
Tecniche di Apprendimento Automatico in Applicazioni Bio-Genetiche

Marco Botta

Dipartimento di Informatica
Università di Torino

www.di.unito.it/~botta/didattica/masterbio.html

botta@di.unito.it

Slides by Botta

Sommario

- Caratterizzazione dei “problemi di apprendimento” riscontrabili in applicazioni Bio-Genetiche e Bio-Mediche.
- Panoramica sulle tecniche di apprendimento disponibili.
- Approccio integrato basato su strategie multiple.
- Alcuni esempi.

Slides by Botta

Tre Aspetti Fondamentali...

Rappresentazione

come presentiamo i dati a disposizione agli algoritmi di apprendimento

Definizione del problema di apprendimento

quale task dobbiamo risolvere

Approccio da utilizzare

quale algoritmo è più adatto al nostro problema

Slides by Botta

Rappresentazione: *il problema principale*

Rappresentazione attributo-valore

| D | Temp | Pres |
|----|------|------|
| 1 | 37 | 125 |
| 2 | 36 | 120 |
| 3 | 39 | 195 |
| 4 | 36 | 140 |
| 5 | 40 | 180 |
| 6 | 37 | 135 |
| 7 | 38 | 170 |
| 8 | 36 | 130 |
| 9 | 37 | 120 |
| 10 | 39 | 135 |
| 11 | 36 | 115 |

Vogliamo un programma che riconosca i casi "preoccupanti"

Rappresentazione con istanze multiple

| D | d | Temp | Pres |
|---|---|------|------|
| 1 | 1 | 37 | 125 |
| 1 | 2 | 36 | 120 |
| 1 | 3 | 37 | 118 |
| 2 | 1 | 38 | 130 |
| 2 | 2 | 37 | 130 |
| 2 | 3 | 39 | 170 |
| 2 | 4 | 38 | 140 |
| 2 | 5 | 37 | 135 |
| 3 | 1 | 36 | 115 |
| 3 | 2 | 37 | 120 |
| 4 | 1 | 36 | 118 |
| 4 | 2 | 37 | 120 |
| 4 | 3 | 39 | 190 |
| 4 | 4 | 40 | 180 |
| 4 | 5 | 36 | 115 |
| 4 | 6 | 37 | 118 |

Casi in cui esiste almeno un record "preoccupante"

Rappresentazione Strutturata

| D | d | Temp | Pres |
|---|---|------|------|
| 1 | 1 | 37 | 125 |
| 1 | 2 | 36 | 120 |
| 1 | 3 | 37 | 118 |
| 2 | 1 | 38 | 130 |
| 2 | 2 | 39 | 133 |
| 2 | 3 | 39 | 170 |
| 2 | 4 | 38 | 140 |
| 2 | 5 | 37 | 135 |
| 3 | 1 | 36 | 115 |
| 3 | 2 | 37 | 120 |
| 4 | 1 | 36 | 118 |
| 4 | 2 | 37 | 120 |
| 4 | 3 | 39 | 140 |
| 4 | 4 | 40 | 180 |
| 4 | 5 | 36 | 115 |
| 4 | 6 | 37 | 118 |

Casi in cui esistono record con temperatura simile e pressione molto diversa

Task di Data Mining

- ✧ Classificazione
- ✧ Dipendenze funzionali / Regressione
- ✧ Clustering / Segmentazione
- ✧ Riassunto / Caratterizzazione
- ✧ Scoperta di Associazioni / Causalità
- ✧ Individuazione di Anomalie
- ✧ Analisi di Serie Temporal

Slides by Botta

Classificazione e Regressione

Classificazione:
predire il valore di
un attributo categorico

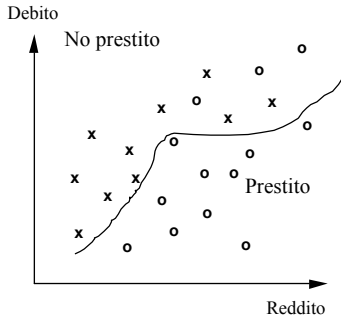
| D | Temp | Pres | C |
|----|------|------|---|
| 1 | 37 | 125 | N |
| 2 | 36 | 120 | N |
| 3 | 39 | 195 | P |
| 4 | 36 | 140 | N |
| 5 | 40 | 180 | P |
| 6 | 37 | 135 | N |
| 7 | 38 | 170 | P |
| 8 | 36 | 130 | N |
| 9 | 37 | 120 | N |
| 10 | 39 | 135 | N |
| 11 | 36 | 115 | N |

Regressione:
predire il valore di
un attributo numerico

| D | Temp | Pres | Fr |
|----|------|------|-----|
| 1 | 37 | 125 | 65 |
| 2 | 36 | 120 | 68 |
| 3 | 39 | 195 | 120 |
| 4 | 36 | 140 | 80 |
| 5 | 40 | 180 | 125 |
| 6 | 37 | 135 | 70 |
| 7 | 38 | 170 | 195 |
| 8 | 36 | 130 | 75 |
| 9 | 37 | 120 | 60 |
| 10 | 39 | 135 | 85 |
| 11 | 36 | 115 | 55 |

Slides by Botta

Classificazione



Problemi tipici affrontati

Individuazione di frodi
Concessione di crediti

Slides by Botta

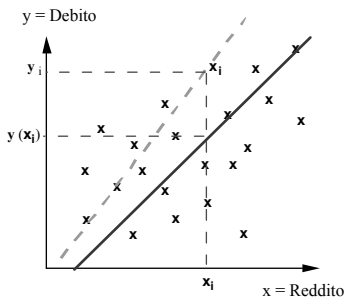
Dipendenze funzionali / Regressione

Individuazione di legami funzionali tra variabili che occorrono in una base di dati

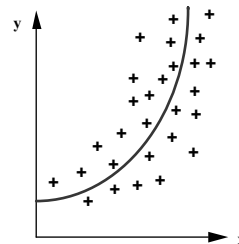
Dato un insieme $E = \{e_1, \dots, e_n\}$, di elementi descrivibili mediante i valori degli attributi $A = \{x_1, \dots, x_k\}$, il task di regressione assegna ad ogni elemento e_i dell'insieme E un valore di una variabile continua f

DM -> Inferisce una "funzione di regressione" direttamente da un sottoinsieme dei dati ("esempi di apprendimento")

$$\forall e_i : f = f(x_1, \dots, x_k)$$

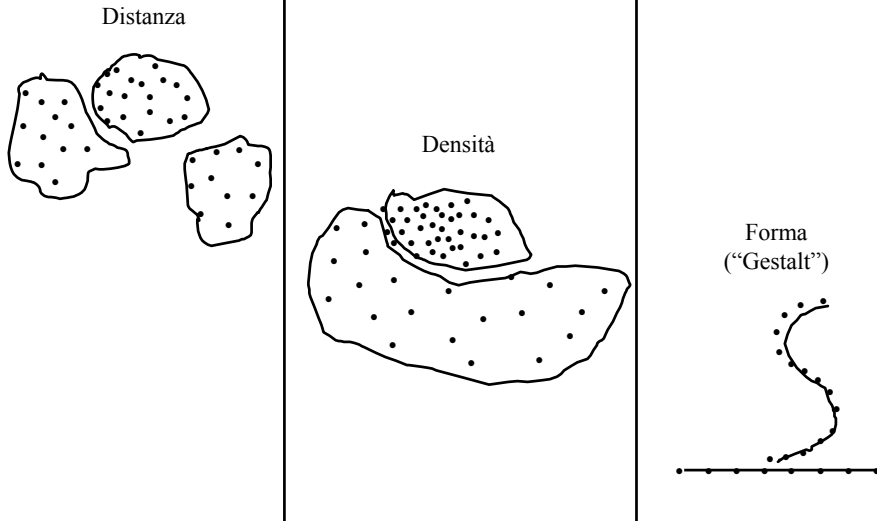


$$\sum_{i=1}^n [y_i - \hat{f}_p(x_i)]^2$$



Slides by Botta

Clustering



Segmentazione

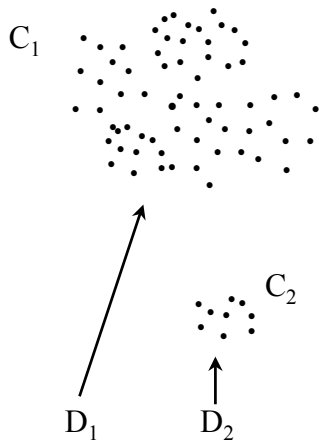
La segmentazione consiste nell'analizzare attuali o potenziali gruppi di clienti ("segmenti") per scoprirne caratteristiche e comportamenti che possano essere sfruttati nel mercato.

La segmentazione porta una organizzazione a vedere, al limite, ognuno dei suoi clienti come un "segmento unitario" ("segment of one"), al fine di stabilire con esso una relazione altamente personalizzata.

Due problemi basilari del marketing

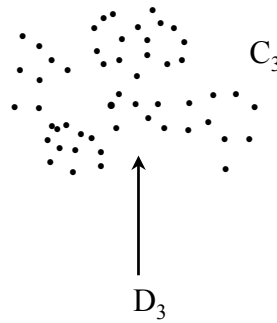
- ✦ Capire le cause dell'abbandono dei clienti ("customer attrition")
- ✦ Individuare nuove fette di mercato ("target marketing" e "cross selling")

Riassunto / Caratterizzazione



Perché sono stati raggruppati?

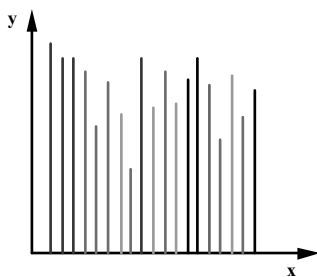
Che cosa hanno in comune?



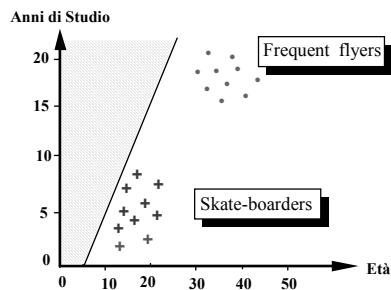
Slides by Botta

Riassunto / Caratterizzazione

Descrizione compatta di un insieme di dati



Media
Deviazione standard



- Persone di mezza età e istruzione universitaria
- + Ragazzi con basso livello di istruzione

Slides by Botta

Scoperta di Associazioni

Scoperta di associazioni tra fatti, proprietà o valori di variabili (“Link analysis”)

Il 72% degli acquirenti di insalata verde, acquista anche un condimento

Problema tipico

Market Basket Analysis

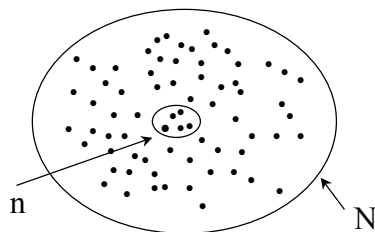


... .. {Pane, Pesche}

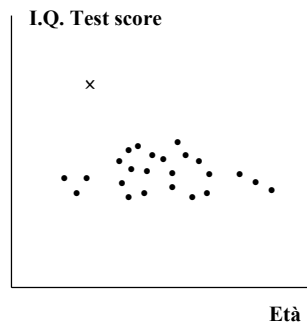
Slides by Botta

Individuazione di Eccezioni

Individuazione di valori devianti dai “normali”
(Eccezioni, Casi particolari, Errori)



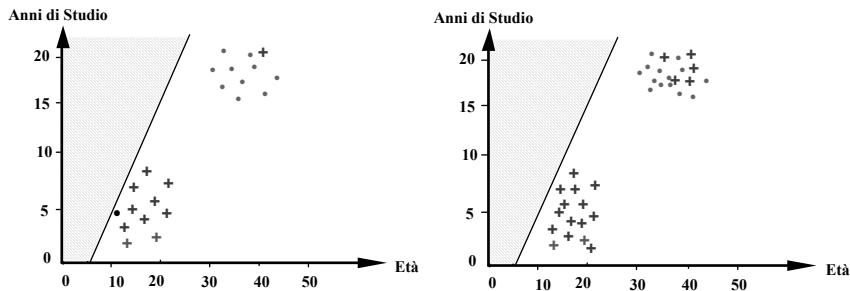
$n \ll N$



Slides by Botta

Individuazione di Anomalie

Individuazione di valori devianti dai “normali”

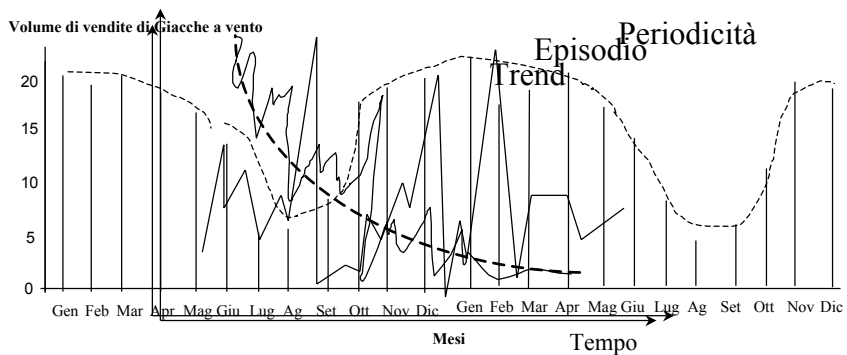


Occorre definire che cosa vuol dire “normale”

Slides by Botta

Analisi di Serie Temporali

- * Individuazione di conformazioni o episodi interessanti
- * Analisi di tendenze
- * Scoperta di periodicità o fenomeni “stagionali”



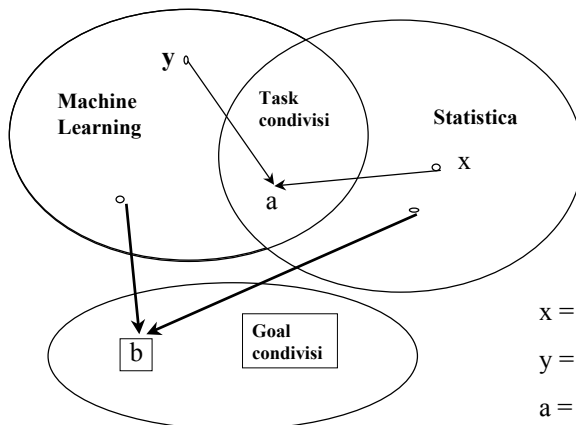
Slides by Botta

Discipline Rilevanti per il Data Mining

- ✳ Statistica
- ✳ Riconoscimento di Forme
- ✳ Intelligenza Artificiale
 - ✳ Apprendimento Automatico
 - ✳ Reti Bayesiane
 - ✳ Agenti Intelligenti
- ✳ Basi di Dati
 - ✳ Query and Reporting
 - ✳ “Data Warehousing” → OLAP
- ✳ Visualizzazione
 - ✳ Grafica
 - ✳ Ambienti multi-mediali
- ✳ Scienze Cognitive

Slides by Botta

Relazioni tra Discipline



x = Progetto di esperimenti

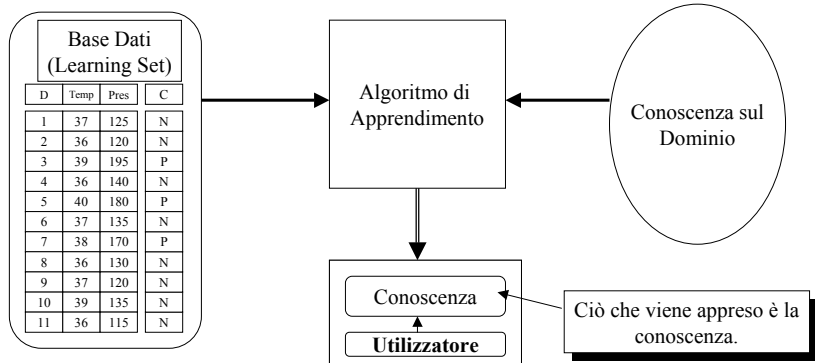
y = Apprendimento di piani

a = Stima dell'errore

b = Classificazione

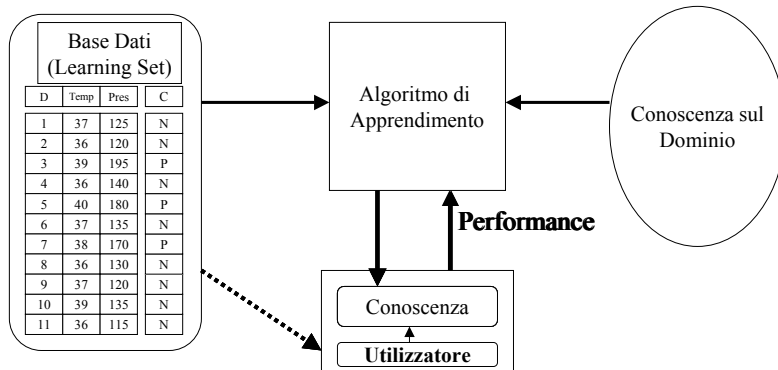
Slides by Botta

Apprendimento Supervisionato



Slides by Botta

Apprendimento come "Hypothesizing and Testing"

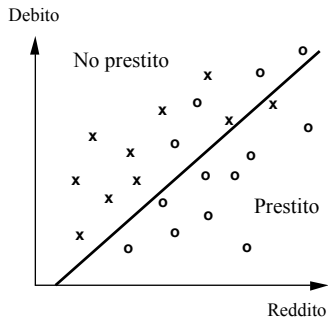


La possibilità di gestire rappresentazioni a istanze multiple o strutturate, dipende in gran parte dall'Utilizzatore

Slides by Botta

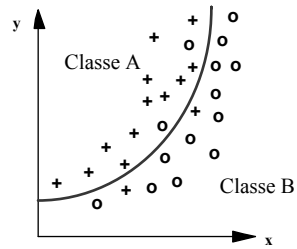
Analisi Statistica: Analisi Discriminante

Funzione discriminante



Lineare

$$\text{Prestito} : y - a x - b < 0$$



Non Lineare

$$\text{Classe A} : y - a x^2 - b > 0$$

Slides by Botta

Analisi Statistica: Clustering

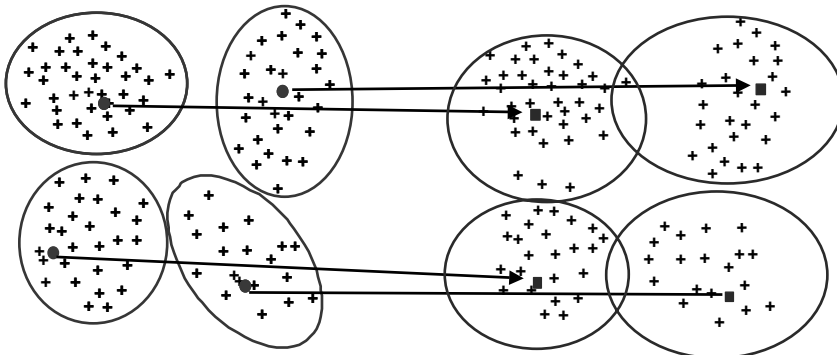
Algoritmo k-Means

Il numero K di cluster desiderato deve essere fornito dall'utente

Funzione distanza

Funzione obiettivo da ottimizzare :

Massimizza la distanza inter-cluster e minimizza la distanza intra-cluster



Slides by Botta

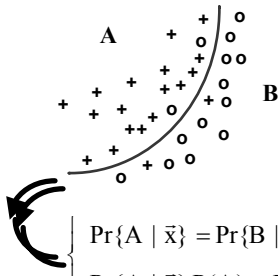
Riconoscimento di Forme: Classificazione

Approccio Statistico

Funzione di discriminazione probabilistica

Classificatore Bayesiano

Classificatore di Massima Verosimiglianza



$$\Pr\{A \mid \bar{x}\} = \frac{\Pr\{\bar{x} \mid A\} P(A)}{\Pr\{\bar{x} \mid A\} P(A) + \Pr\{\bar{x} \mid B\} P(B)}$$

$$\Pr\{B \mid \bar{x}\} = \frac{\Pr\{\bar{x} \mid B\} P(B)}{\Pr\{\bar{x} \mid A\} P(A) + \Pr\{\bar{x} \mid B\} P(B)}$$

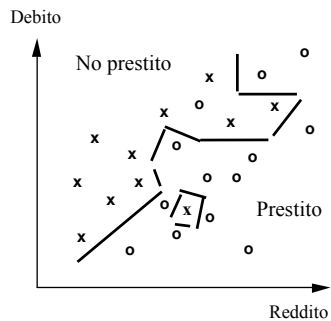
$$\left\{ \begin{array}{l} \Pr\{A \mid \bar{x}\} = \Pr\{B \mid \bar{x}\} \\ \Pr\{A \mid \bar{x}\} P(A) = \Pr\{B \mid \bar{x}\} P(B) \end{array} \right.$$

Slides by Botta

Riconoscimento di Forme: Classificazione

Approccio Basato su Casi "Case-Based"

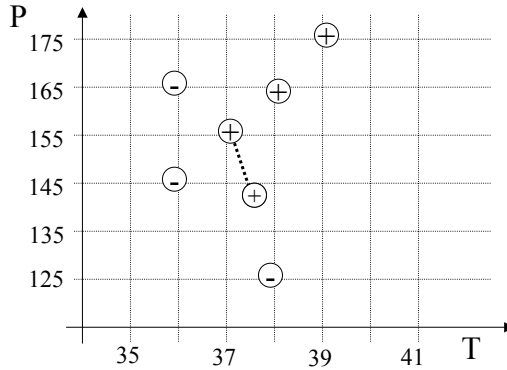
k-Nearest Neighbours



Slides by Botta

Instance Based Learning

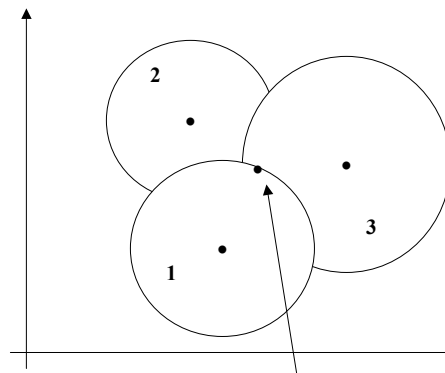
| | |
|---------------------------|---|
| T=37 ⁰ , P=155 | + |
| T=38 ⁰ , P=165 | + |
| T=36 ⁰ , P=145 | - |
| T=36 ⁰ , P=165 | - |
| T=39 ⁰ , P=175 | + |
| T=38 ⁰ , P=125 | - |



Quando si commette un errore si aggiunge un nuovo "caso" o se ne modifica uno esistente

Slides by Botta

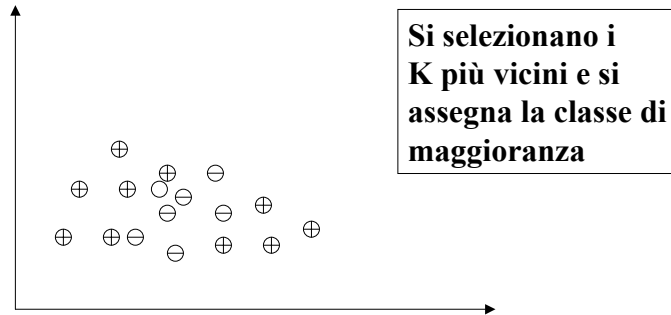
L'idea Base



Una nuova istanza da valutare subisce l'effetto dei campi cui è sottoposta e assume un'etichetta di conseguenza

Slides by Botta

k-NN: per classificazione



K = 1 => classe = +
K = 3 => classe = -

Se scopro che la classe assegnata è sbagliata memorizzo il nuovo caso classificato correttamente da un maestro

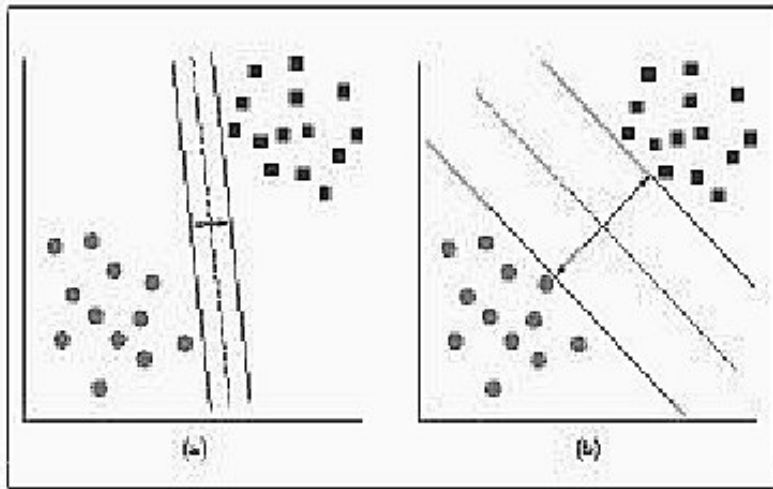
Slides by Botta

Support Vector Machine

- _ Dato un insieme di punti ciascuno appartenente a una di due classi, una SVM trova l'iperpiano che:
 - _ lascia la maggior parte dei punti di una stessa classe nello stesso semipiano
 - _ e massimizza la distanza dei punti delle due classi dall'iperpiano

Slides by Botta

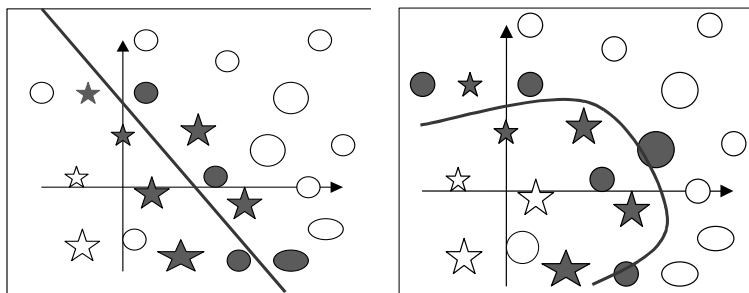
SVM: Idea di base



Slides by Botta

Support Vector Machine

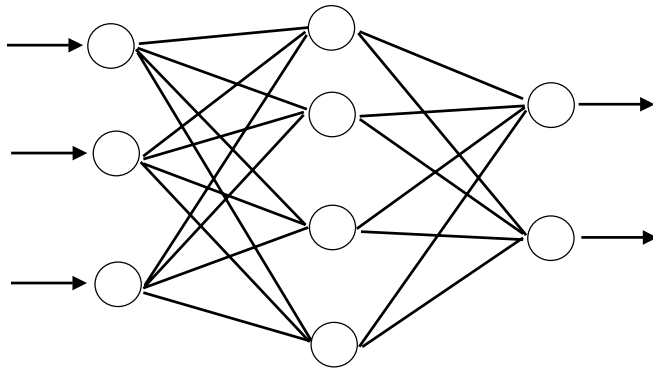
- _ I punti che stanno vicino all'iperpiano sono detti Support Vectors



Slides by Botta

Reti Neurali

Una rete neurale è una struttura composta, formata da elementi computazionali semplici, connessi secondo una topologia "a strati" => Approssimatori universali di funzioni



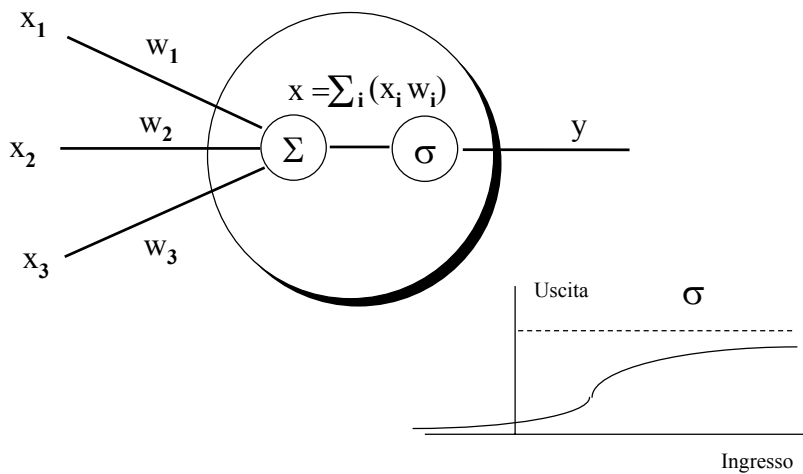
Unità di ingresso

Unità Nascoste

Unità di Uscita

Slides by Botta

Reti Neurali: Funzioni Elementari



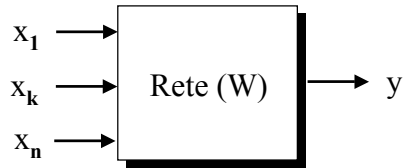
Slides by Botta

Reti Neurali: Addestramento

Algoritmo di “Backpropagation”

Minimizza l’errore quadratico totale

Se la rete è a più strati, l’errore viene propagato “indietro”



$$E = \frac{1}{2} \sum_{k=1}^n (t_k - y_k)^2$$

$$w_j = -\eta \frac{\partial E}{\partial w_j}$$

η = Velocità di apprendimento

Slides by Botta

Intelligenza Artificiale: Apprendimento Automatico Simbolico

- ✧ Alberi di Decisione
- ✧ Regole di Produzione
- ✧ Reti Bayesiane
- ✧ Gerarchie Concettuali

Slides by Botta

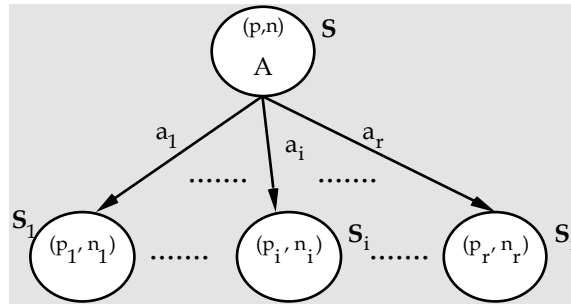
Alberi di Decisione

Date due classi P ed N

Dati p ed n esempi di apprendimento

Dato un insieme di Attributi A

Generare una partizione dello spazio dei possibili esempi, usando un criterio euristico di qualità



Slides by Botta

Alberi di Decisione : Esempio

Attributes

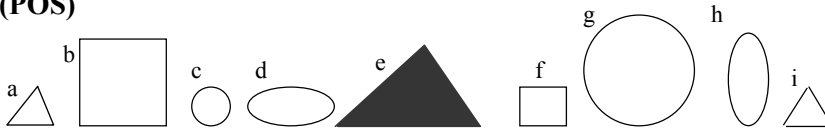
Color = {Red, Blue, Green, White}

Shaded = {Yes, No}

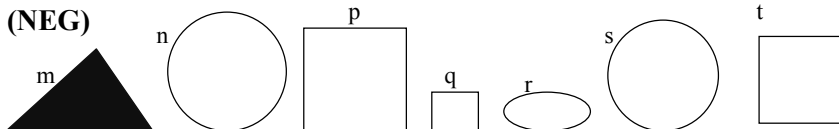
Shape = {Square, Triangle, Circle, Oval}

Size = {Small, Large}

(POS)

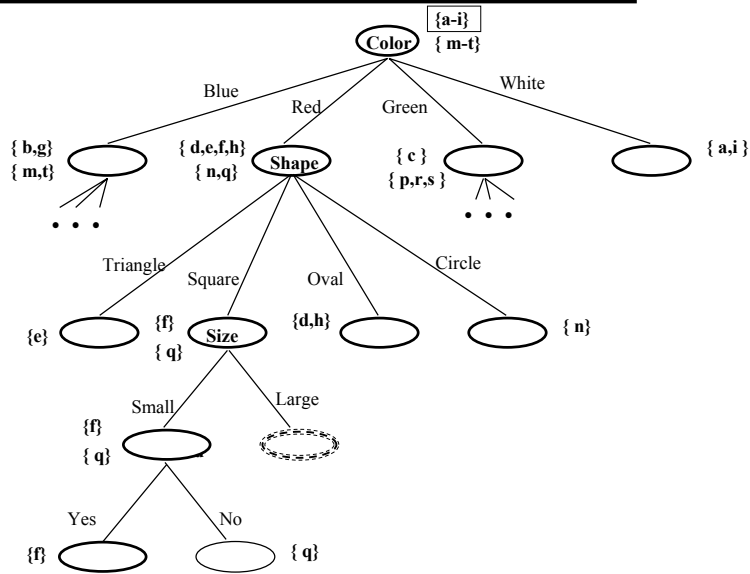


(NEG)



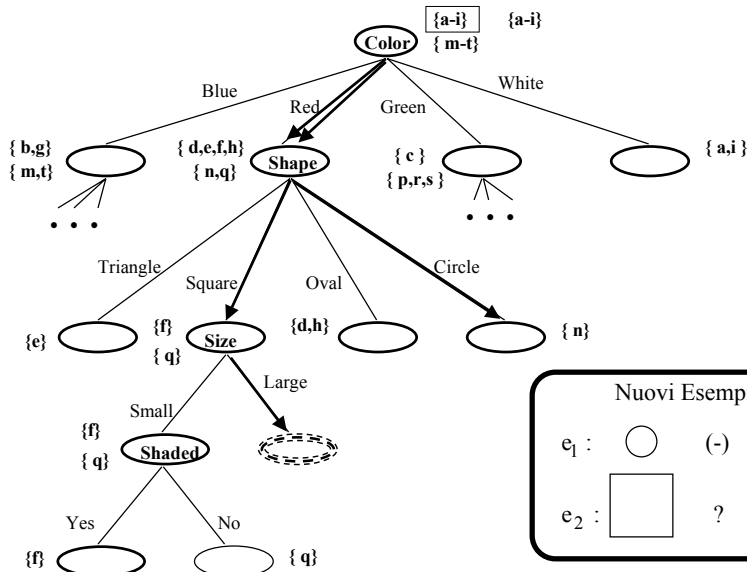
Slides by Botta

Alberi di Decisione: Apprendimento



Slides by Botta

Alberi di Decisione: Esempio (Classificazione)



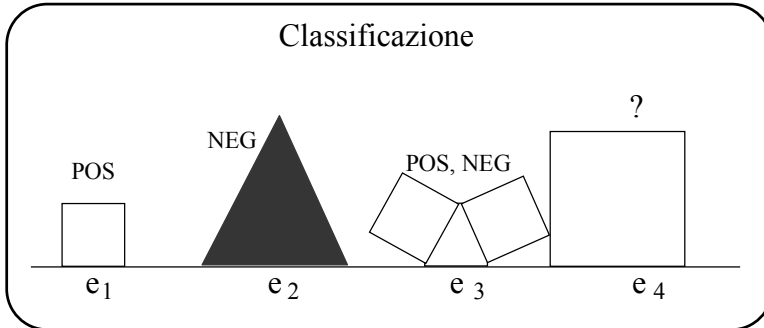
Slides by Botta

Regole di Produzione

Regole di decisione espresse in forma logica:
Calcolo Proposizionale o Calcolo dei Predicati

$(\text{forma} = \text{quadrato} \vee \text{triangolo}) \wedge (\text{dimensione} = \text{piccolo}) \Rightarrow \text{POS}$

$(\text{forma} = \text{triangolo}) \wedge (\text{tratteggiato} = \text{SI}) \Rightarrow \text{NEG}$

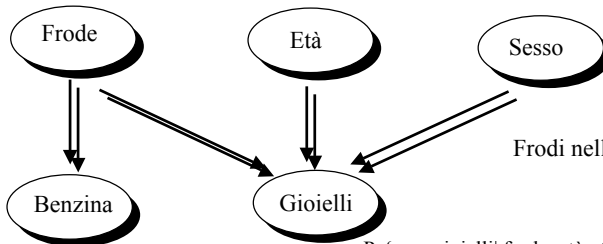


Slides by Botta

Reti Bayesiane

Rete Bayesianiana = Modello grafico di relazioni probabilistiche tra un insieme di variabili
Metodo per rappresentare l'incertezza nel ragionamento

$\Pr\{\text{frode}\} = 0.00001$ $\Pr\{\text{età} < 30\} = 0.25$ $\Pr\{\text{maschio}\} = 0.50$
 $\Pr\{30 < \text{età} < 50\} = 0.40$



Frodi nell'uso di Carte di Credito

$\Pr\{\text{acq. benzina} | \text{frode}\} = 0.2$
 $\Pr\{\text{acq. benzina} | \neg \text{frode}\} = 0.01$

$\Pr\{\text{acq. gioielli} | \text{frode}, \text{età} = *, \text{sesso} = *\} = 0.05$
 $\Pr\{\text{acq. gioielli} | \neg \text{frode}, \text{età} < 30, \text{maschio}\} = 0.0001$
 $\Pr\{\text{acq. gioielli} | \neg \text{frode}, 30 < \text{età} < 50, \text{maschio}\} = 0.0004$
 $\Pr\{\text{acq. gioielli} | \neg \text{frode}, \text{età} > 50, \text{maschio}\} = 0.0002$
 $\Pr\{\text{acq. gioielli} | \neg \text{frode}, \text{età} < 30, \text{femmina}\} = 0.0005$
 $\Pr\{\text{acq. gioielli} | \neg \text{frode}, 30 < \text{età} < 50, \text{femmina}\} = 0.0002$
 $\Pr\{\text{acq. gioielli} | \neg \text{frode}, \text{età} > 50, \text{femmina}\} = 0.001$

Slides by Botta

DM con le Reti Bayesiane

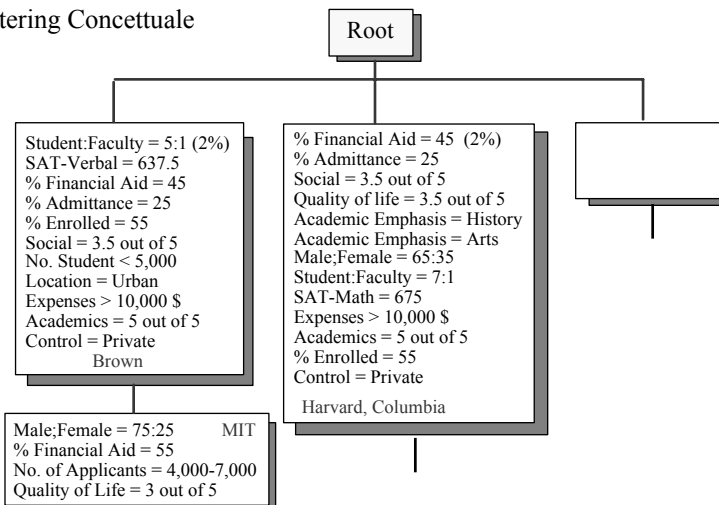
- Codifica della conoscenza dell'esperto mediante una Rete Bayesiana
- Uso della base di dati per aggiornare questa rete, eventualmente creandone di nuove
- Le probabilità si possono apprendere dai dati mediante statistica Bayesiana
- Uso delle reti ottenute simile a quello delle reti neurali

- Metodo robusto rispetto ad errori nella conoscenza iniziale
- Conoscenza interpretabile
- Utile per sfruttare conoscenza a priori

Slides by Botta

Gerarchie Concettuali

Clustering Concettuale



Slides by Botta

Algoritmi Genetici

- ✦ Gli Algoritmi Genetici sono un metodo generale di ricerca stocastica
- ✦ Essi si ispirano ai concetti dell'Evoluzione Darwiniana
- ✦ Possono essere usati nell'ambito di approcci sia simbolici che neurali

Ingredienti

- ✦ Popolazione di soluzioni (Cromosomi)
- ✦ Funzione di "Fitness"
- ✦ Operatori genetici ("Crossover" e Mutazione)

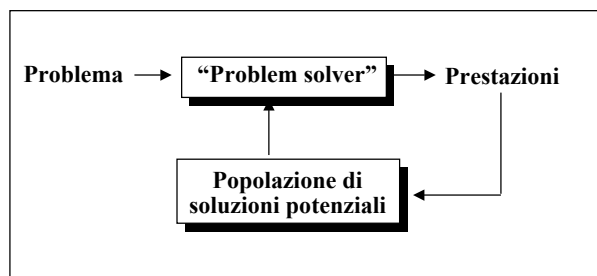
Ciclo di Base

- ✦ Si selezionano dalla popolazione, in numero proporzionale alla loro fitness, gli individui che devono riprodursi
- ✦ Gli individui selezionati si accoppiano e generano due figli, mediante l'applicazione dell'operatore di crossover
- ✦ Ai figli si applica l'operatore di mutazione
- ✦ La popolazione viene rinnovata

Slides by Botta

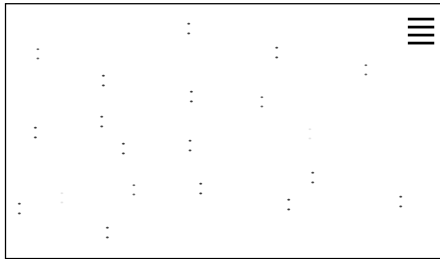
Algoritmi Genetici: Idea di base

La popolazione di potenziali soluzioni al problema migliora nelle generazioni successive

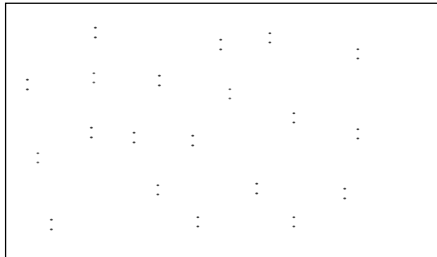


Slides by Botta

Algoritmi Genetici: Selezione

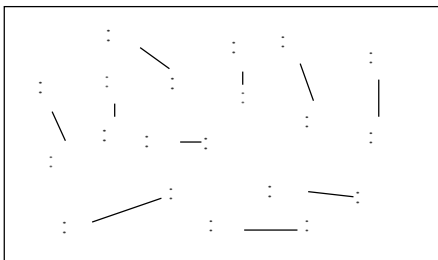


Selezione

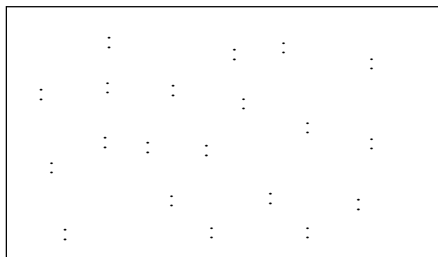


Slides by Botta

Algoritmi Genetici: Riproduzione

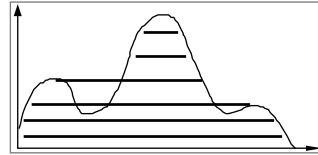
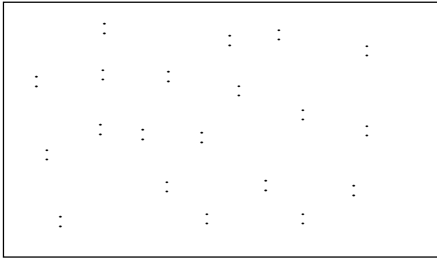


Riproduzione

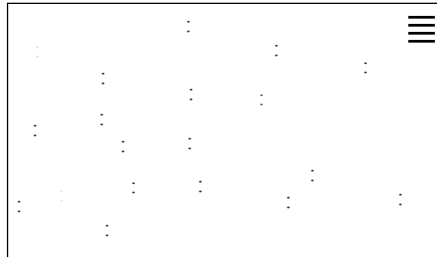


Slides by Botta

Algoritmi Genetici: Valutazione

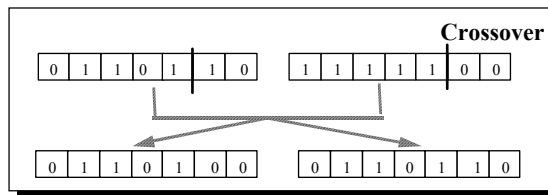


Valutazione

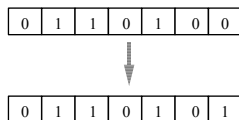


Slides by Botta

Algoritmi Genetici: Operatori Genetici



Mutation



Slides by Botta

Regole di Associazione

Sia I un insieme di items

Sia D un insieme di record, ognuno contenente un sottoinsieme di I

Regola di associazione:

$$r: X \Rightarrow Y$$

X e Y sono sottoinsiemi disgiunti di I

Supporto di un sottoinsieme Z di I: $\text{supp}(Z) = |D(Z)|/|D|$

Confidenza di una regola: $\text{conf}(r) = \text{supp}(X \text{ or } Y)/\text{supp}(X)$

Slides by Botta

Regole di Associazione: Algoritmo *Apriori*

- Algoritmo *Apriori*
 - Fase 1 \Rightarrow Ricerca di tutti gli insiemi frequenti
 - Costruzione incrementale a partire dalla cardinalità 1
 - Generazione dei candidati di cardinalità k a partire dagli insiemi frequenti di cardinalità (k-1)
 - Eliminazione dei candidati spuri
 - Fase 2 \Rightarrow Ricerca di tutte le regole possibili per ogni insieme frequente
- Ottimizzazione del metodo di calcolo del supporto
- Ricerca di regole ottimizzate rispetto la supporto o rispetto alla confidenza \Rightarrow Regioni rettilineari

Slides by Botta

Scoperta di Associazioni / Causalità

La scoperta di associazioni tra variabili è solo il primo passo di analisi. Occorre cercare una spiegazione

- Causalità tra A e B

Una variazione della variabile A “causa” una variazione della variabile B

A = Aumento di dipendenti “a tempo”

B = Aumento delle spese per stipendi

- Risposta comune

Una variazione delle variabili A e B è causata dalla variazione di una terza variabile C

A = Temporale

B = Abbassamento della colonna di mercurio del barometro

C = Arrivo di un’onda di bassa pressione

- Mascheramento

Una variazione della variabile B è causata sia da una variazione di A che da una variazione di una terza variabile C

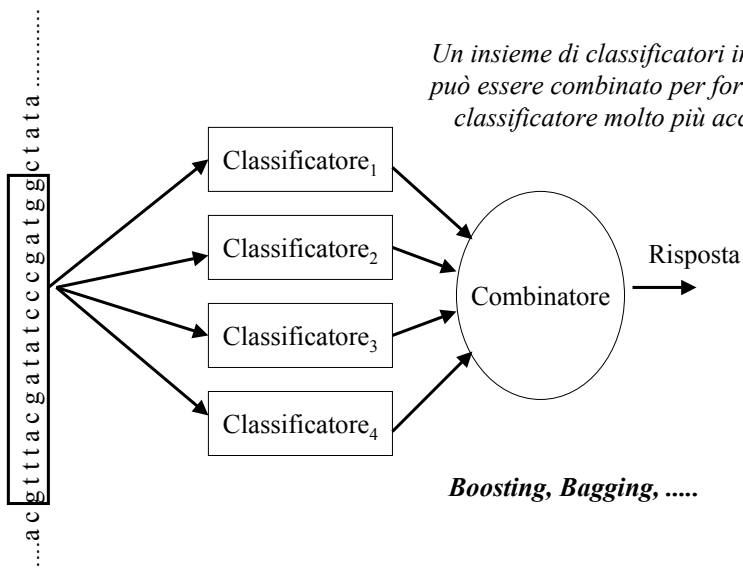
A = Spese per promuovere un prodotto

B = Ricavo dalle vendite

C = Numero di responsabili delle vendite che seguono i clienti

Slides by Botta

Classificatori Compositi



Slides by Botta

Un problema biologico affrontato in Machine Learning Apprendimento Splice-Junctions in sequenze DNA di primati

- _ Dati presi da Genbank 64.1 (ftp site: genbank.bio.net) (risalgono al 1992)
- _ 3190 sequenze nel dataset
- _ 3 categorie:
 - _ "ei" (767) e "ie" (768) includono ogni "split-gene" per i primati in Genbank 64.1
 - _ "n" (1655) non-splice presi da sequenze che non includono uno "splicing site"

Slides by Botta

Un problema biologico affrontato in Machine Learning Apprendimento Splice-Junctions in sequenze DNA di primati

- _ Problema di apprendimento: data una posizione nel centro di una finestra di 60 basi di DNA decidere se
 - _ a) giunzione "intron -> exon" (ie)
 - _ b) giunzione "exon -> intron" (ei)
 - _ c) neither (n)

Slides by Botta

Un problema biologico affrontato in Machine Learning

Apprendimento Splice-Junctions in sequenze DNA di primati

- _ Rappresentazione proposizionale con 62 attributi:
 - _ 1 la classe {n ei ie} della sequenza
 - _ 2 il nome della sequenza
 - _ 3-62 I rimanenti 60 attributi sono le basi della sequenza, in posizioni dalla -30 alla posizione +30 rispetto allo splice site.

ATRINS-DONOR-905,

A,G,A,C,C,C,G,C,C,G,G,G,A,G,G,C,G,G,A,G,G,A,C,C,T,G,C,A,G,G,G,T,G,A,G,C,C,C,A,C,C,G,C,
C,C,C,T,C,C,G,T,G,C,C,C,C,G,C, EI

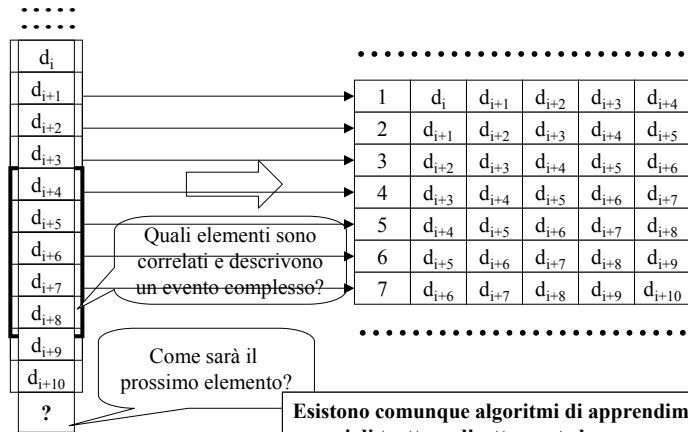
Slides by Botta

Distribuzione dei valori

| Basi | Neither | EI | IE |
|------|---------|---------|---------|
| A | 24.984% | 22.153% | 20.577% |
| G | 25.653% | 31.415% | 22.383% |
| T | 24.273% | 21.771% | 26.445% |
| C | 25.077% | 24.561% | 30.588% |
| D | 0.001% | -- | 0.002% |
| N | 0.010% | 0.010% | -- |
| S | -- | -- | 0.002% |
| R | -- | -- | 0.002% |

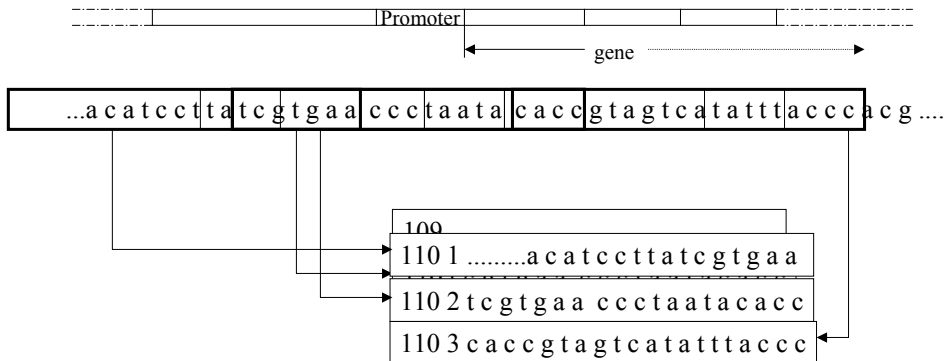
Slides by Botta

Predire in una Sequenza



Slides by Bottà

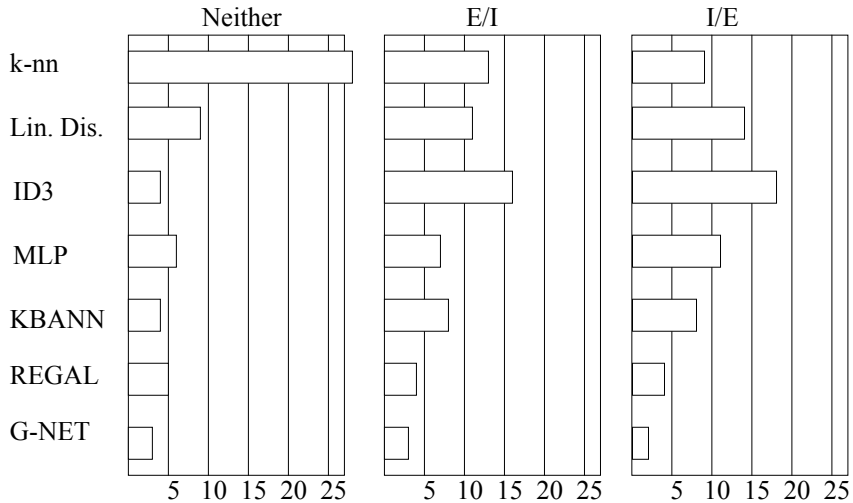
Esempi.....



Un ulteriore passo in avanti consiste nel considerare ogni possibile istanza di promoter come una sequenza, invece che un semplice vettore di attributi.

Slides by Bottà

Apprendimento Splice-Junctions con Rappresentazione Attributo-Valore



Su dataset di 3600 esempi

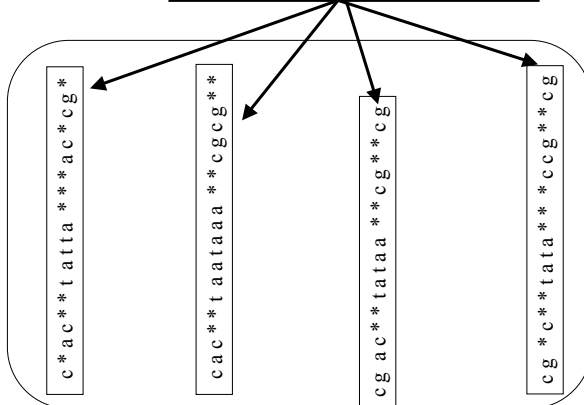
Su altri dataset presi dalla Gene-Bank....!?!

Slides by Botta

Un Problema di Apprendimento

Scegliere / Costruire
i proopti.....

.....aacttta**ccccataataaccg**tctcccacg.....

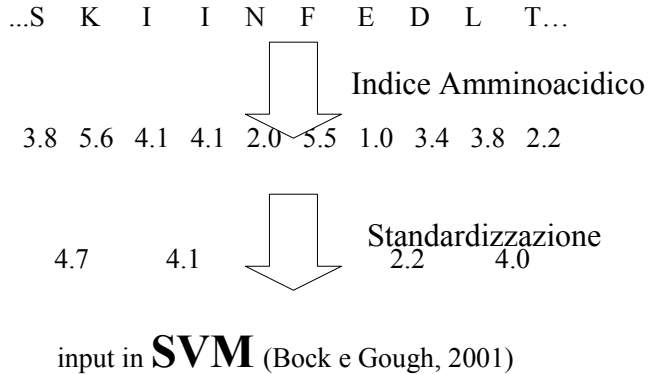


È una forma di
Instance-Based learning

Utilizzato con successo un Algoritmo Genetico

Slides by Botta

Predizione di interazioni dalla sequenza primaria con SVM



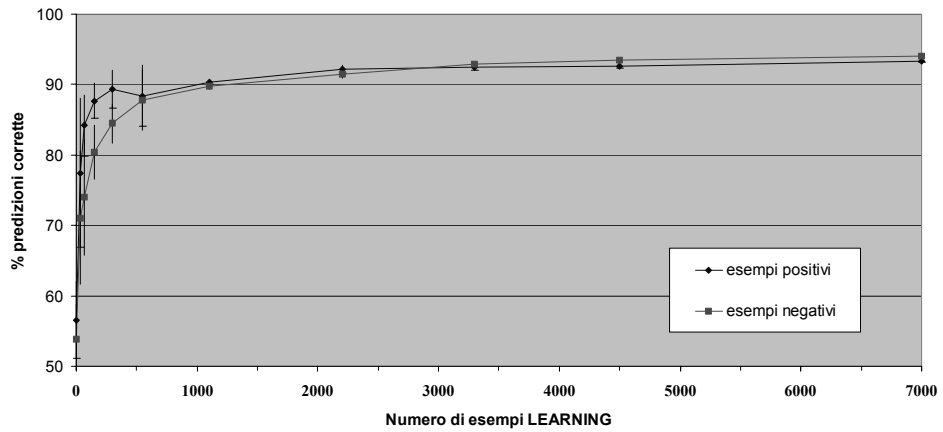
Slides by Botta

Database di interazioni proteiche

- _ BIND classifica interazioni tra proteine ed altre proteine, acidi nucleici, molecole semplici o fotoni
- _ MINT raccoglie essenzialmente interazioni tra proteine, anche se rimane aperto a tutti i tipi di interazione
- _ DIP è il più ricco database di interazioni tra proteine (oltre 13500 al 28/6/2002) ed è in continua e rapida espansione

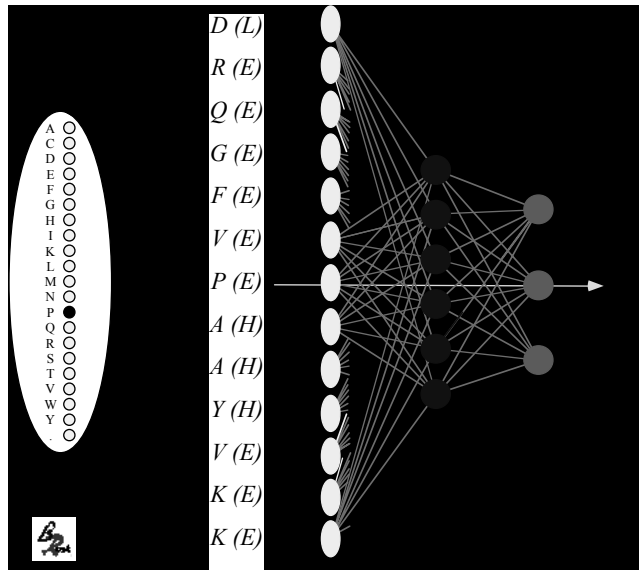
Slides by Botta

Robustezza di SVM



Slides by Botta

Neural Network for secondary structure



Slides by Botta

Software Disponibile

- _ Apprendimento simbolico: Weka Suite
 - _ <http://www.cs.waikato.ac.nz/~ml/weka/>
- _ Neural Networks:
 - _ <http://www.emsl.pnl.gov:2080/proj/neuron/neural/systems/shareware.html>
- _ Algoritmi genetici
 - _ G-net:
 - _ <http://hermes.mfn.unipmn.it/~attilio/PROJECTS/GNET/gnet.html>

Slides by Bottà

Conclusioni

Il Machine Learning come punto di incontro e di integrazione di approcci diversi nati in seno a discipline diverse.

L'apprendimento da dati strutturati e da sequenze è una problematica emergente che risulta essere cruciale per applicazioni avanzate nel settore bio-medico.

Due fattori fondamentali per applicazioni di successo:

- disporre di una equipe che abbia sia competenze informatiche che conoscenze relative al dominio dell'applicazione.
- sapere integrare metodologie diverse in programmi diversi.

Slides by Bottà