

Laurea Specialistica in Informatica
a.a. 2005-2006

Interfacce Intelligenti

Fiorella de Rosis

Introduzione

Prima parte: Formalizzazione e Ragionamento

- 1.1. Ragionamento logico:
 - Formalizzazione
 - Risoluzione
- 1.2. *Ragionamento incerto*
 - Reti Causali Probabilistiche
 - Reti dinamiche
 - *Apprendimento di Reti*

Programma
del Corso

Seconda parte: Modelli di Utente

- 2.1. Modelli logici
- 2.2. Modelli con incertezza

Terza parte: Interazione in linguaggio naturale

- 3.1. Generazione di messaggi
 - Introduzione
 - Teorie
 - Metodi

- 3.2. Comprensione di messaggi

Quarta parte: Simulazione di dialoghi

Costruzione di Reti Causali Probabilistiche

Passi necessari:

- a. Definizione della struttura
- b. Assegnazione dei parametri
- c. Test
- d. Eventuale revisione iterativa del modello

Costruzione di Reti Causali Probabilistiche

Metodi possibili:

- a. Costruzione 'manuale' del modello (struttura e parametri), a partire da conoscenza e ipotesi sul dominio
- b. Apprendimento (dei parametri e/o della struttura) da un insieme di dati.
- c. Combinazione delle due procedure

Nel primo caso, la definizione dei parametri è basata sull'idea soggettivista di probabilità (degree of belief).

Nel secondo caso, è basata sulla sull'idea frequentista.

Nel terzo caso, su un'integrazione dei due concetti.

In ogni caso, il test del modello e la sua revisione iterativa sono passi necessari!

Un Esempio: Es 2.4

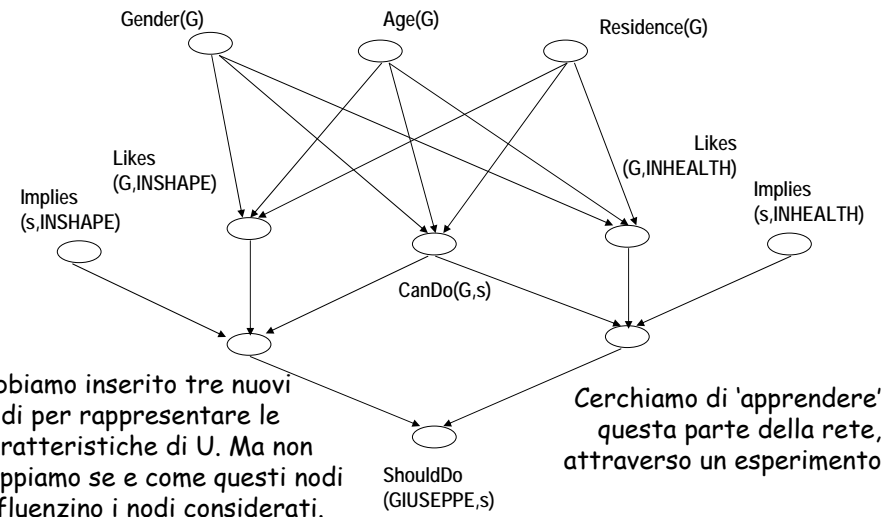
Riprendiamo l'esempio della persuasione.

Ma integriamo la RCP in modo da inserire alcune caratteristiche dell'utente che permettano di scegliere la strategia più promettente (agire sul goal di 'essere in forma' oppure su quello di 'essere in buona salute'?).

Per far questo, inserisco nuove informazioni sull'utente, che permettano al sistema di inferire:

- se è più interessato alla forma fisica o all'essere in buona salute
- e
- se è o no in grado di fare attività fisica.

Un Esempio: Es 2.4



Apprendimento da un insieme di Dati: Passi necessari

- Piano della raccolta dati (esperimento):
 - definizione della popolazione su cui effettuare l'esperimento
 - definizione delle variabili da 'osservare'
 - definizione dei metodi di codifica (var continue vs discrete e valori possibili)
- Codifica dei dati (foglio Excel, file di testo,...) e verifica degli errori
- Discretizzazione delle variabili continue
- Calcolo delle prob a priori e condizionate
- Valutazione dell'*attendibilità* del modello
- Test preliminare del modello
- Eventuale revisione iterativa della struttura e dei parametri
- Test finale del modello in situazioni 'significative'

Apprendimento di RCP

Esperimento:

analisi degli atteggiamenti degli studenti rispetto allo sport

Popolazione coinvolta nell'esperimento: gli studenti del Corso di IntInt

(circa 100 studenti, omogenei per livello culturale, abbastanza equidistribuiti per sesso ed età, diversi per provenienza: Bari-non Bari, personalità, ecc);

Variabili da osservare:

- variabili che definiscono il 'soggetto': età, sesso, luogo di residenza,
- 'scala di valori'
- interesse nei confronti dello sport
- sport eventualmente praticati

Il questionario

Quanti anni hai? __

Qual è il tuo sesso? M __ F __

Dove risiedi? A Bari __

A meno di 100 km da Bari __

A più di 100 km da Bari __

Credi che essere in buona forma fisica sia importante?

molto __ abbastanza __ poco __

Credi che la salute sia un bene fondamentale della vita?

molto __ abbastanza __ poco __

Pratichi qualche sport?

con regolarità __ raramente __ mai __

Se sì, quale?

corsa__ bicicletta __ palestra __ nuoto__

Se raramente o mai, perché?

non m'interessa __ non ho tempo __ non ho soldi __

.....

Somministrazione ad un campione rappresentativo
(quindi, non 'selezionato'...)

Trascrizione dei risultati in un database

Età	Sesso	Resid	Forma fisica	Salute	Sport praticato	Perché no
22	M	B	Y	N	PI	T
25	F	P	N	N	PA	S

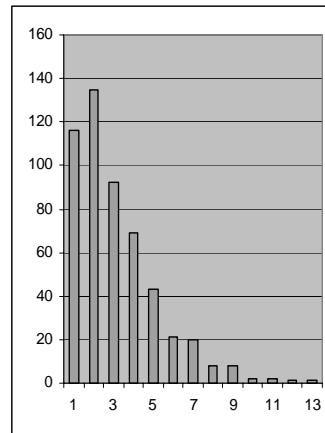
Ma potremmo anche definire le variabili 'interesse per la forma fisica' e 'interesse per la salute' come variabili a più valori, ad esempio inserendo nel questionario una 'scala di Likert'.

Discretizzazione delle variabili continue
(es età):

a. per intervalli di eguale ampiezza ('range');

ma: in una variabile con distribuzione asimmetrica dei valori, una discretizzazione per 'range' produrrebbe alcuni 'dati sparsi', cioè alcuni valori con bassa frequenza

(e vedremo che questo crea problemi nella elaborazione dei dati)



Discretizzazione delle variabili continue:

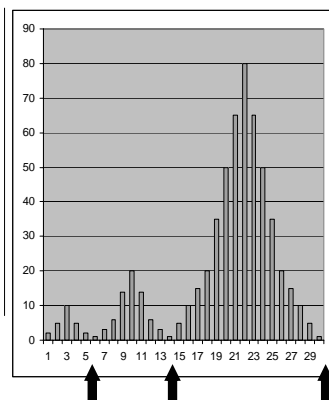
b. per intervalli con eguale numero di casi ('frequency').

Nello stesso esempio precedente, una distribuzione per frequenza, con lo stesso numero di intervalli (ma di ampiezza crescente) produrrebbe una distribuzione 'uniforme'.

I due metodi presentano vantaggi diversi

Discretizzazione delle variabili continue:

Consideriamo la distribuzione seguente:



che potrebbe rappresentare, ad esempio, la distribuzione per età di una popolazione con tre classi di soggetti. Qui, può essere conveniente una discretizzazione in tre valori che rispetti la distribuzione in classi

Calcolo delle Probabilità (a priori e condizionate) con il metodo 'frequentista'

Consideriamo il caso di due variabili, V_h, V_k

la prima a due e la seconda a tre valori:

	V_k^1	V_k^2	V_k^3
V_h^1	a	b	c
V_h^2	d	e	f

$$P(V_k^1) = (a+d)$$

... ecc

$$P(V_h^1|V_k^1) = P(V_h^1 \wedge V_k^1) / P(V_k^1) = a / (a+d)$$

... ecc

$$P(V_h^1) = (a+b+c)$$

... ecc

Notare che: quanto maggiore è il numero di variabili messe in relazione fra loro (e/o il numero di valori che ciascuna di esse può assumere), tanto più complessa è la tabella delle probabilità condizionate.

Problema dei dati 'sparsi' e ambiguità del valore .5 di probabilità

Immaginiamo di aver codificato le età in tre fasce, e di voler calcolare la probabilità del sesso (=M/F) condizionata all'età.

Supponiamo di aver definito le fasce di età in modo sbagliato.

	E1	E2	E3	E4
M	0	11	21	18
F	0	19	21	10

$P(M|E1) = 0/0$ e viene posta, per convenzione, = .5

Ma anche $P(M/E3) = .5$

Una prob = .5 può significare 'incertezza' (su valori osservati) o 'ignoranza' (su valori non osservati)

Apprendimento di RCP

Il problema dell'apprendimento del BN:

Se il database contiene n variabili, un belief network che corrisponde ad un grafo 'completo' (con n nodi) conterrà

$$n*(n-1)/2 \text{ archi (con orientamento da determinare).}$$

Per semplificare tutti gli algoritmi di gestione del BN, è necessario determinare *il BN con il numero minimo di archi che rappresenta in modo 'accettabile' il dominio considerato.*

Apprendimento di RCP

Il problema dell'apprendimento:

Date:

n variabili, rappresentate ciascuna in un nodo di un grafo orientato G;

un dataset $D = \{ \langle v_1^1, \dots, v_n^1 \rangle, \dots, \langle v_1^h, \dots, v_n^h \rangle \}$ relativo all'osservazione dei valori delle n variabili su un numero h di 'soggetti'.

Trovare il modello M che meglio approssima D, e cioè che massimizza la $P(D | M)$.

Individuare il modello corrisponde a individuare:

- quali archi legano i nodi in G (cioè, la sua *struttura*) e
- quali *parametri* sono associati ai nodi di G (probabilità a priori delle radici, e condizionate per gli altri nodi).

Apprendimento di RCP

Una misura di qualità del modello M:

La *probabilità marginale* (PM) di una RCP è il prodotto delle probabilità marginali dei suoi nodi.

Marginal Log Likelihood di un BN:

$$P(D | M) = \prod_k P(v^k | M)$$

Gli algoritmi di apprendimento si propongono di trovare la MLL minima, dati:

- un dominio descritto da un insieme di variabili e
- un set di dati che descrive il comportamento di queste variabili in un numero 'appropriato' di soggetti.

Apprendimento di RCP

Ipotesi di approssimazione:

Le osservazioni sono distribuite in modo indipendente:

$$P(D | M) =$$

$$\prod_k \prod_h P(V_h = v_h^k | \text{Genitori}(V_h, M))$$

... i prodotti diventano somme se si passa ai logaritmi; questo rende le elaborazioni più semplici e aiuta a superare il problema dei numeri molto piccoli.

$$MLL(M) = \sum_k \sum_h \log P(V_h = v_h^k | \text{Genitori}(V_h, M))$$

La MLL di una RCP è quindi uguale alla somma delle MLL dei suoi nodi.

Dipende dal numero di nodi nella RCP e (in misura molto minore) dalla sua struttura.

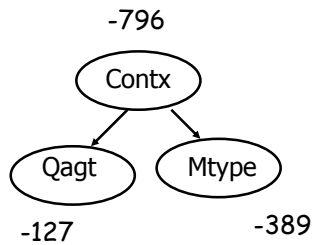
Apprendimento di RCP

Outline dell'algoritmo:

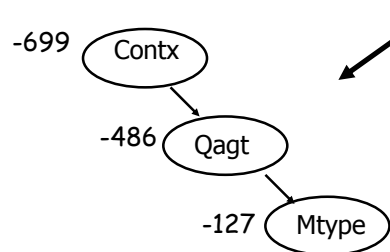
- \forall nodo della rete, si calcolano le correlazioni con gli altri nodi (candidati genitori), fino ad un numero massimo di genitori prefissato;
- per ogni combinazione (nodo considerato, genitori candidati), si calcola la MLL del nodo;
- si ordinano le combinazioni per valori decrescenti della MLL;
- si costruisce iterativamente la rete collegando i nodi con i genitori individuati;
- si calcola la MLL della rete.



Un esempio



MLL(RCP) = -1313
in entrambi i casi



La MLL è una misura importante per *confrontare modelli alternativi* sullo stesso set di dati, più che per valutare singoli modelli

Alternative nell' Apprendimento

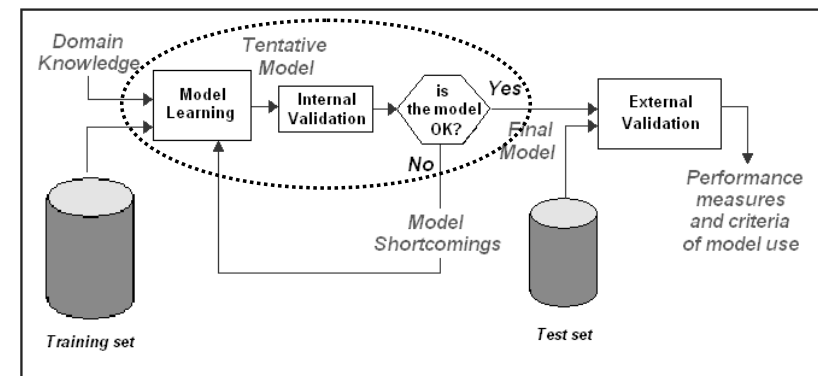
Possiamo affidare l'intero compito di definizione della struttura e apprendimento dei parametri all'algoritmo

Oppure

Possiamo adottare una strategia 'mista':
Definizione di 'vincoli' iniziali,
Completamento della struttura con algoritmo di apprendimento,
Revisione iterativa della struttura e dei parametri.

Vediamo come....

Outline del processo iterativo di apprendimento



Apprendimento di RCP

L'algoritmo non tiene conto del significato delle variabili né dell'uso che verrà fatto del modello.

Quindi un modello 'ottimale' in termini di MLL può non essere *interpretabile*.

Questo limite implica che un apprendimento totalmente automatico del modello produce, in genere, risultati insoddisfacenti

e che la strategia da adottare per la costruzione di un buon modello debba essere basata, piuttosto, su una *integrazione di procedura manuale e automatica*, e su un *processo iterativo di costruzione del modello* (per 'raffinamenti successivi').

Apprendimento di RCP

La MLL di una RCP è funzione della complessità della rete (numero di nodi e di archi).

Il confronto fra due RCP apprese deve quindi tener conto:

- della loro complessità e, a parità di complessità,
- della loro plausibilità teorica e
- dei vantaggi ai fini dell'uso previsto.

Apprendimento di RCP

Riprendiamo il nostro esempio.

Età	Sesso	Resid	Forma fisica	Salute	Pratichi sport	Quale	Perché no	...
22	M	B	Y	N	Y	PI	T	...
25	F	P	N	N	Y	PA	S	...

Numero possibile di nodi nel BN: 8

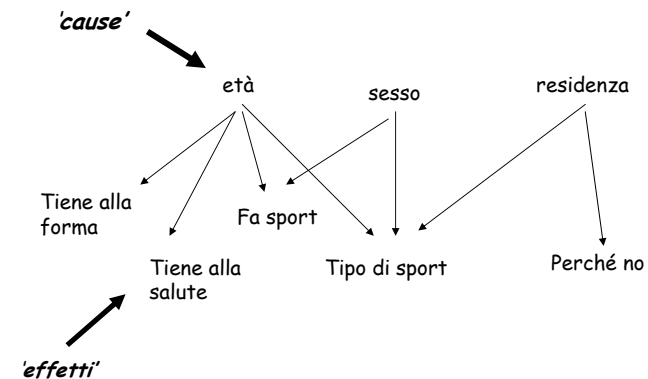
N massimo possibile di archi: 28

**Ma esiste una struttura 'teorica',
che tenga conto delle ipotesi sul dominio?**

Apprendimento di RCP

Possiamo semplicemente distinguere le variabili in due grandi categorie: 'cause' e 'effetti'.

Oppure, possiamo ipotizzare legami ben precisi fra alcune variabili.



Apprendimento di RCP

1 Caso: apprendimento (*quasi*) completamente automatico

Si può isolare un *sottoinsieme delle variabili*, su cui effettuare l'analisi (ad esempio, eliminando l'identificatore del soggetto).

Si possono *imporre vincoli* su:

- Il *numero di classi* in cui discretizzare le variabili continue e il *criterio* da applicare (per range o per frequenza).
- L'*ordine* di considerazione delle variabili: questo determina un ordine di considerazione dei potenziali genitori per ogni nodo.
- Il *numero massimo di 'genitori'* per ogni nodo.

Nel nostro esempio

Possiamo provare ad apprendere la rete assegnando:

Vincoli sull'ordine di considerazione delle variabili:

Var. 'causali'

Effetti

Discretizzazione di age per frequenza

Nessun limite sul numero di genitori.

... applicare l'algoritmo di apprendimento...

... e osservare struttura e parametri della rete appresa.

Come verificare se la rete 'appresa' è soddisfacente?

Come penso di usare la rete?

Quali delle variabili nella RCP precedente

Sono nascoste ('hidden')

e quali sono 'osservabili'?

Nel nostro esempio:

Età, sesso e residenza sono 'osservabili'.

Le altre, sono 'nascoste'.

Verifica di RCP

1. Verificare la plausibilità dei parametri calcolati dall'algoritmo (probabilità a priori e condizionate)

Esempio:

$P(\text{Tiene alla forma} | \text{Gender} = \text{M}) = .2$

$P(\text{Tiene alla forma} | \text{Gender} = \text{F}) = .8$

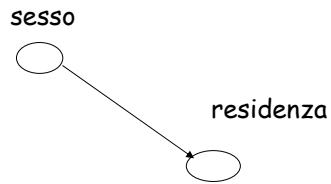
$P(\text{Luogo di Residenza} = \text{Bari}) = .5$

Se un parametro è poco plausibile, lo si corregge

Verifica di RCP

3. Verificare la plausibilità di link 'appresi'

E' plausibile che il sesso (Gend) influenzi la residenza?



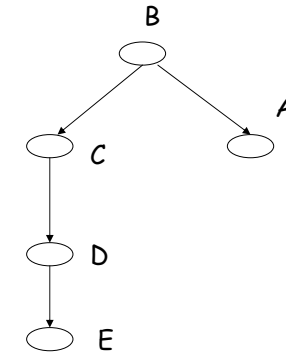
4. Verificare la mancanza di link 'attesi'

E' plausibile che manchi un link che lega l'interesse rispetto alla forma fisica al praticare o no uno sport?



Verifica di RCP

5. Verificare l'effetto della 'lontananza' fra nodi



In questa rete, la probabilità di A influenza la probabilità di E, ma debolmente, vista la lunghezza del cammino che lega questa coppia di nodi.

Verifica di RCP

6. Verificare l'effetto della proprietà di d-separazione

Consideriamo una RCP che contenga 5 nodi: A, B, C, D, E.

La sua probabilità congiunta (o *marginale*) sarà in genere uguale (come abbiamo visto nell'Unità 4) a:

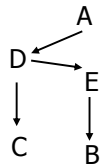
$$P(A, B, C, D, E) = P(A | B, C, D, E) * P(B | C, D, E) * P(C | D, E) * P(D | E) * P(E)$$

... e idem per i diversi valori delle 5 variabili...

La struttura di una rete appresa introduce implicitamente delle ipotesi di *indipendenza condizionale* (o d-separazione) fra alcune delle variabili che (come abbiamo visto nell'Unità precedente) semplificano il calcolo delle prob marginali.

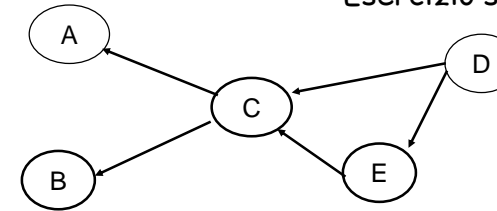
Ad esempio, se B è indipendente da A, C e D dato E, avremo che:

$$P(B | A, C, D, E) = P(B | E)$$



Questo significa che, se introduco una 'evidenza' su E, (ad es, E è 'vero'), la probabilità di B varierà. Ma a questo punto, introdurre evidenza su A, D o C non avrà più nessun effetto su B!!

Esercizio sulla d-separazione



a) Which of the following statements are implied by the indicated network structure; answer yes and no; and give a brief reason for your answer!

i) $P(A, B | C) = P(A | C) * P(B | C)$
yes, because...

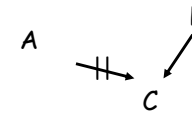
ii) $P(C, E | D) = P(C | D) * P(E | D)$
no, because...

iii) $P(C | A) = P(C)$
no, because...

Verifica di RCP

Come correggere eventuali errori rilevati nel modello?

1. Errori sui parametri:
facili da correggere, in modo iterativo
2. Errori sui link:
si possono aggiungere o togliere dei link, ma *bisogna rivedere i parametri*. Processo iterativo.
3. Errori dovuti alla d-separazione:
sono i più complessi da gestire.



Come correggere i parametri quando si cancella un link

A	B	Y	N
H	H	0.115	0.885
H	M	0.000	1.000
H	L	0.500	0.500
C	H	0.043	0.957
C	M	0.040	0.960
C	L	0.000	1.000


$A = H/C$; $B = H/M/L$; $C = Y/N$

Obiettivo:

cancellare il nodo A conservando la relazione fra il nodo C e il nodo genitore B.

Metodo:

media dei valori corrispondenti



B	Y	N
H	.079	.921
M	.020	.980
L	.250	.750

Come correggere problemi di d-separazione: Apprendimento Semi-Automatico

E' possibile imporre vincoli sulla struttura della rete,
e lasciare all'algorithm il compito di completarla.

L'algorithm aggiungerà dei link a quelli imposti,
calcolando i valori di tutti i parametri.

Il rischio è di ottenere una struttura *molto complessa*.

Apprendimento Semi-Automatico di RCP

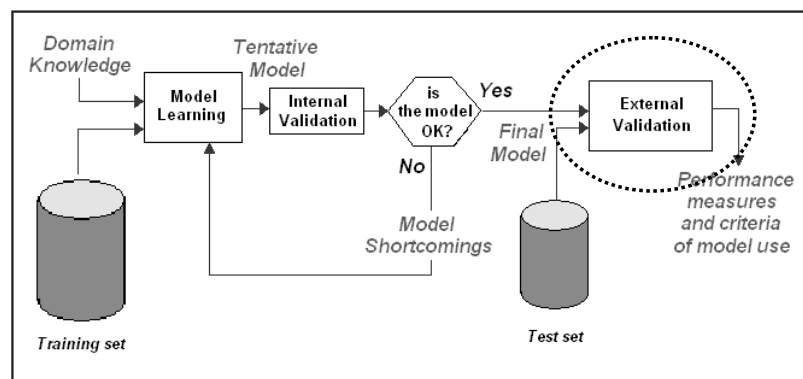
Esempio: per correggere il problema di d-separazione visto 6 lucidi fà,
potremmo inserire un link tra i nodi D e B.

Mi chiedo allora:
l'algorithm aveva considerato questa ipotesi,
nella costruzione della rete?

Che livello di MLL aveva ottenuto?

Provo ad imporre questo vincolo...

Outline del processo iterativo di apprendimento



Test del Modello

Ci sono due modi per testare la bontà di un modello:

- Il modo più classico (*test interno*) consiste nel dividere l'insieme D dei dati, in modo casuale, in un *training set*, (che comprenda ad es, il 75% dei casi) e in un *test set* (che comprenda il restante 25 %), di 'costruire' il modello sul training set e di verificare poi la accuratezza delle previsioni sul test set.
- Un modo più pragmatico (*test esterno*) addestra il modello su tutti i dati di D e lo sottopone poi a valutazione verificandone il comportamento su un *nuovo gruppo di casi reali* (ad es, di utenti).

Misure di performance

E' possibile confrontare le risposte fornite dal modello su una o più variabili assunte come 'particolarmente rilevanti' nel modello.

Il confronto si basa sul calcolo del 'tasso di veri positivi' (TPR) e di 'falsi positivi' (FPR)

Realtà Modello	T	F	tot
T	a	b	a+b
F	c	d	c+d
tot	a+c	b+d	

$$TPR = a/(a+c)$$

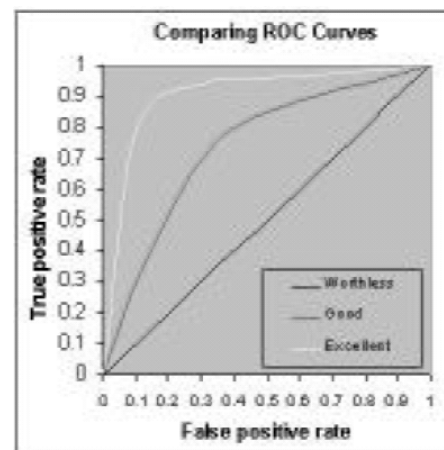
$$FPR = b/(a+c)$$

Chi definisce la 'realtà'?
Una misura 'oggettiva' esterna
Al modello

Le curve ROC

('Receiver Operating Characteristics')

Mettono in relazione il 'tasso di veri positivi' con il 'tasso di falsi positivi'.



Maggiore è l'area compresa dalla curva, migliore è l'approssimazione del modello.

Esercizio

Applicate al nostro esempio (o ad un altro esempio a vostra scelta) la metodologia che abbiamo descritto

Quali Tool possiamo utilizzare per l'apprendimento automatico di RCP?

- Hugin
- Bayesware
- BNJ
- Netica
-

... alcuni dei quali in versione free...

Riferimenti