Concetti di base sulla Teoria delle Probabilità e sulle Reti bayesiane

Mariagrazia Semenza

Concetti preliminari

Ambito di applicazione: **l'incertezza, le possibilità**

Oggetto della teoria delle probabilità: **"variabili random" in un dominio di interesse**

Esempi di "variabili random" o aleatorie:

- numeri aleatori
- funzioni aleatorie
- variabili categoriche o, in generale, enti aleatori
- eventi aleatori

Definizioni di probabilità

Approccio oggettivista o frequentista, che considera la probabilità come un'**entità misurabile**, legata alla frequenza di accadimento

Approccio soggettivista, che considera la probabilità una misura del **grado di attesa** soggettivo del verificarsi di un evento (schema delle scommesse)

Teoremi delle probabilità totali

Sono i teoremi che nel caso degli eventi stabiliscono le proprietà **necessarie e sufficienti** per definire la proprietà additiva delle probabilità rispettando le **condizioni di coerenza**.

Teorema 1: Nel caso di **eventi incompatibili**, la probabilità dell'evento somma deve coincidere con la somma delle probabilità

Teorema 2: Nel caso di **partizioni** (finite) le probabilità devono dare come somma 1.

In particolare, per due eventi opposti (partizione con n=2) le probabilità di eventi opposti devono essere complementari

Partizione

E' un **insieme di eventi incompatibili ed esaustivi**, dei quali cioè è **certo** che deve verificarsi **uno ed uno solo**.

Come si ricava una partizione partendo da un insieme qualunque di eventi?

Se E1,...,En sono incompatibili ma non esaustivi si completa la partizione con
Eo=1- (E1+...+En) = notE1· notE2⋯ notEn

In generale i termini della partizione si possono ricavare da
(E1+notE1)(E2+notE2)(En+notEn) = 1

Ecco i termini, detti costituenti:

E1·E2···En,
E1·E2··· notEn,
E1· notE2···En,
...,
notE1· notE2··· notEn

Il numero massimo di costituenti della partizione determinata da E1, E2, ...,En è $\mathbf{2^n}$

Geometricamente possono essere visti come versori che individuano gli assi ortogonali di uno spazio ad $\mathbf{s \leq 2^n}$ dimensioni

Corollario al teorema 2:

perchè risulti determinata la probabilità di tutti gli eventi E logicamente dipendenti da E1,...,En è **necessario e sufficiente** attribuire delle probabilità a tutti i costituenti C1,...,Cs.
La p(E) ne risulta linearmente dipendente.

Le tre leggi alla base della teoria delle probabilità:

- Condizione di convessità: $0 \leq p(E) \leq 1$
- Additività semplice: $p(A \cup B) \leq p(A)+p(B)$
  (l'uguaglianza stretta vale solo nel caso di eventi incompatibili)
- Regola del prodotto: $p(AB) = p(A|B)p(B) = p(B|A)p(A)$

Dalla terza proprietà deriva direttamente:

$$\mathbf{p(A|B)} = p(B|A)p(A)/p(B) = \mathbf{K\, p(B|A)p(A)}$$

Teorema di Bayes

Significato del Teorema di Bayes: **apprendimento dall'esperienza**

$$p(A|B) = K\, p(B|A)\, p(A)$$

p(A) probabilità a priori
p(A|B) probabilità a posteriori
p(B|A) verosimiglianza
K fattore di normalizzazione

E se p(A) = 0?

Definizioni utili per le reti bayesiane

Regola del prodotto: $p(A,B) = p(A|B)\, p(B)$

$p(A,B)$ è detta **probabilità congiunta**

La probabilità congiunta di un insieme di eventi può essere espressa come una **catena di probabilità condizionali**:

$$p(A,B,C,D) = p(A|B,C,D)p(B|C,D)p(C|D)p(D)$$

In generale:

$$p(E1,E2,...,Ek) = \prod_{i=1}^{k} p(Ei|Ei-1,...E1)$$

Definizioni utili per le reti bayesiane

Inferenza probabilistica

Consideriamo un insieme di eventi E1,E2,..,Ek e tutte le possibili combinazioni dei loro valori Vero e Falso

Conosciamo tutti i valori: p(E1,E2,...,Ek)

Supponiamo che un sottoinsieme di questi presenti un valore definito, per es. Ej = Vero = **e**

Chiamiamo inferenza probabilistica il processo di calcolo del valore

$$p(Ei=Vero|Ej=\mathbf{e})$$

Inferenza probabilistica: p(Ei=Vero|Ej=**e**)

Come si calcola?

p(Ei=Vero|Ej=**e**) = p(Ei=Vero,Ej=**e**)/p(Ej=**e**) (def. probabilità condizionali, o regola del prodotto)

Per calcolare p(Ei=Vero,Ej=**e**) ricordiamo che abbiamo tutti i valori p(E1,E2,...,Ek) e quindi

$$p(Ei=Vero,Ej=\mathbf{e}) = \sum_{Ei=Vero,\ Ej=\mathbf{e}} p(E1,...,Ek)$$

Allo stesso modo si calcola p(Ej=**e**)

"In generale l'inferenza probabilstica non è trattabile usando questo metodo, perchè, avendo k eventi, dovremmo avere una lista di $2^{\mathbf{k}}$ probabilità congiunte p(E1,E2,...,Ek). Per molti problemi che ci interessano non potremmo scrivere una lista del genere anche se l'avessimo (e che generalmente non abbiamo)"

Definizioni utili per le reti bayesiane

Secondo Pearl gli umani organizzano la loro conoscenza utilizzando il concetto di

Indipendenza condizionale

Diciamo che l'evento A è condizionalmente indipendente da un evento B, dato l'evento C, se

p(A|B,C) = p(A|C) (indipendenza stocastica subordinata)

In pratica A è condizionalmente indipendente da B, dato C, se la conoscenza di B non porta a nessuna ulteriore variazione della probabilità di A rispetto a quella apportata dall'avverarsi di C.

Così, per quanto riguarda A, se conosciamo C possiamo ignorare B.

Dall'indipendenza condizionale di A e B dato C otteniamo:

p(A,B|C) = p(A|C)p(B|C) (= p(B,A|C))

In generale B e C possono essere insiemi di eventi; se C è un insieme vuoto:

p(A,B) = p(A)p(B) (noncorrelazione)

Le **reti bayesiane** sono una struttura che rappresenta le indipendenze condizionali

Sono DAG (grafi diretti aciclici) i cui nodi sono gli eventi

Una rete bayesiana stabilisce che ogni nodo, dati i suoi immediati ascendenti (parents *P*), è **condizionalmente indipendente** da ogni altro che **non** sia suo discendente

Cioè, per ogni E del grafo p(E|A,*P*(E)) = p(E|*P*(E)) per ogni A, evento o insieme di eventi, che non sia discendente di E. Possiamo anche scrivere I(E,A|*P*(E))

Le reti bayesiane sono chiamate anche **reti causali**, perchè gli archi che connettono i nodi possono essere visti come se rappresentassero relazioni causali dirette

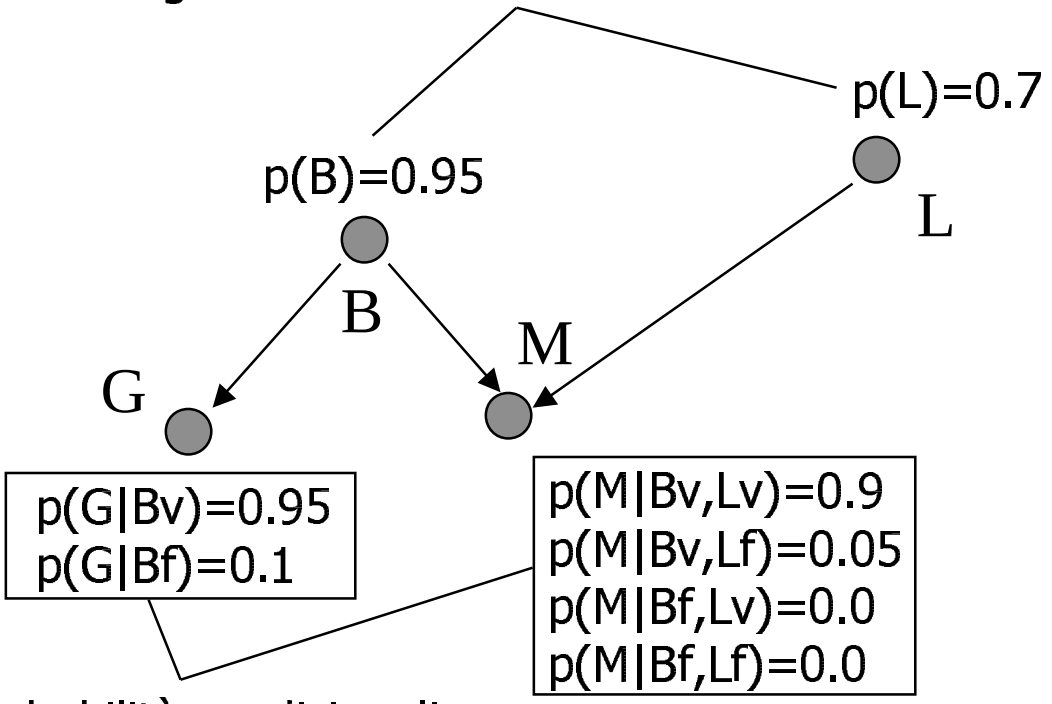
"Gli umani esperti sono **spesso** capaci di collegare cause ed effetti in modo tale da evidenziare indipendenze condizionali rappresentabili con una rete bayesiana".

ATTENZIONE: ogni strumento è limitato; non esiste uno strumento universale di rappresentazione della conoscenza.

Nel tentativo di costruire un modello semplifichiamo, schematizziamo la realtà o quello che della realtà abbiamo percepito.

Lo strumento non deve forzarci a distorcere ulteriormente il modello per esigenze di calcolo.

Probabilità a priori, associata ad ogni nodo che non ha ascendenti



Tavole delle probabilità condizionali associate ad ogni nodo ed ai suoi ascendenti

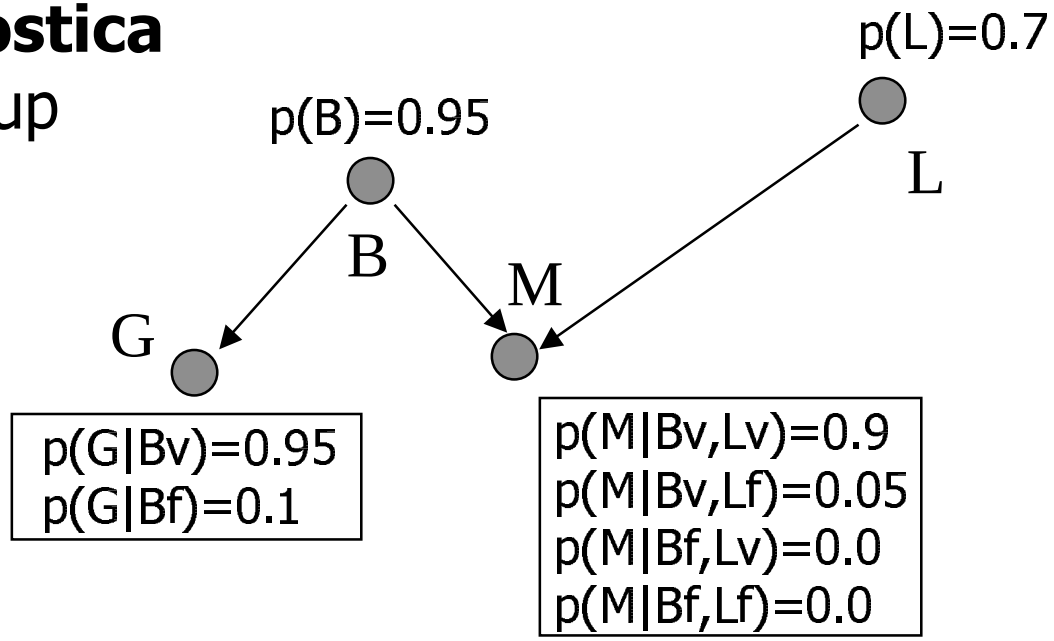
Formula della rete:

p(B,G,M,L) = p(G|B)p(M|B,L)p(L), invece di

p(B,G,M,L) = p(G|B,M,L)p(B|M,L)p(M|L)p(L)

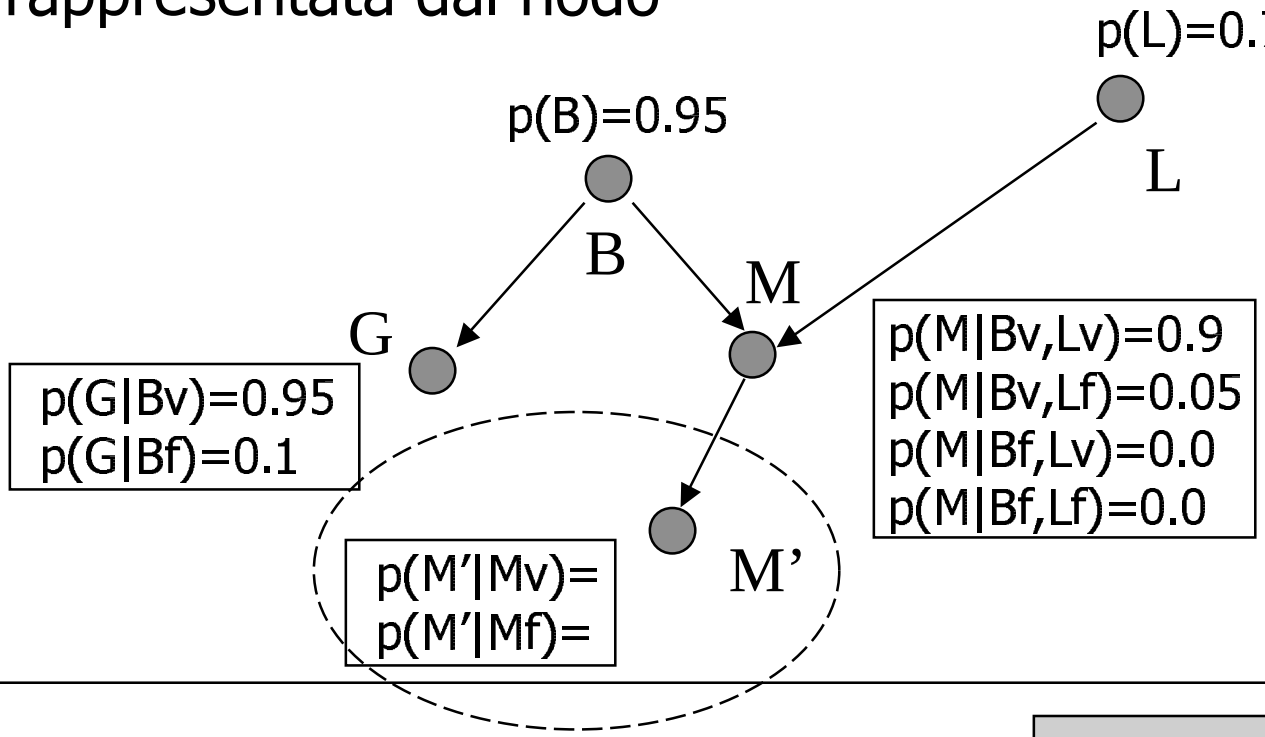
Cammini inferenziali nelle reti di Bayes

- **Inferenza causale**
  o top-down
- **Inferenza diagnostica**
  o bottom-up
- **Explaining away**



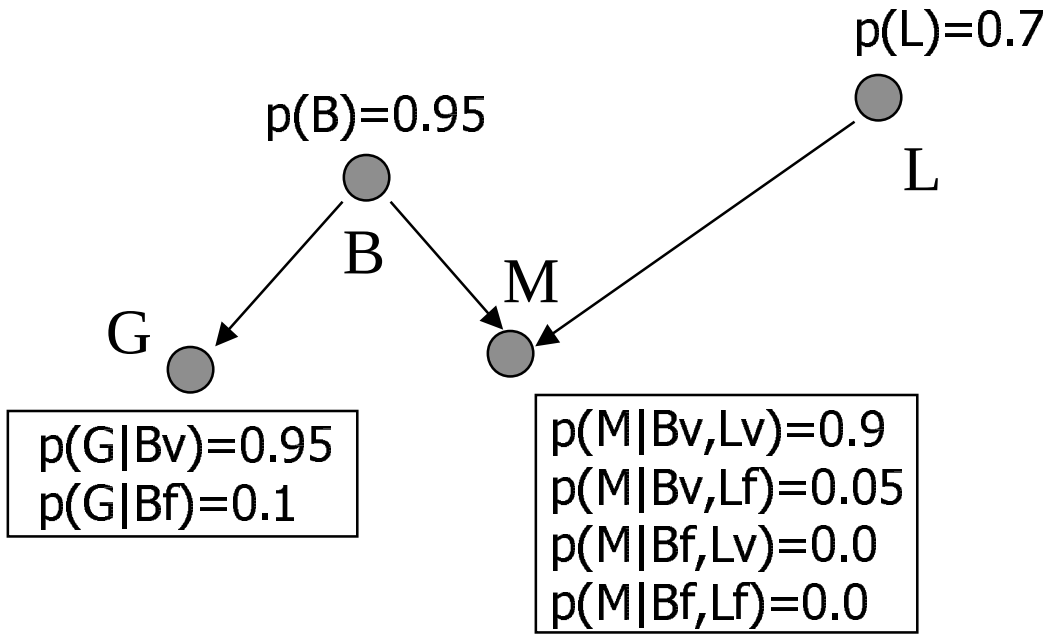
Evidenza incerta

"Nel calcolo delle reti bayesiane, perchè possa essere considerata per "data" l'evidenza di un nodo, dobbiamo esser certi della verità o falsità della proposizione rappresentata dal nodo"



D-Separation

(Separazione direction-dependent)



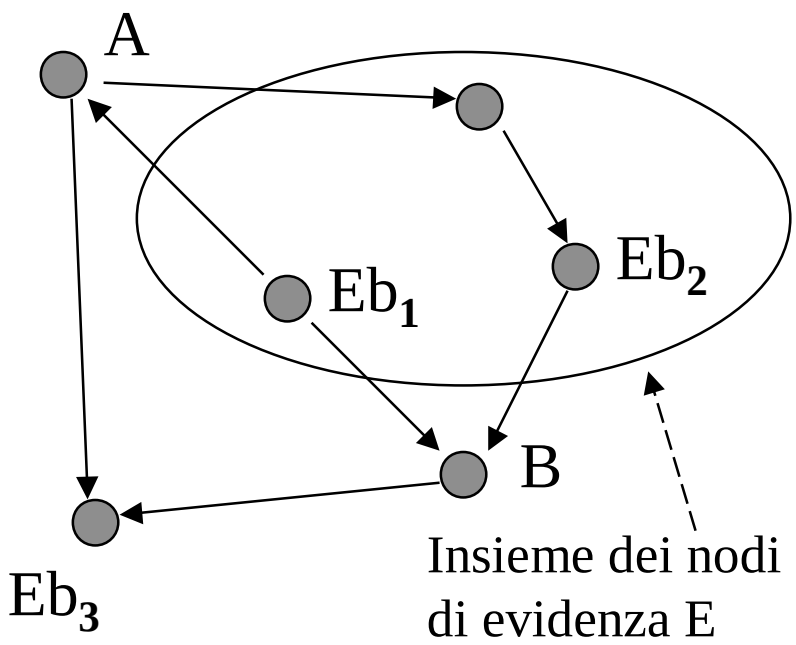
Le tre regole che definiscono la d-separation:

Due nodi A e B sono condizionalmente indipendenti, dato un insieme di nodi di evidenza E, se su ogni cammino che unisce A con B esiste un nodo Eb che gode di una delle seguenti proprietà:

1 - Eb è in E ed entrambi gli archi del cammino sono diretti fuori da Eb

2 - Eb è in E ed un arco va verso Eb mentre l'altro esce

3 - Eb non è in E ed entrambi gli archi vanno verso Eb



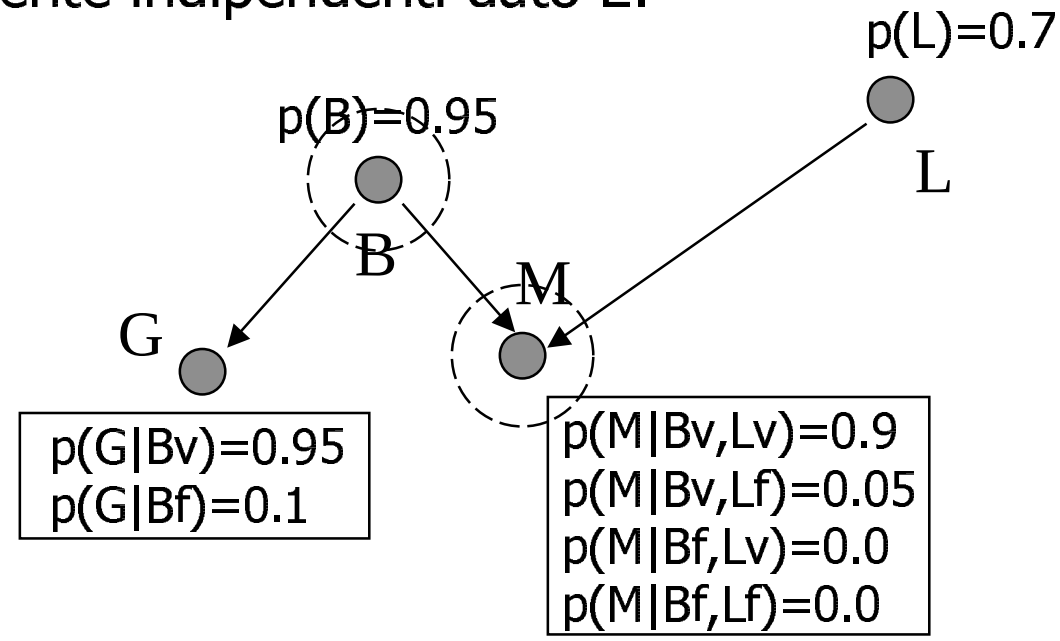
Quando per un cammino vale una di queste condizioni diciamo che **Eb blocca il cammino**, e se tutti i cammini che collegano A e B sono bloccati diciamo che **E d-separa A e B**.

A e B sono condizionalmente indipendenti dato E.

Esempio:

I(G,L|B)
per la regola 1, **|B**
per la regola 3, **M**

mentre non vale
I(B,L|M)



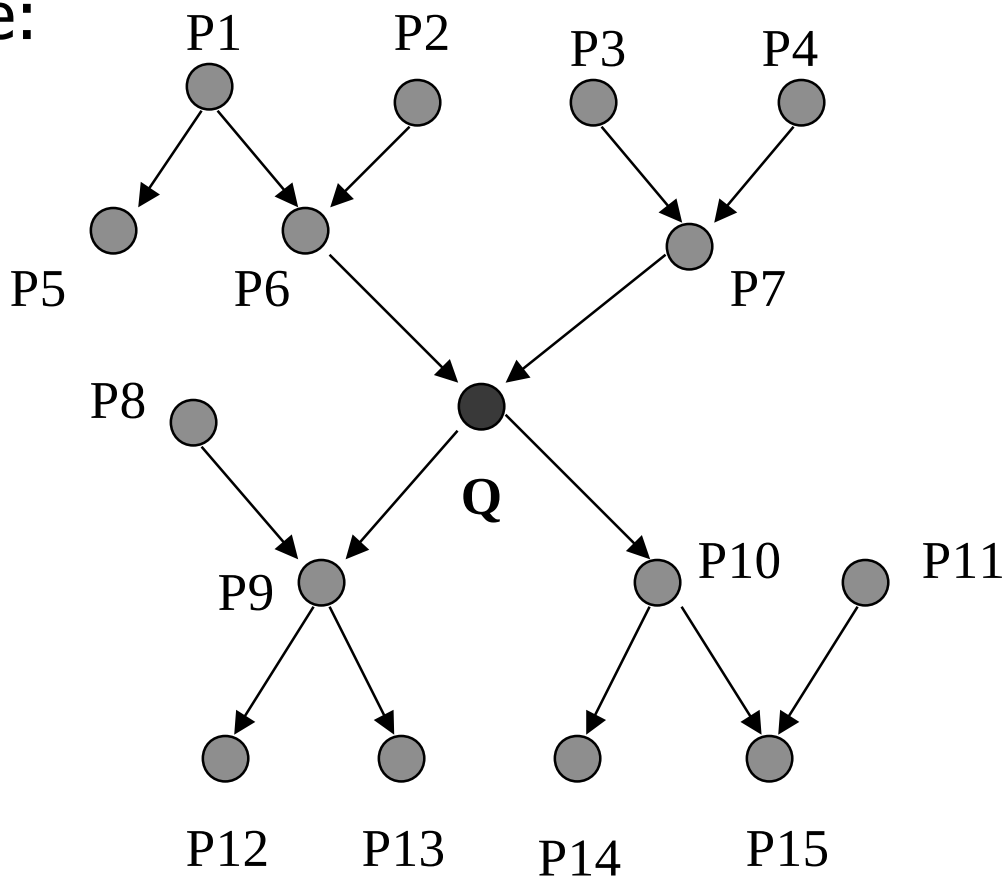
Politree

Anche con le semplificazioni introdotte dalle reti bayesiane, strutture del genere, se di grandi dimensioni, non sono trattabili.

Fortunatamente, però, per un'importante classe di reti, chiamata politree, esistono numerose scorciatoie per calcolare le probabilità condizionali, cioè per compiere inferenza probabilistica.

Un politree è un DAG (grafo diretto aciclico) tale che esista un solo cammino, lungo gli archi in qualunque direzione, tra due nodi del DAG.

Ecco un "tipico" politree:



Vogliamo calcolare la p(Q|qualche altro nodo)

Si usano algoritmi **ricorsivi** bottom-up e/o top-down

La complessità degli algoritmi cresce **linearmente** con il numero di nodi, ma questo vale **solo** per i politree

Quando una rete non è un politree le procedure ricorsive non possono essere utilizzate a causa dei cammini multipli che collegano i nodi.

Approcci possibili:

Monte Carlo - ai nodi privi di ascendenti si assegnano random, in funzione delle rispettive probabilità, i valori Vero e Falso.
A partire da questi valori, ed usando le CPT dei discendenti, si assegnano a questi ultimi valori random, e così via fino a completare tutta la rete.
Si ripete il processo per molte ("infinite") volte e si memorizzano i risultati.

p(Qv|Ev) si calcola dividendo il numero di volte in cui Q=Vero ed E=Vero per il numero di volte in cui E=Vero

Altro approccio possibile:

Clustering - consiste nel raggruppare un certo numero di nodi in supernodi, in modo tale che la rete alla fine risulti un politree.

I possibili valori di un supernodo sono ora tutte le combinazioni dei possibili valori dei nodi che lo compongono.

L'inconveniente è che ora ad ogni supernodo sono associate molte CPT, che ora danno le probabilità condizionali di tutti i valori dei supernodi, condizionate da tutti i valori dei nodi parents, che possono essere supernodi a loro volta.

Costruire una rete bayesiana

Si parte da un training set di dati, cioè da un insieme di valori (evidenze, osservazioni, misure) associati a tutte o ad una parte delle "variabili".

Bisogna trovare una rete che si accordi nel **modo migliore** con il training set, una volta stabilito il **metodo di punteggio** per ognuna delle reti ipotizzate.

Per "trovare una rete" si intende trovare sia la struttura del DAG, sia le CPT (tavole di probabilità condizionali) associate ad ogni nodo.

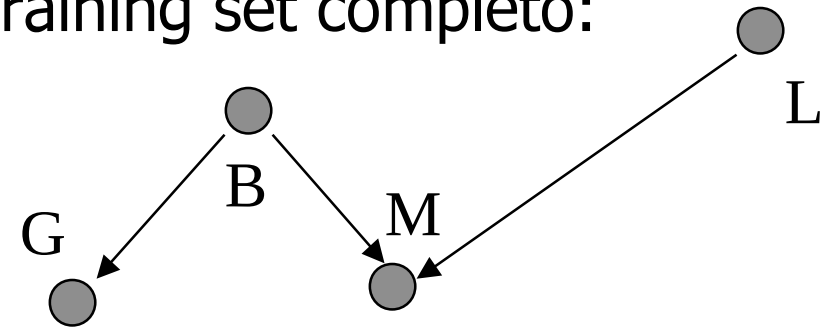
1° caso: la struttura della rete è nota

Dobbiamo solo trovare le CPT (Spesso gli esperti riescono a trovare la struttura appropriata per il problema ma non le CPT)

Se siamo fortunati abbiamo un training set completo, che cioè contiene un valore per ogni variabile rappresentata nella rete.

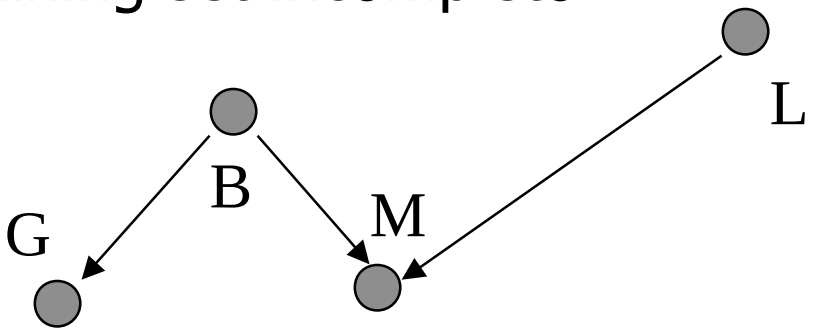
Altrimenti il problema si complica.

Esempio di struttura nota con training set completo:



G	M	B	L	N° osservazioni
Vero	Vero	Vero	Vero	54
Vero	Vero	Vero	Falso	1
Vero	Falso	Vero	Vero	7
Vero	Falso	Vero	Falso	27
Falso	Vero	Vero	Vero	3
Falso	Falso	Vero	Falso	2
Falso	Falso	Falso	Vero	4
Falso	Falso	Falso	Falso	2
				100

Esempio di struttura nota con training set incompleto



G	M	B	L	N° osservazioni
Vero	Vero	Vero	Vero	54
Vero	Vero	Vero	Falso	1
*	*	Vero	Vero	7
Vero	Falso	Vero	Falso	27
Falso	Vero	*	Vero	3
Falso	Falso	Vero	Falso	2
Falso	Falso	Falso	Vero	4
Falso	Falso	Falso	Falso	2
				100

2° caso: bisogna trovare anche la struttura

Cioè bisogna trovare quella struttura, associata alle relative CPT, che meglio si adatta al training set.

Per far questo bisogna trovare una metrica per assegnare un punteggio ad ogni struttura candidata e specificare una procedura di ricerca fra le possibili strutture, in quanto non è possibile esaminare in modo esaustivo tutte le posibili strutture.

Spesso si procede per piccoli cambiamenti: si può partire da una struttura qualunque (anche quella senza archi, dove cioè i nodi sono indipendenti) e si aggiunge o si toglie un arco, o se ne modifica la direzione. Si calcola il punteggio per ogni modifica.

2° caso: bisogna trovare anche la struttura

In alcuni casi la struttura della rete può essere semplificata in modo significativo introducendo un nodo i cui valori non sono contenuti nel training set, un cosiddetto nodo nascosto

Esempio:

La bontà dell'introduzione di un nodo nascosto va valutata usando la stessa metrica stabilita per gli archi, considendo che i dati associati a questo nodo non sono misurabili, quindi sono da considerare "mancanti"

Esempi di applicazioni delle reti bayesiane

- Diagnosi mediche (Pathfinder, 1990)
- Analisi decisionale (Si incorporano *nodi dei valori* e *nodi di decisione*; si parla allora di diagrammi di influenza)
- Diagnosi di reti mobili (cellulari)
- Analisi di dati finanziari
- Previsioni meteorologiche
- Esplorazioni in campo petrolifero
- Esplorazione del deep space (NASA AutoClass, open source in LISP e C)
- Interazione software-utente (Microsoft, progetto Lumiere)
- Interazione pilota-aereo
- Data mining
- Costruzione di mappe (robotica)
- Comprensione di storie

Ambienti free di sviluppo di reti bayesiane

MSBNx di Microsoft
(http://research.microsoft.com/adapt/MSBNx/)

Un minimo di riferimenti bibliografici:

"Probabilistic reasoning in intelligent systems:networks of plausible inference" di Judea Pearl Ed. Morgan Kaufmann

"Intelligenza artificiale" di N. Nilsson Ed. Apogeo

Per pubblicazioni in rete
v. elenco in http://research.microsoft.com/adapt/
oppure
http://www.kddresearch.org/Resources/Papers/Intro/pearl-bbn2000.pdf