Reti Bayesiane

Capitolo 14

Outline

- ◊ Sintassi
- ◊ Semantica
- ◊ Inferenza esatta tramite enumerazione
- ◊ Inferenza esatta tramite eliminazione di variabile
- ◊ Inferenza approssimata tramite simulazione stocastica

Reti Bayesiane (Bayesian networks)

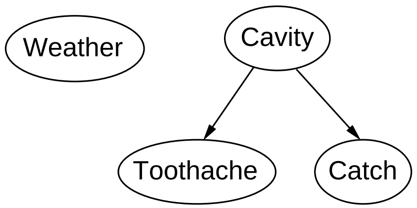
Una semplice notazione grafica per asserzioni condizionalmente indipendenti e quindi per specifiche di distribuzioni condizionali complete

Sintassi:
un insieme di nodi, uno per variabile
un grafo diretto aciclico (link ≈ "influenza direttamente")
una distribuzione condizionale per ogni nodo dati i suoi genitori:
$\mathbf{P}(X_i|Parents(X_i))$

Nel caso più semplice, distribuzione condizionale rappresentata come una tabella della probabilità condizionale (CPT) data la distribuzione su X_i per ogni combinazione di valori assunti dai genitori

Esempio

La topologia della rete codifica asserzioni di indipendenza condizionale:



$Weather$ è indipendente dalle altre variabili

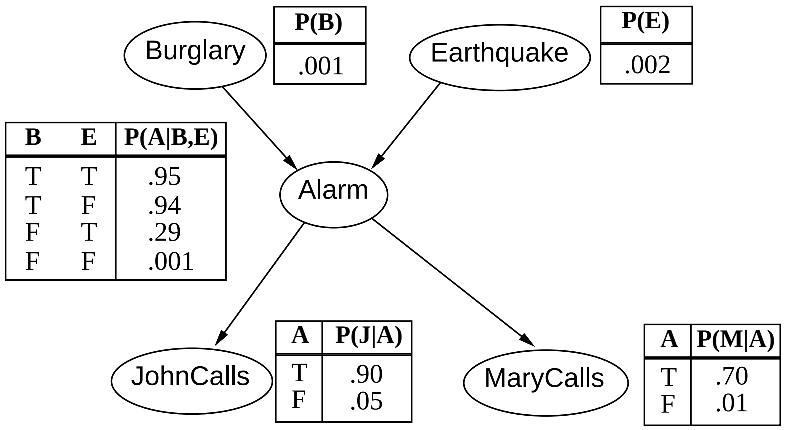
$Toothache$ e $Catch$ sono condizionalmente indipendenti data $Cavity$

Esempio

Sono al lavoro, il vicino John chiama per dire che il mio allarme $Alarm$ è entrato in funzione, ma la vicina Mary non chiama. Alcune volte l'allarme è attivato da piccole scosse di terremoto. C'è un ladro in casa ?

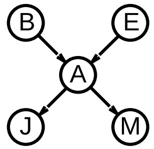
Variabili: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
La topologia della rete riflette conoscenza "causale":
- Un ladro può attivare l'allarme
- Un terremoto può attivare l'allarme
- L'attivazione dell'allarme può indurre Mary a chiamare
- L'attivazione dell'allarme può indurre John a chiamare

Esempio



Compattezza

Una CPT per variabili Booleane X_i con k genitori Booleani ha 2^k righe per le combinazioni di valori dei genitori



Ogni riga richiede un numero p per $X_i = vero$
(il numero per $X_i = falso$ è $1 - p$)

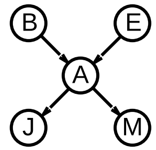
Se ogni variabile non ha più di k genitori,
la rete completa richiede $O(n \cdot 2^k)$ numeri

Cioè, cresce linearmente con n, vs. $O(2^n)$ per la distribuzione congiunta completa

Per la rete precedente, $1 + 1 + 4 + 2 + 2 = 10$ numeri (vs. $2^5 - 1 = 31$)

Semantica globale

La semantica globale definisce la distribuzione congiunta completa come il prodotto delle distribuzioni condizionali locali:



$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

p.e., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$=$

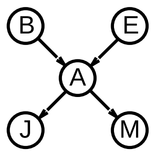
Semantica globale

La semantica globale definisce la distribuzione congiunta completa come il prodotto delle distribuzioni condizionali locali:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

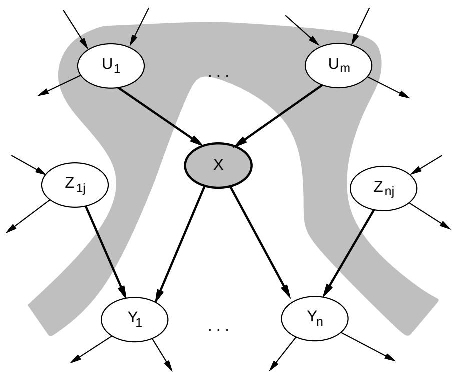
p.e., $P(j \land m \land a \land \neg b \land \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$



Semantica locale

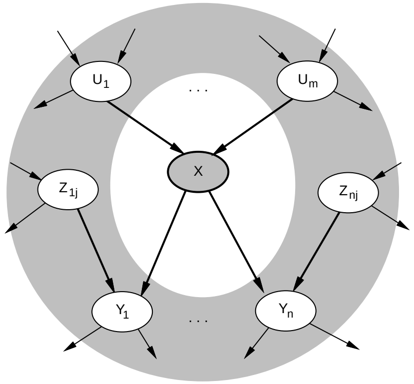
Semantica locale: ogni nodo è condizionalmente indipendente dai suoi non discendenti dati i genitori



Teorema: Semantica locale $\Leftrightarrow$ semantica globale

Markov blanket

Ogni nodo è condizionalmente indipendente da tutti gli altri dato il suo Markov blanket: genitori + figli + genitori dei figli



Costruzione di Reti Bayesiane

Necessità di un metodo tale che data una serie di asserzioni di indipendenza condizionale localmente controllabili, garantisca la semantica globale desiderata

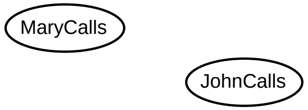
1. Scegliere un ordinamento di variabili $X_1, \ldots, X_n$
2. For $i = 1$ to n
   aggiungi X_i alla rete
   seleziona genitori da $X_1, \ldots, X_{i-1}$ tali che
   $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$

Questa scelta di genitori garantisce la semantica globale:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \prod_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1}) \quad \text{(chain rule)} \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i)) \quad \text{(per costruzione)}\end{aligned}$$

Esempio

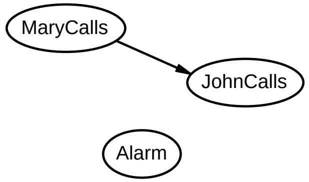
Supponiamo di scegliere l'ordine M, J, A, B, E



$P(J|M) = P(J)$?

Esempio

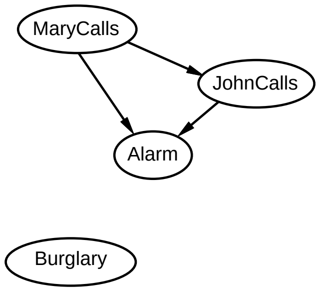
Supponiamo di scegliere l'ordine M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

Esempio

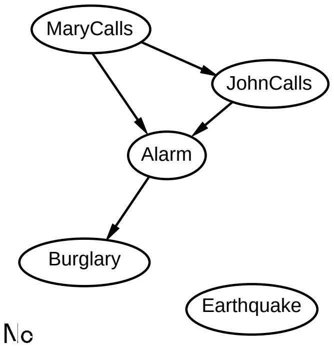
Supponiamo di scegliere l'ordine M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$?
$P(B|A, J, M) = P(B)$?

Esempio

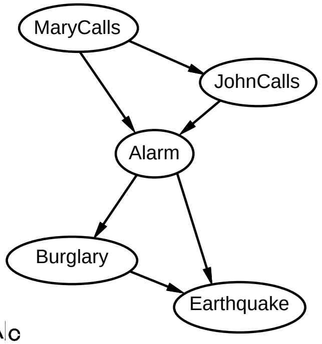
Supponiamo di scegliere l'ordine M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$? Yes
$P(B|A, J, M) = P(B)$? No
$P(E|B, A, J, M) = P(E|A)$?
$P(E|B, A, J, M) = P(E|A, B)$?

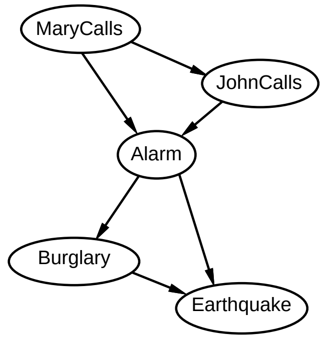
Esempio

Supponiamo di scegliere l'ordine M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? No
$P(B|A, J, M) = P(B|A)$? Yes
$P(B|A, J, M) = P(B)$? No
$P(E|B, A, J, M) = P(E|A)$? No
$P(E|B, A, J, M) = P(E|A, B)$? Yes

Esempio



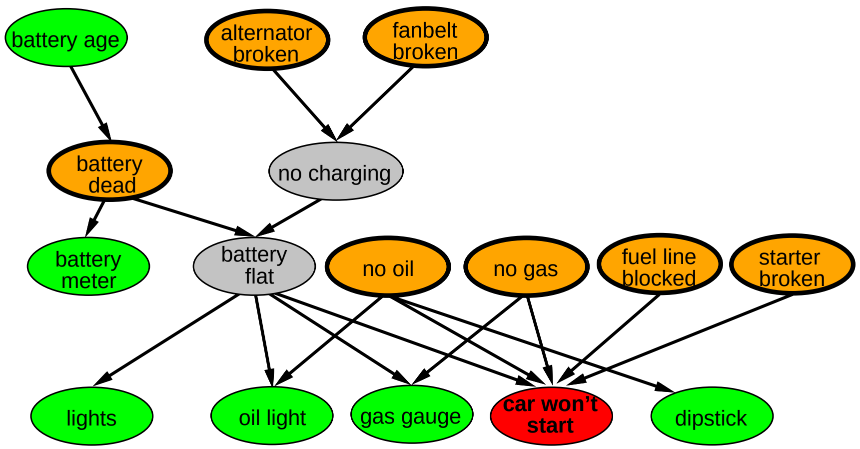
Decidere l'indipendenza condizionale è difficile nelle direzioni non causali

Valutare le probabilità condizionali è difficile in direzioni non causali

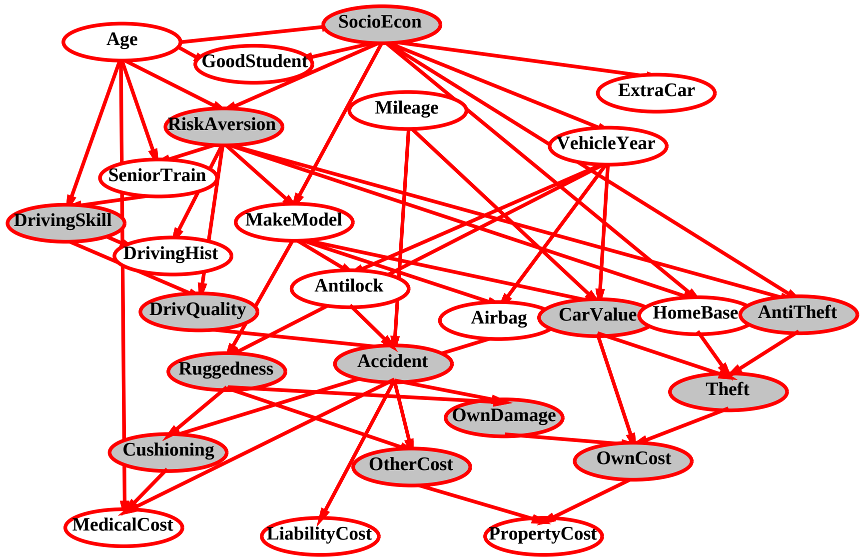
La rete è meno compatta: $1 + 2 + 4 + 2 + 4 = 13$ numeri necessari

Esempio: diagnosi per automobile

Evidenza iniziale: auto non parte
Variabili controllabili (in verde), variabili "rotte, da aggiustare" (in arancio)
Variabili nascoste (in grigio) assicurano struttura sparsa, riducono i parametri



Esempio: assicurazione dell'automobile



Compiti di inferenza

Query semplici: calcolare la probabilità a posteriori marginale $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$
p.e., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$

Query congiuntive: $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

Decisioni ottimali: reti di decisioni includono informazioni di utilità;
inferenza probabilistica richiesta per $P(outcome|action, evidence)$

Recupero informazione: quale evidenza si deve cercare?

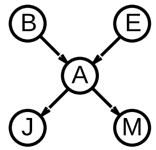
Analisi della sensitività: quali valori di probabilità sono i più critici?

Spiegazione: perché ho bisogno di un nuovo motore di avviamento?

Inferenza tramite enumerazione

Modo un pò più furbo per marginalizzare alcune variabili dalla distribuzione congiunta senza costruire esplicitamente la sua rappresentazione

Query semplice sulla rete dell'allarme:



$\mathbf{P}(B|j,m)$
$= \mathbf{P}(B,j,m)/P(j,m)$
$= \alpha\mathbf{P}(B,j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B,e,a,j,m)$

Riscrittura di entrate della distribuzione congiunta usando il prodotto di entrate di CPT:
$\mathbf{P}(B|j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$
$= \alpha\mathbf{P}(B)\Sigma_e P(e)\Sigma_a\mathbf{P}(a|B,e)P(j|a)P(m|a)$

Enumerazione ricorsiva depth-first: $O(n)$ in spazio, $O(d^n)$ in tempo

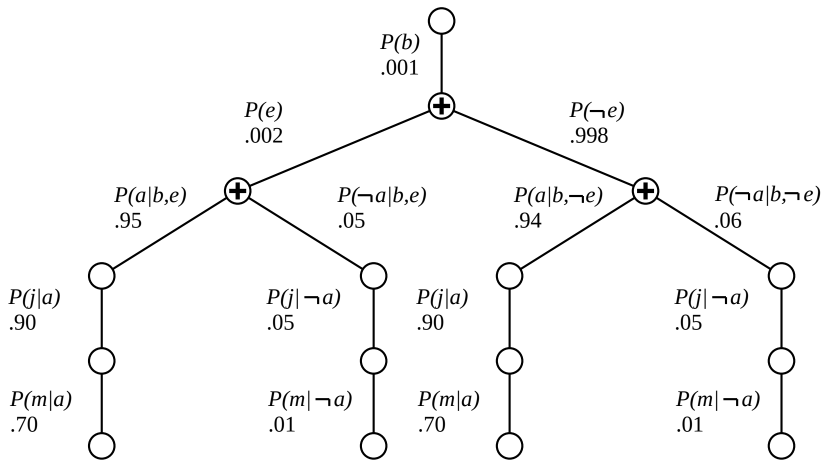
Algoritmo di enumerazione

```
function Enumeration-Ask(X, e, bn) returns a distribution over X
   inputs: X, the query variable
           e, observed values for variables E
           bn, a Bayesian network with variables {X} ∪ E ∪ Y
   Q(X) ← a distribution over X, initially empty
   for each value x_i of X do
       extend e with value x_i for X
       Q(x_i) ← Enumerate-All(Vars[bn], e)
   return Normalize(Q(X))
```

```
function Enumerate-All(vars, e) returns a real number
   if Empty?(vars) then return 1.0
   Y ← First(vars)
   if Y has value y in e
       then return P(y | Pa(Y)) × Enumerate-All(Rest(vars), e)
       else return Σ_y P(y | Pa(Y)) × Enumerate-All(Rest(vars), e_y)
           where e_y is e extended with Y = y
```


Albero di valutazione

L'enumerazione è inefficiente: calcoli ripetuti
p.e., calcola $P(j|a)P(m|a)$ per ogni valore di e



Inferenza tramite eliminazione di variabile

Eliminazione di variabile: effettuare le somme da destra a sinistra, memorizzare i risultati intermedi (fattori) per evitare di ricalcolarli

$$
\begin{aligned}
&\mathbf{P}(B|j,m)\\
&\quad = \alpha \underbrace{\mathbf{P}(B)}_{B} \Sigma_e \underbrace{P(e)}_{E} \Sigma_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M}\\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) P(j|a) f_M(a)\\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) f_J(a) f_M(a)\\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a f_A(a,b,e) f_J(a) f_M(a)\\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) f_{\bar{A}JM}(b,e) \text{ (elimina } A\text{)}\\
&\quad = \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (elimina } E\text{)}\\
&\quad = \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

Eliminazione di variabile: operazioni base

Eliminare una variabile da un prodotto di fattori:

1. muovere i fattori costanti al di fuori della somma
2. aggiungere le sottomatrici al prodotto “pointwise” dei fattori rimanenti

$$\Sigma_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \,\Sigma_x\, f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_{\bar{X}}$$

assumendo che $f_1, \ldots, f_i$ non dipendano da X

Prodotto pointwise di fattori f_1 e f_2:

$$
\begin{aligned}
&f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l)\\
&\quad = f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l)
\end{aligned}
$$

P.e., $f_1(a,b) \times f_2(b,c) = f(a,b,c)$

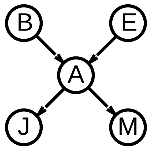
Algoritmo di eliminazione di variabile

```
function ELIMINATION-ASK(X, e, bn) returns a distribution over X
   inputs: X, the query variable
           e, evidence specified as an event
           bn, a belief network specifying joint distribution P(X_1, ..., X_n)

   factors ← []; vars ← REVERSE(VARS[bn])
   for each var in vars do
       factors ← [MAKE-FACTOR(var, e)|factors]
       if var is a hidden variable then factors ← SUM-OUT(var, factors)
   return NORMALIZE(POINTWISE-PRODUCT(factors))
```


Variabili irrilevanti

Consideriamo la query $P(JohnCalls|Burglary = true)$



$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(J|a) \sum_m P(m|a)$$

La somma su m è uguale a 1; M è **irrilevante** per la query

Thm 1: Y è irrilevante a meno che $Y \in Ancestors(\{X\} \cup \mathbf{E})$

Qui, $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$, e
$Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$
quindi M è irrilevante

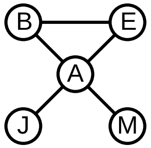
(Confrontare con backward chaining a partire dalla query in KB con clausole di Horn)

Variabili irrilevanti

Defn: grafo moralizzato di una rete bayesiana: sposare tutti i genitori ed eliminare la direzione degli archi

Defn: $\mathbf{F}$ è m-separato da $\mathbf{G}$ tramite $\mathbf{H}$ sse è separato tramite $\mathbf{H}$ nel grafo moralizzato

Thm 2: Y è irrilevante se m-separato da X tramite $\mathbf{E}$



Per $P(JohnCalls|Alarm = true)$, sia $Burglary$ che $Earthquake$ sono irrilevanti

Complessità dell'inferenza esatta

Reti singolarmente connesse (o polytree):
- ogni coppia di nodi è connessa da al più un cammino (non diretto)
- il costo in tempo e spazio della eliminazione di variabile è $O(d^k n)$

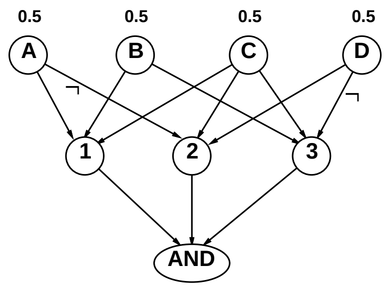
Reti connesse più che singolarmente:
- possibile ridurre 3SAT alla inferenza esatta $\Rightarrow$ NP-hard
- equivalente a modelli 3SAT con **conteggio** (del numero di soluzioni)

$\Rightarrow$ #P-complete

1. A v B v C

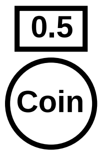
2. C v D v ¬A

3. B v C v ¬D



Inferenza tramite simulazione stocastica

Idea base:
1) Estrarre N campioni da una distribuzione di campionamento S
2) Calcolare la probabilità a posteriori approssimata $\hat{P}$
3) Mostrare che converge alla vera probabilità P



Outline:
- Campionamento da una rete vuota
- Rejection sampling: rigettare i campioni in disaccordo con l'evidenza
- Likelihood weighting: usare l'evidenza per pesare i campioni
- Markov chain Monte Carlo (MCMC): campiona in accordo ad un processo stocastico la cui distribuzione stazionaria è la vera probabilità

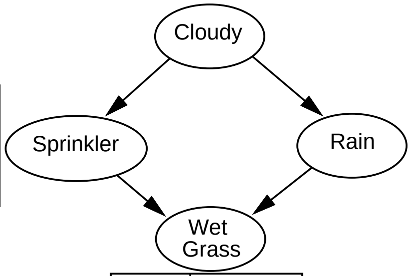
Campionamento da una rete vuota

function PRIOR-SAMPLE(bn) **returns** an event sampled from bn
  inputs: bn, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $\mathbf{x} \leftarrow$ an event with n elements
  for $i = 1$ **to** n **do**
    $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid Parents(X_i))$
  return $\mathbf{x}$

Esempio

P(C)
.50



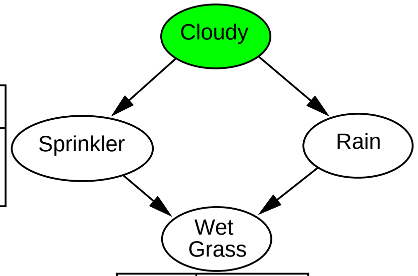
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Esempio

P(C)
.50



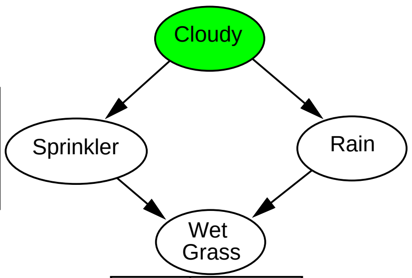
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Esempio

P(C)
.50



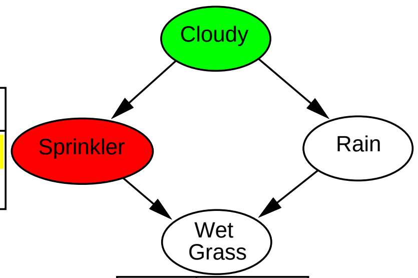
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Esempio

P(C)
.50



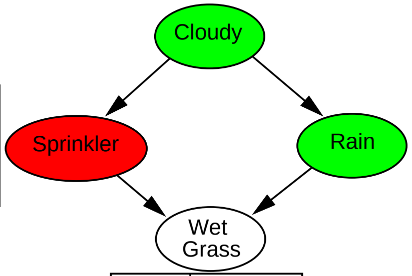
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Esempio

P(C)
.50



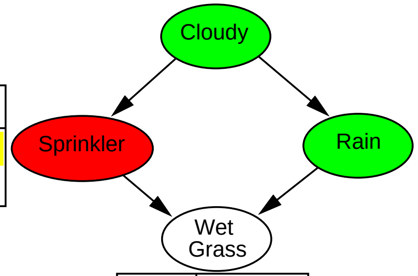
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Esempio

P(C)
.50



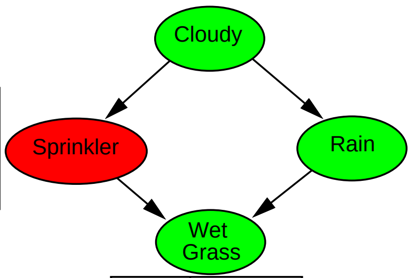
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Esempio

P(C)
.50



C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

Campionamento da una rete vuota

Probabilità che PRIORSAMPLE generi un evento particolare
$S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i|Parents(X_i)) = P(x_1 \ldots x_n)$
cioè, la vera probabilità a priori

P.e., $S_{PS}(t,f,t,t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t,f,t,t)$

Posto $N_{PS}(x_1 \ldots x_n)$ essere il numero di campioni generati per l'evento $x_1, \ldots, x_n$

Allora abbiamo

$$\begin{aligned} \lim_{N\to\infty} \hat{P}(x_1,\ldots,x_n) &= \lim_{N\to\infty} N_{PS}(x_1,\ldots,x_n)/N \\ &= S_{PS}(x_1,\ldots,x_n) \\ &= P(x_1 \ldots x_n) \end{aligned}$$

Cioè, stime derivate da PRIORSAMPLE sono consistenti

In breve: $\hat{P}(x_1,\ldots,x_n) \approx P(x_1 \ldots x_n)$

Rejection sampling

$\hat{\mathbf{P}}(X|\mathbf{e})$ stimate da campioni in accordo con $\mathbf{e}$

```
function REJECTION-SAMPLING(X, e, bn, N) returns an estimate of P(X|e)
   local variables: N, a vector of counts over X, initially zero
   for j = 1 to N do
       x ← PRIOR-SAMPLE(bn)
       if x is consistent with e then
           N[x] ← N[x]+1 where x is the value of X in x
   return NORMALIZE(N[X])
```

P.e., stimare $\mathbf{P}(Rain|Sprinkler = true)$ usando 100 campioni
   27 campioni hanno $Sprinkler = true$
      Di questi, 8 hanno $Rain = true$ e 19 hanno $Rain = false$.

$\hat{\mathbf{P}}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19\rangle) = \langle 0.296, 0.704\rangle$

Analisi di rejection sampling

$\hat{\mathbf{P}}(X|\mathbf{e}) = \alpha \mathbf{N}_{PS}(X, \mathbf{e})$ (def. algoritmo)
$= \mathbf{N}_{PS}(X, \mathbf{e})/N_{PS}(\mathbf{e})$ (normalizzato tramite $N_{PS}(\mathbf{e})$)
$\approx \mathbf{P}(X, \mathbf{e})/P(\mathbf{e})$ (properietà di PRIORSAMPLE)
$= \mathbf{P}(X|\mathbf{e})$ (def. di probabilità condizionale)

Quindi rejection sampling restituisce stime consistenti della prob. a posteriori

Problemi: costosissimo se $P(\mathbf{e})$ è piccola

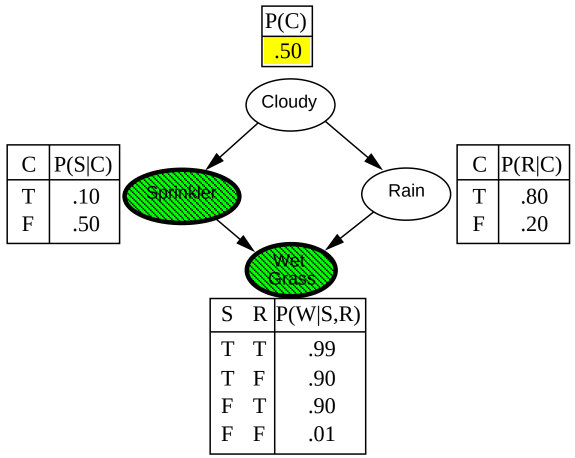
$P(\mathbf{e})$ converge esponenzialmente con il numero di variabili di evidenza!

Likelihood weighting

Idea: fissare le variabili di evidenza, campionare solo variabili non di evidenza, e pesare ogni campione con la likelihood accordata dall'evidenza

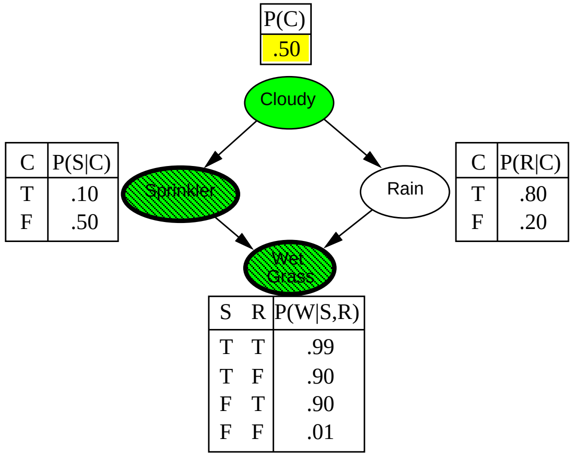
```
function LIKELIHOOD-WEIGHTING(X, e, bn, N) returns an estimate of P(X|e)
   local variables: W, a vector of weighted counts over X, initially zero
   for j = 1 to N do
       x, w ← WEIGHTED-SAMPLE(bn)
       W[x] ← W[x] + w where x is the value of X in x
   return NORMALIZE(W[X])
--------------------------------------------------------------------------
function WEIGHTED-SAMPLE(bn, e) returns an event and a weight
   x ← an event with n elements; w ← 1
   for i = 1 to n do
       if X_i has a value x_i in e
           then w ← w × P(X_i = x_i | Parents(X_i))
           else x_i ← a random sample from P(X_i | Parents(X_i))
   return x, w
```


Esempio di likelihood weighting



$w = 1.0$

Esempio di likelihood weighting



P(C)
.50

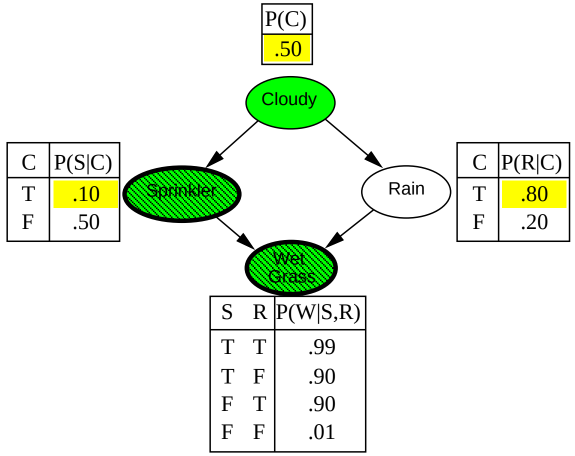
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Esempio di likelihood weighting



P(C)
.50

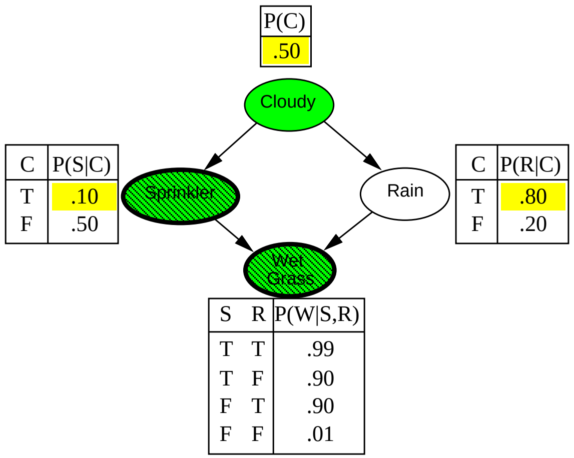
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0$

Esempio di likelihood weighting



P(C)
.50

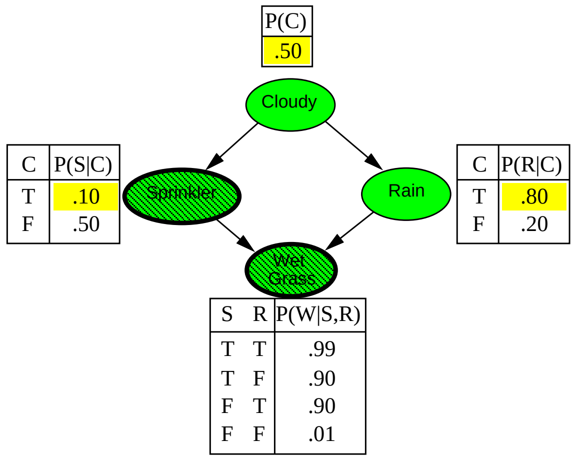
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$w = 1.0 \times 0.1$

Esempio di likelihood weighting



P(C)
.50

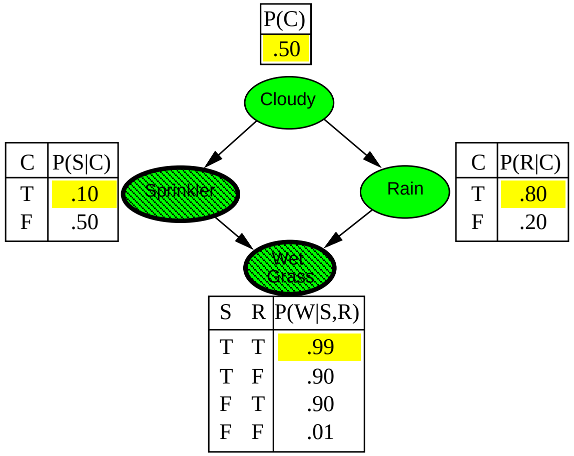
C	P(S|C)
T	.10
F	.50

C	P(R|C)
T	.80
F	.20

S	R	P(W|S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

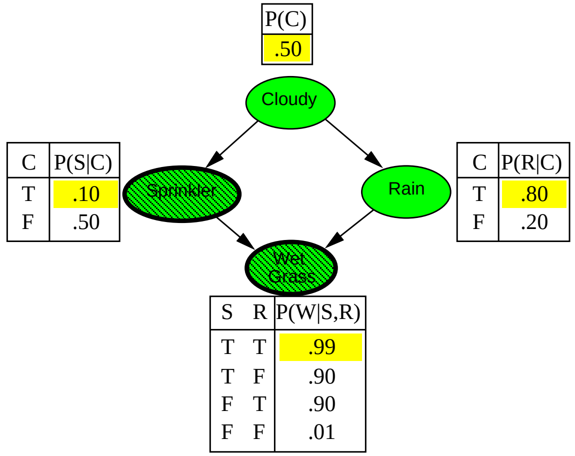
$w = 1.0 \times 0.1$

Esempio di likelihood weighting



$w = 1.0 \times 0.1$

Esempio di likelihood weighting



$w = 1.0 \times 0.1 \times 0.99 = 0.099$

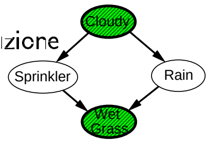
Analisi di likelihood weighting

La probabilità di campionamento per WEIGHTEDSAMPLE è

$S_{WS}(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{l} P(z_i | Parents(Z_i))$

Nota: pone attenzione solo all'evidenza negli **antenati**

$\Rightarrow$ da qualche parte "nel mezzo" fra la distribuzione a priori e quella a posteriori



Il peso per un dato campione $\mathbf{z}, \mathbf{e}$ è

$w(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{m} P(e_i | Parents(E_i))$

La probabilità di campionamento pesata è

$S_{WS}(\mathbf{z}, \mathbf{e}) w(\mathbf{z}, \mathbf{e})$

$= \prod_{i=1}^{l} P(z_i | Parents(Z_i)) \quad \prod_{i=1}^{m} P(e_i | Parents(E_i))$

$= P(\mathbf{z}, \mathbf{e})$ (per la semantica standard globale della rete)

Quindi likelihood weighting restituisce stime consistenti
però le prestazioni degradano con la presenza di tante variabili di evidenza
poiché pochi esempi hanno quasi tutto il peso totale

Riassunto

Inferenza esatta tramite l'eliminazione di variabile:
- polinomiale su polialberi, NP-hard in generale
- spazio = tempo, dipendente dalla topologia

Inferenza approssimata tramite LW:
- LW si comporta male quando c'è molta evidenza (soprattutto a "valle")
- LW in genere indipendente dalla topologia
- La convergenza può essere molto lenta per probabilità vicine a 1 o 0
- Può trattare combinazioni arbitrarie di variabili discrete e continue