita il vettore y dove $y_i=1$ se l'oggetto con input x appartiene alla classe i e $y_j=0$ per $i \neq j$, per i,j=1..N.

E' quindi un mapping, o corrispondenza tra valori di input ed output, che può essere modellato con una funzione non lineare; data la non linearità, conviene impiegare una rete neurale.

Scelta delle feature

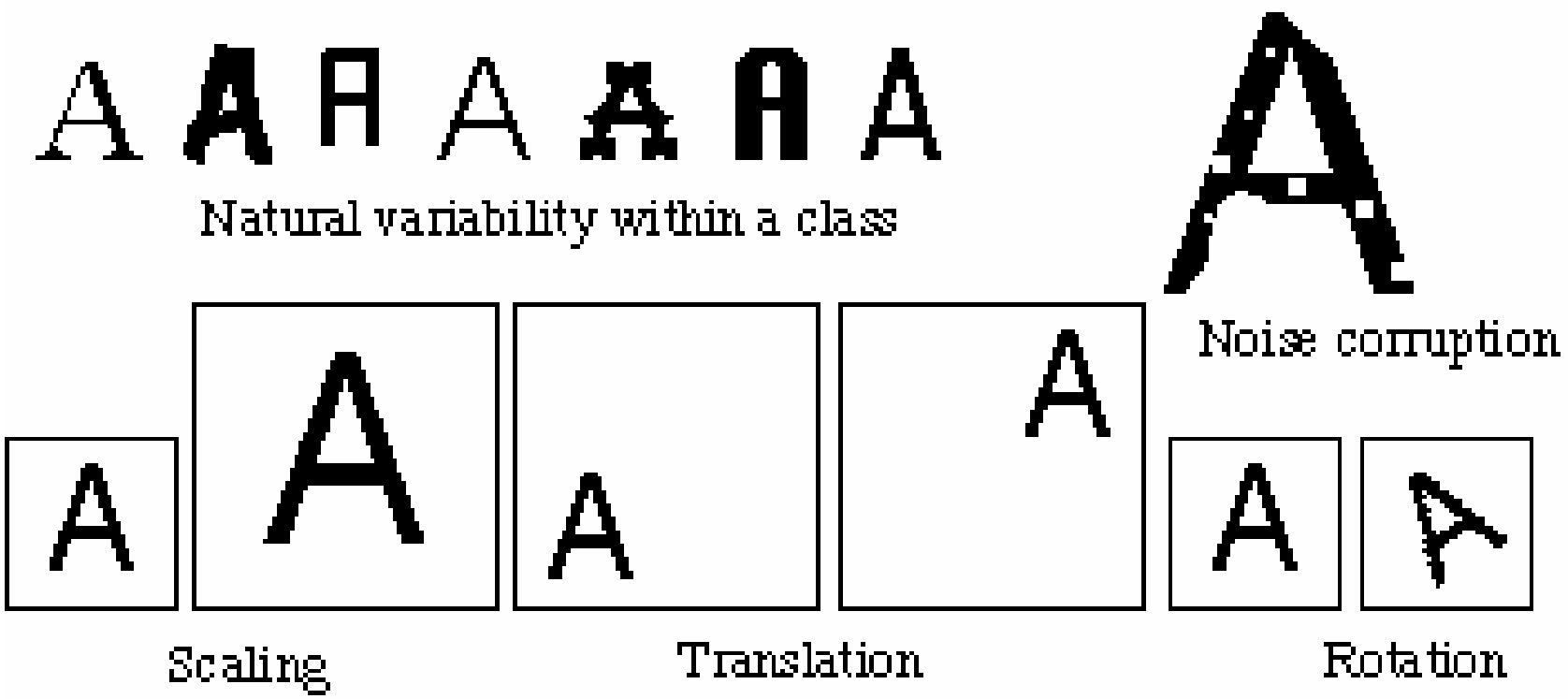
Le feature devono essere scelte con le proprietà:

- discriminanza: i valori delle feature sono simili per oggetti appartenenti alla stessa classe e sono molto diversi per oggetti appartenenti a classi diverse
- indipendenza: i valori delle feature non devono essere correlati tra loro
- minimalità: devono essere il minimo numero possibile di proprietà
- disponibilità: facili e veloci da calcolare.

L'intervallo dei valori di ogni feature deve essere normalizzato in [0,1] o [-1,1] per non avere ordini di grandezza troppo diversi nell'apprendimento.

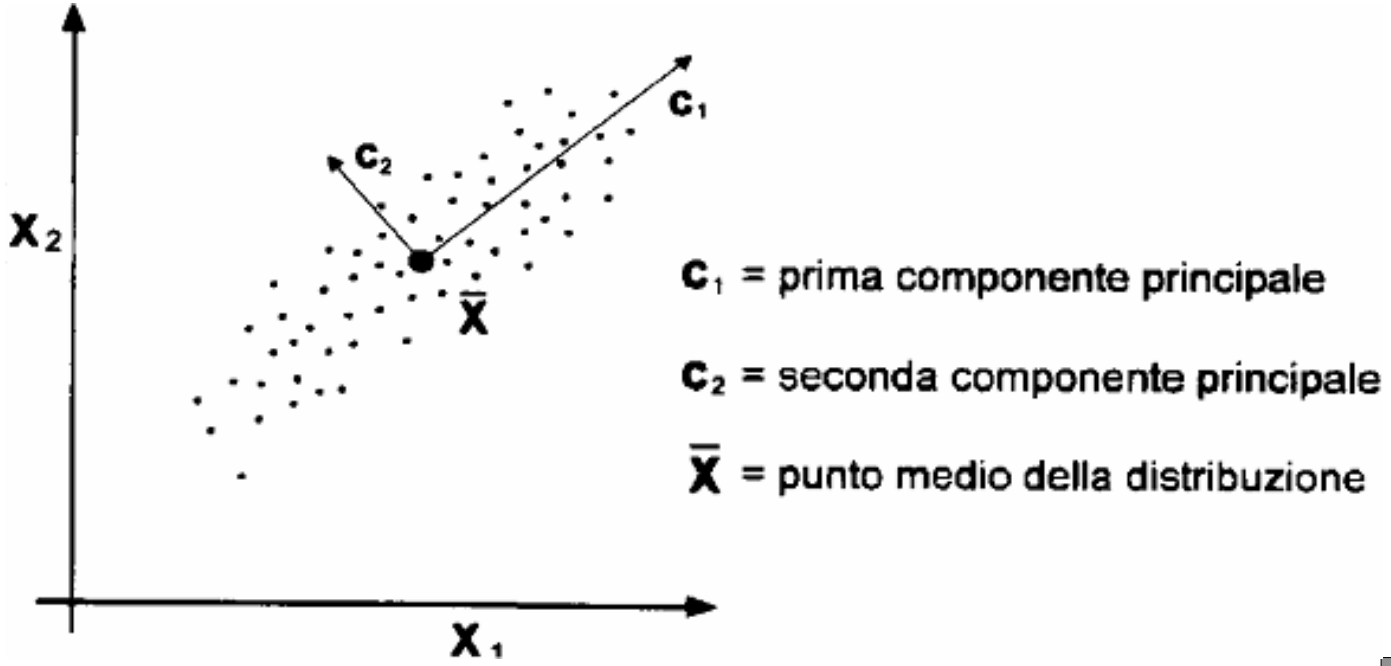
Proprietà feature per pattern recognition

- le feature che descrivono con numeri la lettera A devono assumere valori simili anche per forme della lettera A leggermente diverse ed in posizione diversa



Riduzione delle feature

Ridurre il numero delle feature velocizzare l'apprendimento. Una tecnica statistica efficace è la P.C.A.: analisi delle componenti principali dell'insieme dei valori delle feature. Per descrivere l'insieme nella figura bastano i valori delle prime due componenti.



Progettare la rete

Per un tipo di apprendimento supervisionato:

- individuare le classi in cui dividere l'input secondo il tipo di problema
- scegliere le feature analizzando matematicamente gli oggetti in input
- definire molte coppie (input, output) per i set di training (60%), validation (20%), test (20%)
- definire la codifica numerica: per l'input valori in [-1,1]; per l'output valori binari {0,1}
- scegliere alcuni criteri per valutare la qualità della risposta globale della rete sul test set

Progettare la rete

- scegliere un modello e definire l'architettura con:
  - funzione di attivazione per ogni neurone
  - numero di livelli hidden e numero di neuroni per ogni livello hidden numero di neuroni per lo strato input: tanti quanti i valori delle feature
  - numero di neuroni per lo strato output: tanti quante sono le classi
- scegliere valori piccoli dei pesi per favorire l'apprendimento
- scegliere un algoritmo di apprendimento e i suoi parametri di controllo (es. back propagation)
- scegliere una tecnica per controllare l'apprendimento

Usare la rete neurale

Dopo l'addestramento e la prova col test set, non si modificano più i pesi. Per usarla occorre:

- calcolare le feature del nuovo input di cui non si conosce l'output
- passarle in input alla rete neurale che le elabora calcolando le risposte dai neuroni dal livello input verso il livello output
- la risposta della rete va interpretata per decidere quale classe ha scelto; non si ha un neurone con valore 1 e tutti gli altri 0 per effetto delle funzioni di attivazione e di apprendimento: la classe assegnata è quella con valore di output più alto
- secondo il problema, se l'output è basso (es. 0,7 invece di 1), si può decidere accettarlo o di rifiutarlo facendo classificare ad un esperto

Testi consigliati

- Bishop, *Neural networks for pattern recognition*, Clarendon Press, Oxford, 1996
- Domeniconi, Jordan, *Discorsi sulle reti neurali e l'apprendimento,* Franco Angeli, 2001
- Cammarata, *Reti neuronali,* Etas
- Floreano, *Manuale sulle reti neurali*, Il Mulino
- *L'apprendimento delle reti artificiali di neuroni,* Le Scienze n.291, novembre 1992
- De Luca, Caianiello, *Introduzione alla cibernetica*, Franco Angeli
- ftp://ftp.sas.com/pub/neural/FAQ.html risposte a domande frequenti
- diwww.epfl.ch/mantra/tutorial/english/index.html un tutorial
- digitare "rete neurale" o "reti neurali" nei motori di ricerca

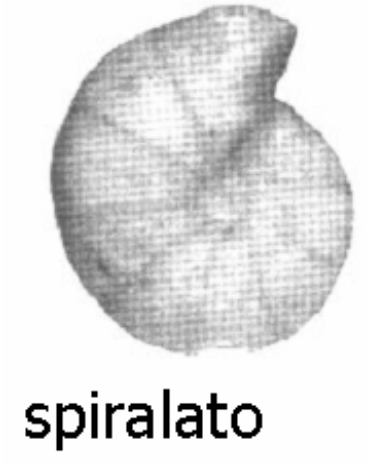
Strumenti software

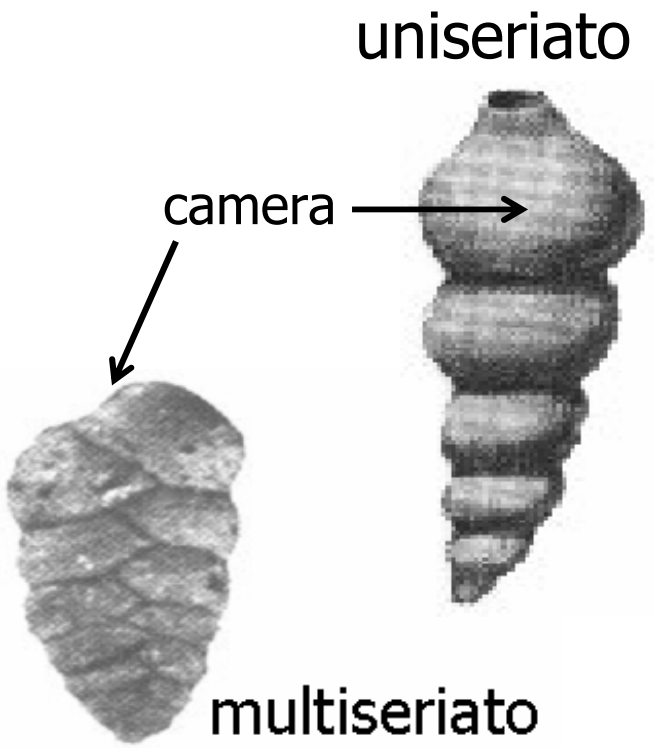
- Matlab Neural Toolbox, www.mathworks.com/products/neuralnet/
- Matlab NetLab, www.ncrg.aston.ac.uk/netlab/
- Java, esistono vari pacchetti gratuiti su Internet, link su queste pagine:
- www.geocities.com/fastiland/NNwww.html
- www.mathtools.net/Java/Neural_Networks/
- www-ra.informatik.uni-tuebingen.de/software/JavaNNS/welcome_e.html

Parte 5

Esempi di applicazione e commento di articolo scientifico in lingua inglese.

Esempio di classificazione: modi di accrescimento di camere in foraminiferi





Obiettivo della classificazione

Classificare il modo di accrescimento delle camere di un foraminifero intero setacciato da arenaria e fotografato da microscopio

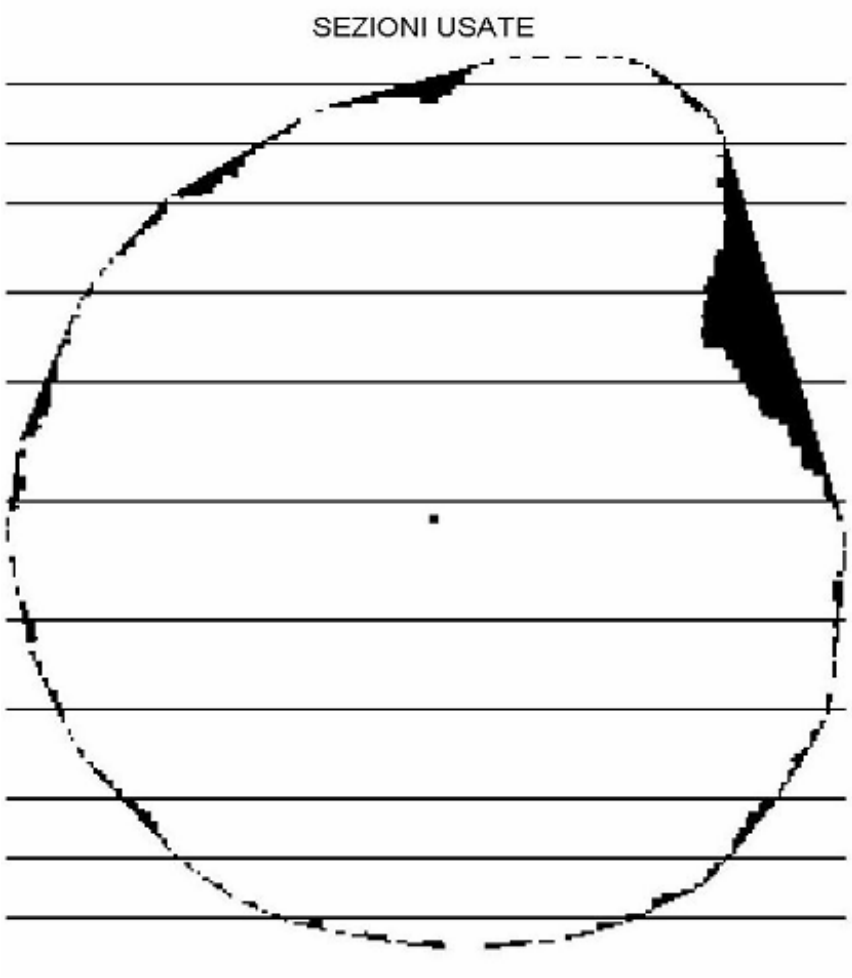
Si sceglie un apprendimento supervisionato e la Mlp perché si conoscono già le classi ed esistono numerosi esempi di classificazione presi da cataloghi di immagini.

I profili degli oggetti sono molto diversi quindi è inutile cercarne una descrizione matematica; non è possibile individuare le singole camere e contarle. Quindi si può usare:

-un campionamento non uniforme e adattivo dello spessore del guscio in varie parti

- feature calcolate su tutto il guscio

Feature scelte



In alto e in basso si misurano più spessori perché in tali zone si differenziano maggiormente gli spessori dei gusci. Altre feature:

- rapporto tra area di cerchio e area del fossile
- eccentricità del fossile (0=retta, 1=cerchio)
- spessore del guscio in pixel in 11 sezioni
- totale di 13 feature ridotte a 10

Rete neurale per classificare

- ogni valore di input normalizzato in [-1,1]
- ogni valore di output assume valore in [0,1] e somma 1

- rete Mlp: 10 nodi input , 8 nodi interni, 5 nodi output
- funzione attivazione neurone: softmax
- funzione di errore: cross-entropy
- funzione minimizzazione errore: quasi-newton
- apprendimento con riduzione di errore di generalizzazione tramite early stopping
- training set di 209 immagini
- validation set di 68 immagini

Risultati della classificazione

test set di 70 immagini, 1 classificato male, percentuale correttezza 98,57%

matrice di confusione per mostrare gli errori di classificazione

	1	2	3	4	5
1	17	0	0	0	0
2	0	17	1	0	0
3	0	0	15	0	0
4	0	0	0	4	0
5	0	0	0	0	16

riga=classe da rete
colonna=classe reale
1=agatostego
2=seriale
3=multiseriale
4=irregolare
5=spiralato

Sequenze di numeri

- Predizione: una serie di dati può essere data in input ad una rete neurale per fargli decidere quale sarà il prossimo valore, in base al modello statistico che la rete neurale si crea con i dati precedentemente usciti

- Interpolazione: un sensore fornisce un dato ma ogni tanto si guasta e non fornisce il dato; una rete neurale può prendere in input tutti i dati validi forniti dal sensore e ricavare i dati mancanti, in base al modello statistico che la rete neurale si crea con tutti i dati

Compressione dati

- uso di Multi Layer Perceptron
- N neuroni input e N neuroni output, gli N valori di input coincidono con gli N valori di output
- numero di pesi W molto minore di N, numero di neuroni nell'unico livello hidden è molto minore di N
- i pesi tra input e neuroni hidden sono uguali ai pesi tra neuroni hidden e neuroni output
- dopo l'apprendimento vengono conservati o trasmessi solo i valori dei pesi W invece dei N valori

Data Mining

Un processo di

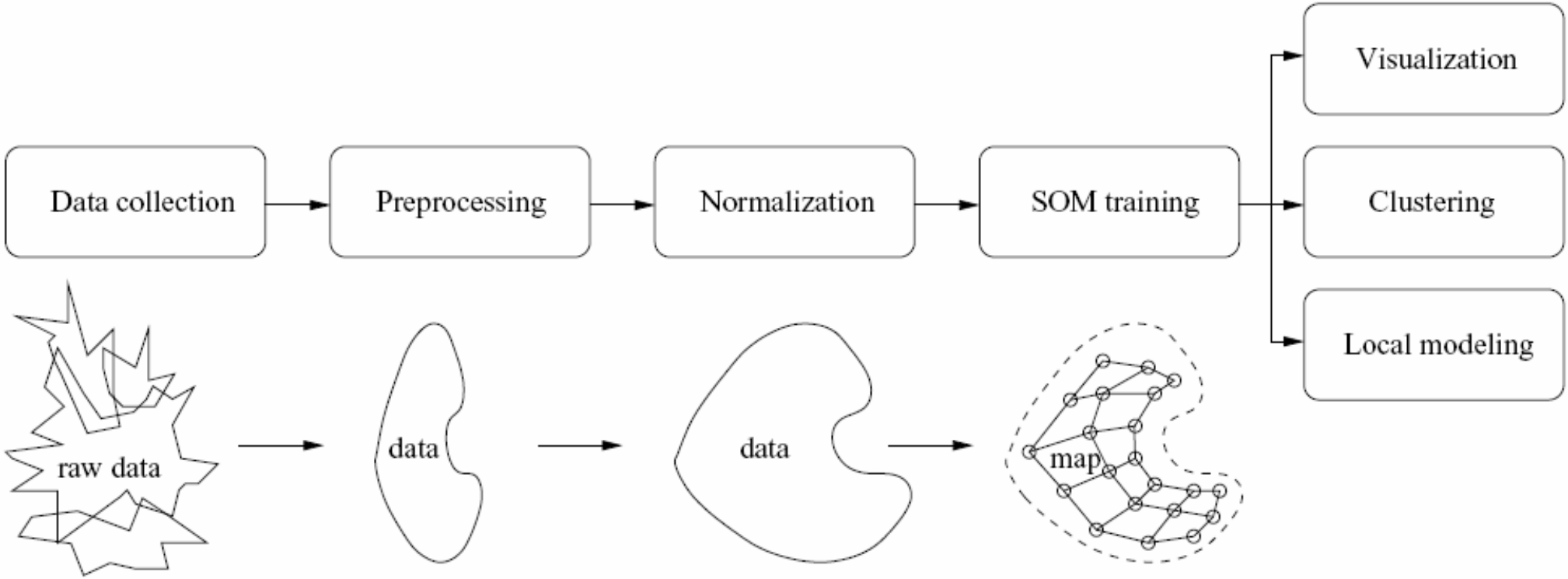
- selezione,
- esplorazione,
- modellazione

di grandi masse di dati per scoprire regolarità o relazioni non note a priori, allo scopo di ottenere un risultato chiaro e utile al proprietario del database.

Significa scavare nella miniera di tanti dati per trovare le informazioni utili.

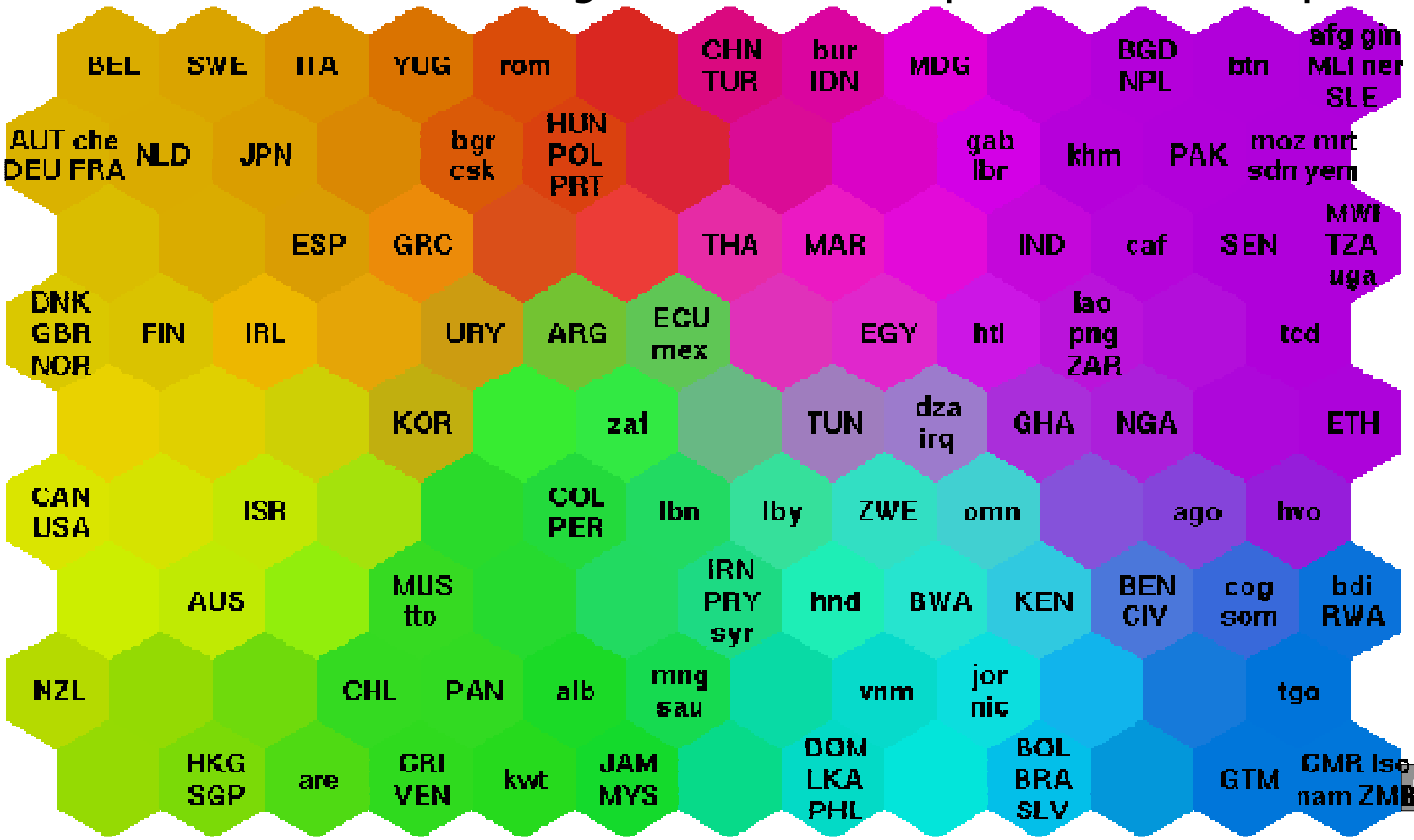
Esempio: supermercato ha database con acquisti dei clienti e vuole sapere cosa compra un tipo di cliente.

Data Mining: processo



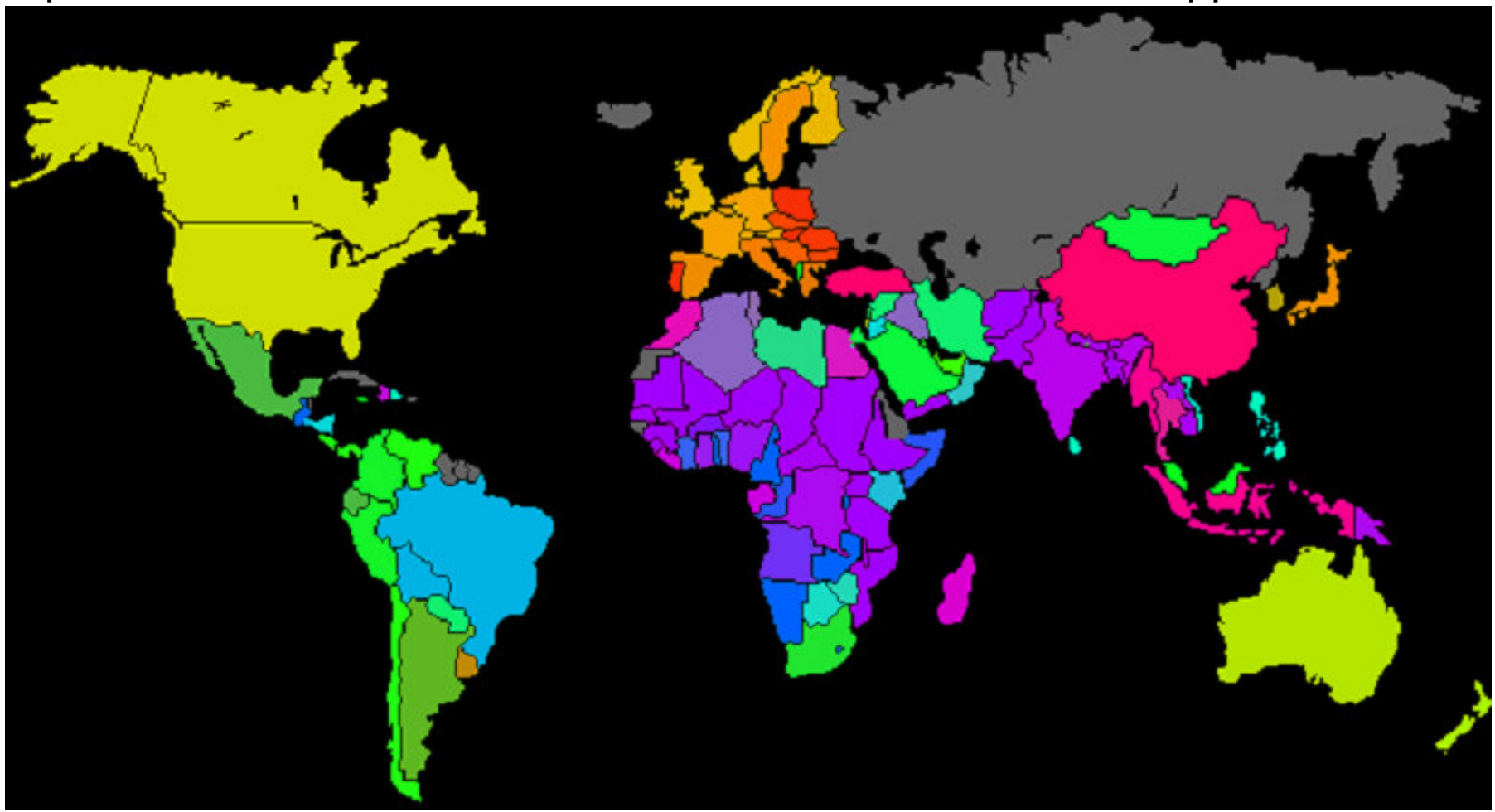
Esempio: Mappa mondiale della povertà

- Per ogni nazione del mondo c'è una sigla e 39 misure numeriche di: qualità della vita, nutrizione, educazione, reddito, ecc.
- La SOM ha analizzato questi dati e ha prodotto una matrice di neuroni output colorata in modo che ad ogni numero corrisponde un colore particolare.



Esempio: Mappa mondiale della povertà

- Si notano i cluster di paesi poveri molto vicini geograficamente. La contro prova si ottiene visualizzando i colori della SOM su una mappa.



Helsinki University Technology www.cis.hut.fi/research/som-research/worldmap.html