9. Ragionamento su Modelli di Conoscenza Incerta

Dispensa ICon

versione: 06/12/2024, 00:23

Probabilità · Semantica della Probabilità · Indipendenza · Belief Network · Inferenza Probabilistica · Modelli Probabilistici Sequenziali · Simulazione Stocastica

1 Probabilità

Nel mondo reale spesso si devono prendere decisioni in base a informazioni che sono però *incomplete*. In tali casi è difficile capire l'*esatto* stato del mondo. Ad esempio un medico non conosce le esatte condizioni (interne) del paziente, un docente non sa con precisione quanto ciò che ha spiegato sia stato compreso dai discenti, un robot non conosce quanto sia successo in una stanza che ha lasciato da poco. Per prendere decisioni un KBS deve cercare di sfruttare *tutta* la conoscenza disponibile. Essendo tale conoscenza limitata il sistema dovrà essere in grado di ragionare sotto **incertezza**, facendo ipotesi e valutandone l'attendibilità.

Nel seguito si richiameranno concetti legati alla *probabilità* e alle assunzioni di *indipendenza* sulla rappresentazione del mondo per poi esaminare le modalità di *ragionamento* probabilistico su tali rappresentazioni.

Se non si può assumere una *conoscenza completa* del mondo, per prendere decisioni spesso si formulano diverse ipotesi.

> **Esempio** — *Cinture di sicurezza*: abbassano il rischio di danni gravi
>
> Casi (estremi) di mancato uso delle cinture: assumendo l'impossibilità di incidenti, non servirebbe usarle; assumendo, invece, di doverne avere sicuramente, non si userebbe l'auto.
>
> La decisione (compromesso a seconda dei casi) dipende dalla *possibilità* che occorra un incidente e dai relativi vantaggi e svantaggi, quali: l'*utilità* delle cinture in caso d'incidente, la *scomodità* del loro uso e l'importanza della *mobilità*.

L'**incertezza** può essere caratterizzata come **ontologica** o **epistemologica**. Nel primo caso essa riguarda il mondo in sé (si tratta di una forma di *imprecisione* nella rappresentazione), ad esempio *"persona molto alta"* rappresenta una conoscenza *vaga* (*fuzzy*) sul valore esatto dell'altezza. Nel secondo caso l'incertezza concerne le credenze sullo stato del mondo: essendo spesso le situazioni non completamente conosciute ci si possono comunque fare idee/valutazioni (*belief*) soggettive. Il problema è come aggiornare tale *credito* in presenza di nuova informazione disponibile? Si devono considerare valutazioni nei due momenti successivi: **a priori**, ossia prima di qualsiasi osservazione, e **a posteriori**, ossia dopo la scoperta di informazione, tipicamente da osservazioni, il che porta all'aggiornamento delle credenze sullo stato del mondo (così come modellato).

Il *ragionamento* in presenza di *incertezza* è un problema studiato in *Teoria della probabilità* e in *Teoria delle decisioni*. Si basa sul calcolo dell'*azzardo*, sulla valutazione dell'incertezza sulle conseguenze di decisioni da prendere.

Esempio — Lancio di un dado:

- A effettua il lancio, osservando ⚅, ma dice a B solo che il risultato è *pari*, mentre a C non viene detto nulla;
- a una diversa conoscenza riguardo uno stesso evento corrisponderanno diverse probabilità soggettive:
  - per A: $P_A(⚅) = 1$ (certezza dal risultato osservato);
  - per B: $P_B(⚅) = \frac{1}{3}$ (se disposto a credere ad A);
  - per C: $P_C(⚅) = \frac{1}{6}$ (ha la massima incertezza)

Si noti che tali credenze riguardano un lancio specifico non uno generico.

La misura della credibilità rientra nella prospettiva che si chiama **bayesiana** o **soggettiva**: lo studio di come la conoscenza impatti sulle credenze del soggetto. Soggetti diversi potranno assumere diverse probabilità. Tale prospettiva è diversa da quella *oggettiva* detta anche *frequentista*.

La *credibilità* di una proposizione α viene misurata (*convenzionalmente*) con un valore in $[0, 1]$: Si ha probabilità *nulla* se si crede che α sia del tutto falsa e che nessuna nuova evidenza possa cambiarla; si avrà una probabilità pari a 1 se α si ritiene assolutamente vera; altrimenti la probabilità sarà un valore in $]0, 1[$ che misura l'incertezza sulla credibilità della sua verità; ciò non significa che α sia solo parzialmente vera, bensì che la sua verità sia sconosciuta.

2 Semantica della Probabilità

Nozioni base:

- **variabile**: *aleatoria* / *casuale*, denotata con iniziale maiuscola e dotata di *dominio* di valori; per il momento si considereranno variabili: *booleane* con dominio $\{true, false\}$, denotando, per brevità, $X = true$ con x, ad esempio analogamente $fire$ per $Fire = true$; variabili *discrete* con dominio finito o almeno enumerabile;
- **mondo**: funzione che associa a ogni variabile un valore, viceversa una variabile può essere vista come una funzione dai mondi al suo dominio, ad esempio, sintomi, malattie, risultati di esami, nel tempo, per tutti i pazienti e il personale sanitario di un ospedale;
- **proposizione primitiva**: assegnazione di un valore a una variabile o diseguaglianza tra variabili e valori/variabili, come ad esempio: $A = true$, $X < 7$, $Y > Z$;
- **proposizione**: si costruisce applicando i *connettivi* logici a proposizioni primitive.

Dato l'insieme di *mondi possibili* Ω, *finito* (assumendo un numero finito di variabili), una **misura di probabilità** $\mu : \Omega \to \mathbb{R}_+$ verifica le seguenti proprietà:

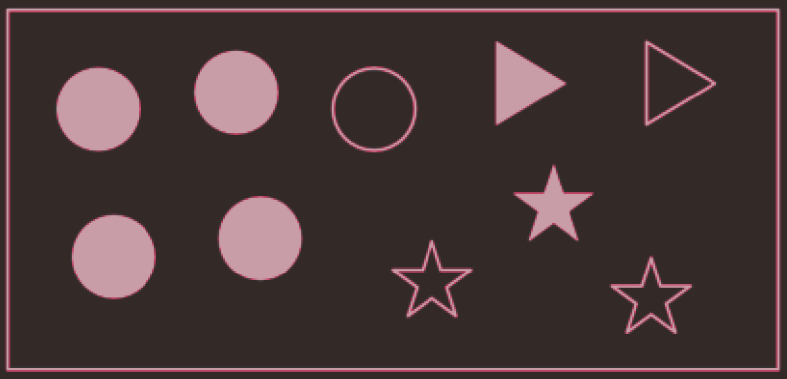
- $\mu(\omega_1 \cup \omega_2) = \mu(\omega_1) + \mu(\omega_2)$, se ω_1 e ω_2 disgiunti;
- $\mu(\Omega) = 1$, valore *convenzionale*.

La **probabilità** di una proposizione α sarà definita come

$$P(\alpha) = \mu(\{w \in \Omega : \ \alpha \text{ vera in } w\})$$

una misura coerente con la probabilità dei mondi.

Esempio — Si considerino i 10 mondi rappresentati in figura



- essi possono essere descritti dalle variabili:
  - *Shape* con dominio $\{circle, triangle, star\}$;
  - *Filled* booleana;
  - (posizione)
- se sono *equiprobabili* ossia se $\forall w\colon\ P(w) = 0.1$ (per tutti gli altri probabilità nulla) allora:
  - $P(Shape = circle) = 0.5$;
  - $P(Filled = false) = 0.4$;
  - $P(Shape = circle \wedge Filled = false) = 0.1$.

Data una variabile *discreta* X, la **distribuzione di probabilità** di X è la funzione funzione $P(X) : dom(X) \to \mathbb{R}$, definita ponendo $P(x)$ come probabilità della proposizione $X = x$, con $x \in dom(X)$.

Questa definizione può essere estesa al caso di un *insieme di variabili* $\{X_1, \dots, X_n\}$: la distribuzione $P(X_1, \dots, X_n)$ va dai valori assegnati alle variabili alla probabilità; ad esempio, per $P(X, Y)$, distribuzione su X e Y: dati $x \in dom(X)$ e $y \in dom(Y)$

$$P(X = x, Y = y) = P(X = x \wedge Y = y)$$

i.e. definisce la probabilità per la proposizione (congiuntiva) $X = x \wedge Y = y$. La **distribuzione di probabilità congiunta** per l'insieme di variabili $\{X_1, \dots, X_n\}$ è la distribuzione su tutti i mondi possibili (le proposizioni sono equivalenti alle assegnazioni totali).

Nel caso di variabili continue si devono considerare infiniti mondi, per la precisione un insieme non enumerabile di insiemi di mondi: in caso di dominio *infinito* di una variabile e numero *infinito* di variabili. Esso non può essere descritto con un linguaggio finito.

In realtà non serve una misura definita su *tutti* gli insiemi di mondi ma solo su quelli rappresentabili con formule logiche. Questo costituisce la base per la definizione di una cosiddetta σ-algebra.

Per una variabile X continua si definisce una *funzione* di **densità di probabilità** $p : \mathbb{R} \to \mathbb{R}_+$ con integrale *unitario*. La misura della probabilità su *intervalli* sarà

$$P(a \leq X \leq b) = \int_a^b p(X) dX$$

Anche se $\forall v : P(X = v) = 0$ è possibile avere $P(v_0 \leq X \leq v_1) > 0$, essendo $v_0 \leq v_1$.

Una distribuzioni (densità / probabilità) si dice:

- *parametrica*: se la formula ha un numero finito di parametri prefissati;
- *non parametrica*: se non si fissa il numero dei parametri, potenzialmente anche infiniti

Altro metodo: *discretizzazione* in un numero finito di mondi, ad esempio, altezze limitate approssimate al centimetro.

2.1 Probabilità Condizionata

Con la disponibilità di fatti nuovi si può operare un *aggiornamento* della credibilità di proposizioni.

La **probabilità condizionata** $P(h \mid e)$ di h *data* e è una misura del credito di h (*ipotesi*), assumendo la verità di e (*evidenza*). $P(h \mid e)$ si chiama anche probabilità **a posteriori** di h: e rappresenta la congiunzione di *osservazioni* dirette del mondo, *tutte* le osservazioni riguardanti una data situazione, non solo una selezione, per la correttezza della probabilità condizionata. $P(h)$ si dice anche probabilità **a priori** di h e corrisponde a $P(h \mid true)$, ossia precede qualunque osservazione.

> **Esempio** — diagnostica
>
> *Prima* di prendere in considerazione un dato paziente,
> si può usare la distribuzione di probabilità a priori sulle possibili malattie.
>
> *Poi* si raccoglie evidenza con visite, esami di lab, ecc. Le informazioni specifiche su un paziente costituiscono l'evidenza. Si aggiornano le probabilità per riflettere le nuove conoscenze e prendere decisioni informate.

> **Esempio** — robot consegne
>
> - evidenza raccolta via via dai sensori
> - se sono rumorosi, il robot può sbagliarsi su come sia fatto il mondo
>   - pur essendo consapevole dell'informazione ricevuta

2.1.1 Semantica della Probabilità Condizionata

L'evidenza e (proposizione) elimina i mondi incompatibili. Come nella conseguenza logica, essa seleziona i mondi possibili, quelli in cui e vera.

Si definisce una nuova misura μ_e su insiemi di mondi, data l'evidenza e:

$$\mu_e(S) = \begin{cases} 0 & e \text{ falsa in } \omega,\ \forall \omega \in S \\ c \cdot \mu(\omega) & e \text{ vera in } \omega,\ \forall \omega \in S \end{cases}$$

dove c è una *costante di normalizzazione* (dipende da e).

Ogni mondo in cui e sia falsa ha probabilità condizionata nulla. Per gli altri mondi, si considerano probabilità normalizzate: perché risulti una misura di probabilità, dato e:

$$\begin{aligned} 1 = \mu_e(\Omega) &\overset{\text{per casi}}{=} \mu_e(\{\omega \mid e \text{ vera in } \omega\}) + \mu_e(\{\omega \mid e \text{ falsa in } \omega\}) \\ &= c \cdot \mu(\{\omega \mid e \text{ vera in } \omega\}) + 0 = c \cdot P(e) \end{aligned}$$

per cui $c = 1/P(e)$, quindi la misura può essere definita solo se $P(e) > 0$; difatti $P(e) = 0$ indica che e è impossibile.

La **probabilità condizionata** della *proposizione* h data e può essere definita sommando le probabilità condizionate dei mondi in cui h è vera

$$\begin{aligned} P(h \mid e) &= \mu_e(\{\omega \mid h \text{ vera in } \omega\}) \\ &= \mu_e(\{\omega \mid h \wedge e \text{ vera in } \omega\}) + \mu_e(\{\omega \mid h \wedge \neg e \text{ vera in } \omega\}) \\ &= \frac{1}{P(e)} \cdot \mu(\{\omega \mid h \wedge e \text{ vera in } \omega\}) + 0 = \frac{P(h \wedge e)}{P(e)} \end{aligned}$$

(altrove questa formula viene assunta come definizione di $P(h \mid e)$, qui è stata ricavata da una misura più semplice).

Per estensione si definisce la **distribuzione di probabilità condizionata** $P(X|Y)$ in funzione di X e Y (insiemi di) variabili: dati $x \in dom(X)$ e $y \in dom(Y)$, si considera la probabilità condizionata delle proposizioni $P(X = x \mid Y = y)$.

Esempio — Considerano i mondi della figura precedente con probabilità pari a 0.1, data l'evidenza $Filled = false$, si hanno 4 mondi con probabilità a posteriori non nulla:

- $P(Shape = circle \mid Filled = false) = 0.25$
- $P(Shape = star \mid Filled = false) = 0.5$

Si considererà la *decomposizione* di una congiunta per semplificare i calcoli:

Proposizione (*chain rule*) — Date le proposizioni $\alpha_1, \ldots, \alpha_n$:

$$\begin{aligned} P(\alpha_1 \wedge \alpha_2 \wedge \cdots \wedge \alpha_n) = {} & P(\alpha_1) \cdot \\ & P(\alpha_2 \mid \alpha_1) \cdot \\ & P(\alpha_3 \mid \alpha_1 \wedge \alpha_2) \cdot \\ & \vdots \\ & P(\alpha_n \mid \alpha_1 \wedge \cdots \wedge \alpha_{n-1}) \\ = {} & \prod_{i=1}^{n} P(\alpha_i \mid \alpha_1 \wedge \cdots \wedge \alpha_{i-1}), \end{aligned}$$

La parte destra si annulla se è nullo uno dei prodotti, anche in caso alcuni fossero indefiniti; il caso base è $P(\alpha_1|true) = P(\alpha_1)$, con $true$ = congiunzione vuota.

2.1.2 Teorema di Bayes

Come *aggiornare* le probabilità in base a nuova evidenza?

Proposizione (*Regola di Bayes*[1])

Se $P(e) \neq 0$, si può scrivere:

$$P(h \mid e) = \frac{P(e \mid h) \cdot P(h)}{P(e)}$$

dove $P(h \mid e)$ è la probabilità **a posteriori**, $P(e \mid h)$ si chiama **verosimiglianza** o *likelihood* e $P(h)$ è la probabilità **a priori** di h.

Estendendo a distribuzioni condizionate da k, evidenza pregressa o conoscenza di fondo *implicita*, se $P(e \mid k) \neq 0$ si ha:

$$P(h \mid e \wedge k) = \frac{P(e \mid h \wedge k) \cdot P(h \mid k)}{P(e \mid k)}$$

ossia partendo da $P(h \mid k)$, data nuova evidenza e, si ha la nuova misura $P(h \mid e \wedge k)$.

In genere, una tra $P(e \mid h)$ o $P(h \mid e)$ può essere *stimata* (dai dati) più facilmente dell'altra, che potrà essere ricavata dal teorema.

Esempio — Diagnosi medica: osservati i sintomi (S) si vogliano determinare potenziali malattie (M):

- servirebbe calcolare $P(M \mid S)$, difficile perché dipende dal *contesto*, ad esempio alcune malattie sono più frequenti negli ospedali;
- tipicamente è più facile determinare $P(S \mid M)$, relazione che lega i sintomi alle malattie, che dipende meno dal contesto;
- le probabilità possono essere messe in relazione attraverso il teorema di Bayes, dove $P(M)$ rappresenta l'influenza del contesto.

Esempio ☿ — Diagnostica dell'impianto luci:

Si vuole sapere se l'*interruttore* s_1 funzioni. L'elettricista che ha realizzato l'impianto non può saperlo in anticipo, ma è in grado di specificare che l'output d'un interruttore dipende dal passaggio di corrente, dalla sua posizione e dal suo stato: la probabilità a priori che sia rotto dipende dal costruttore e dall'usura. Con il teorema di Bayes si può inferire lo stato, date probabilità a priori ed evidenza.

Esempio — Affidabilità degli impianti d'allarme antincendio: quanto è verosimile che in caso d'incendio ($Fire$) scatti l'allarme ($Alarm$)?

Probabilità di incendio dato l'allarme scattato (Bayes):

$$P(fire \mid alarm) = \frac{P(alarm \mid fire) \cdot P(fire)}{P(alarm)} = \\ = \frac{P(alarm \mid fire) \cdot P(fire)}{P(alarm \mid fire) \cdot P(fire) + P(alarm \mid \neg fire) \cdot P(\neg fire)}$$

- $P(alarm \mid fire)$ allarme scattato, noto che un incendio sia in atto;
- $P(fire)$ incendio (senza altre condizioni), misura di quanto l'edificio sia soggetto ad incendi;
- $P(alarm)$ allarme scattato (senza condizioni);
- $P(fire \mid alarm)$ incendio dato l'allarme, più difficile da rappresentare direttamente, ad esempio può dipendere dalla frequenza di atti vandalici nella zona.

Probabilità Condizionata e Implicazione ☌

Ci si può chiedere se valga sempre l'uguaglianza: $P(f \leftarrow e) \stackrel{?}{=} P(f \mid e)$

Ciò non è necessariamente vero. Si osservi che, per l'equivalenza $e \rightarrow f \equiv \neg e \vee f$, si ha che $P(f \leftarrow e) = P(f \vee \neg e)$ misura delle interpretazioni (mondi) per le quali f è vera oppure e è falsa. Un caso in cui l'uguaglianza non vale è il seguente:

- $P(oviparo \mid mammifero)$ risulta *bassa*, i mammiferi che depongono uova sono rari (e.g. l'ornitorinco);
- $P(oviparo \leftarrow mammifero) = P(oviparo \vee \neg mammifero)$ è *alta* essendo la proporzione dominata da animali non-mammiferi; per lo stesso motivo, anche $P(\neg oviparo \leftarrow mammifero)$ risulta *alta*.

2.2 Valori Attesi

Per le predizioni, si possono definire *medie ponderate* su tutti i mondi possibili dei valori di funzioni numeriche.

Data f, funzione sui mondi, il **valore atteso** di f rispetto a P è definito come segue:

$$\mathcal{E}_P(f) = \sum_{w \in \Omega} f(w) \cdot P(w)$$

Se f è definita (come un'interpretazione π) sulle proposizioni, ossia data α:

$$f(\alpha) = \begin{cases} 1 & \alpha \; vera \\ 0 & \alpha \; falsa \end{cases}$$

allora:

$$\mathcal{E}_P(f) = P(\alpha)$$

Esempi: f funzione che restituisca il valore di una variabile, il numero di bit usati per descrivere un mondo o una misura di gradimento.

Esempio — Domotica. Sia $\mathcal{E}_P(number\_of\_broken\_switches)$ il numero atteso di deviatori rotti, ossia il numero *medio*, nel lungo periodo, di deviatori non funzionanti secondo una certa distribuzione P.

Impianto con 3 deviatori in totale, ognuno con probabilità 0.7 di essere rotto, il numero atteso di deviatori rotti sarebbe:

$$0 \cdot (0.3^3) + 1 \cdot (3 \cdot 0.7 \cdot 0.3^2) + 2 \cdot (3 \cdot 0.7^2 \cdot 0.3) + 3 \cdot (0.7^3) = 2.01$$

- Nota: 3 casi/mondi con un deviatore rotto e gli altri funzionanti e 3 casi con due deviatori rotti e uno solo funzionante.

Nel caso di distribuzioni condizionate, si definisce il *valore atteso* **condizionato** di f data l'evidenza e:

$$\mathcal{E}(f \mid e) = \sum_{\omega \in \Omega} f(\omega) \cdot P(\omega \mid e)$$

Esempio — Valore atteso di deviatori rotti essendo l_1 spenta:

$$\mathcal{E}(number\_of\_broken\_switches \mid \neg lit(l_1))$$

calcolata ediando i numeri di interruttori rotti su tutti i mondi in cui l_1 è spenta.

Data una variabile booleana su $\{0, 1\}$, il *valore atteso* coincide con la probabilità della variabile.

Gli algoritmi per i valori attesi possono calcolare probabilità e i teoremi sui valori attesi si applicano anche alle distribuzioni di probabilità.

3 Indipendenza

Per definire distribuzioni su molte variabili serve molta conoscenza. Gli assiomi della probabilità risultano deboli per certi scopi: pongono pochi vincoli sulle probabilità condizionate da sfruttare, ad esempio, con n variabili binarie, ci saranno $2^n - 1$ probabilità da conoscere per avere una completa distribuzione dalla quale derivare quelle condizionate. Serve un DB molto grande per determinarle anche in forma approssimata.

Per limitare il quantitativo d'informazione richiesta, si può assumere che una variabile dipenda direttamente *solo* da alcune altre. Altre assunzioni dovrebbero far sì che si richiedano meno dati per specificare un modello: con strutture più semplici il ragionamento sarà svolto in modo più efficiente.

In generale $P(h \mid e) \in]0, 1[$ non determina vincoli sui valori di $P(h \mid f \wedge e)$, se non in casi-limite: $P(h \mid f \wedge e) = 1$ quando f implica h oppure $P(h \mid f \wedge e) = 0$ quando f implica $\neg h$.

Spesso è però disponibile un tipo comune di *conoscenza qualitativa* riassumibile con:

$$P(h \mid e) = P(h \mid f \wedge e)$$

cioè f risulta irrilevante per la probabilità di h data e.

Estendendo l'idea a un dato insieme di variabili Zs: X si dice *condizionatamente* **indipendente** da Y dato Zs sse

$$P(X \mid Y, Zs) = P(X \mid Zs)$$

cioè $\forall x \in dom(X)\ \forall y \in dom(Y)\ \forall z \in dom(Zs)$, se $P(Y = y \wedge Zs = z) > 0$ allora

$$P(X = x \mid Y = y \wedge Zs = z) = P(X = x \mid Zs = z)$$

ossia noto il valore di ogni variabile in Zs, conoscere il valore di Y non influenza la credibilità dell'assegnazione di un valore a X. L'indipendenza condizionata è spesso *facile* da accertare e torna utile nell'*inferenza*. È, invece, più raro disporre di una tabella delle probabilità dei mondi da cui verificare l'indipendenza per via numerica.

Esempio — Modello probabilistico per studenti ed esami

- si assume a priori che $Intelligent$ sia indipendente da $Works\_hard$: il fatto che si studi molto non dipende dall'intelligenza;
- le risposte date ($Answers$) potrebbero dipendere dall'assiduità nello studio e dall'intelligenza, quindi, noto il valore di $Answers$, $Intelligent$ e $Works\_hard$ diventano reciprocamente dipendenti; inoltre casi di risposte approfondite fornite senza aver studiato molto fanno aumentare il credito da attribuire all'intelligenza del/la candidato/a;
- il voto ($Grade$) dovrebbe dipendere dalle risposte, non da intelligenza o studio profuso quindi $Grade$ dovrebbe essere indipendente da $Intelligent$ (e $Works\_hard$) noto il valore di $Answers$;
- in mancanza delle risposte, $Intelligent$ influenzerebbe $Grade$, in generale: studenti più intelligenti forniscono risposte migliori quindi $Grade$ dipende da $Intelligent$ se non sono disponibili osservazioni.

Proposizione — Se le probabilità condizionate sono ben definite, i seguenti enunciati sono equivalenti:

1. X è condizionatamente indipendente da Y data Z;
2. Y è condizionatamente indipendente da X data Z;
3. $\forall x, y, y', z$:

$$P(X = x \mid Y = y \wedge Z = z) = P(X = x \mid Y = y' \wedge Z = z)$$

   i.e. noto il valore di Z, cambiare Y non influenza la credibilità del valore di X;
4. $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$.

X e Y sono *incondizionatamente* **indipendenti** se

$$P(X, Y) = P(X)P(Y)$$

ossia condizionatamente indipendenti ma senza osservazioni date: ciò non implica che siano pure condizionatamente indipendenti, data qualche altra evidenza Z. Raramente tale proprietà può essere determinata per via numerica da dati disponibili.

4 Belief Network

Poter fare assunzioni di indipendenza condizionata consente una rappresentazione concisa dei domini. Data una variabile X, solo alcune variabili influenzano *direttamente* il suo valore. L'insieme di variabili che la influenzano localmente si chiama *Markov blanket* (descritto più avanti) e tale località può essere sfruttata: date queste variabili, X sarà condizionatamente indipendente dalle altre. Le assunzioni di *indipendenza* condizionata inducono un *ordinamento* delle variabili che può essere rappresentato attraverso un *grafo orientato*[2].

Una *rete bayesiana* è un modello (a grafo) che evidenzia la dipendenza condizionata fra variabili. Definito un *ordinamento totale* sull'insieme delle sue variabili $\{X_1, \ldots, X_n\}$, la *distribuzione* di probabilità *congiunta* potrà essere decomposta in termini di probabilità condizionate, tramite la *chain rule*:

$$P(X_1 = v_1 \wedge X_2 = v_2 \wedge \cdots \wedge X_n = v_n) = \prod_{i=1}^{n} P(X_i = v_i \mid X_1 = v_1 \wedge \cdots \wedge X_{i-1} = v_{i-1})$$

ossia:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$$

I **genitori** di ciascuna variabile X_i saranno costituiti dal sottoinsieme minimale di *predecessori* di X_i nell'ordinamento totale, $parents(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$, tale che gli altri predecessori di X_i siano condizionatamente indipendenti da X_i dato $parents(X_i)$. Così X_i dipende dai genitori, ma è indipendente dagli altri predecessori:

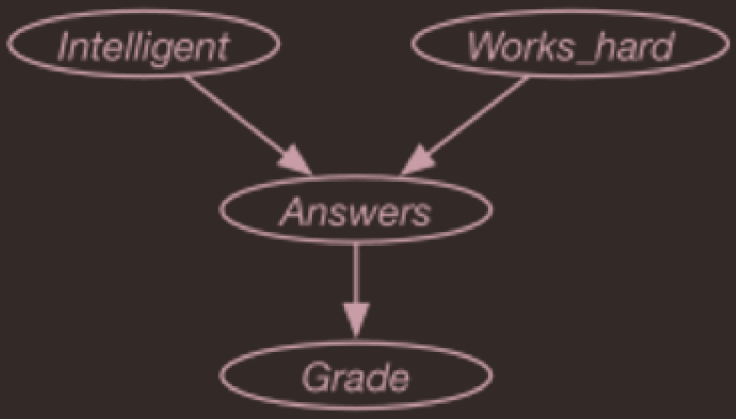
$$P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid parents(X_i))$$

Quando più insiemi minimali soddisfano tale condizione, allora si può operare una scelta casuale. Alcuni dei predecessori possono dipendere deterministicamente da altri. Riscrivendo la *chain rule*, la **fattorizzazione** della *congiunta* sarà:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

Formalmente, una **belief network** o **rete bayesiana** (BN) è rappresentata da un grafo aciclico orientato (DAG) con una variabile, X, (con un suo dominio) per ciascun nodo, archi verso X da ciascuno dei nodi relativi ai suoi genitori $parents(X)$ e quindi una distribuzione condizionata di probabilità (CPD) $P(X \mid parents(X))$ (a priori in caso di X senza genitori). Il grafo è *aciclico* per costruzione con una fattorizzazione che dipende dall'ordinamento (i genitori vanno scelti tra i predecessori): a diversi ordinamenti corrisponderanno diverse BN (alcuni ordinamenti portano a reti con meno archi). Una BN codifica relazioni di *indipendenza* condizionata fra variabili: dati i genitori, ogni variabile è indipendente da tutte le variabili che non sono sue discendenti.

Esempio — Dato l'ordinamento $Intelligent$, $Works\_hard$, $Answers$, $Grade$ delle variabili in un esempio precedente:



- $Intelligent$ non ha predecessori, quindi nemmeno genitori;
- $Works\_hard$ è indipendente da $Intelligent$ e $parents(Works\_hard) = \emptyset$;
- $Answers$ dipende da $Intelligent$ e $Works\_hard$:
  $parents(Answers) = \{Intelligent, Works\_hard\}$;
- $Grade$ è indipendente da $Intelligent$ e $Works\_hard$ data $Answers$:
  $parents(Grade) = \{Answers\}$;

- la fattorizzazione della distribuzione congiunta è
$P(Intelligent, Works\_hard, Answers, Grade) =$
$P(Intelligent) \cdot P(Works\_hard)$
$\cdot P(Answers \mid Intelligent, Works\_hard)$
$\cdot P(Grade \mid Answers)$;
- per le risposte si possono considerare domini semplificati, come ad esempio $Answers = \{insightful, clear, superficial, vacuous\}$, oppure anche il testo stesso delle risposte.

4.1 Osservazioni e Query

Data una BN che specifica una distribuzione congiunta, il problema di *inferenza probabilistica* più comune è il calcolo della *distribuzione a posteriori* di (alcune) **variabili di query** data l'*evidenza*, ossia la congiunzione di assegnazioni di valori osservate riguardanti altre variabili.

Esempio — Prima delle osservazioni, $P(Intelligent)$ fornita dalla BN, dopo si può usare l'*inferenza* per determinare $P(Grade)$:

- query $P(Intelligent \mid Grade = \texttt{A})$: distribuzione a posteriori, ossia saputo che il voto conseguito è `A` (un valore di $Grade$);
- query $P(Intelligent \mid Grade = \texttt{A} \wedge Works\_hard = false)$: distribuzione a posteriori, noto pure che $Works\_hard = false$;
- $Intelligent$ e $Works\_hard$ sono indipendenti quando non siano date altre informazioni ma, se si conosce il voto, esse diventano *dipendenti*; ciò spiega perché ci si vanti se si hanno buoni voti senza studiare molto: aumenta la probabilità di risultare intelligenti.

Le dipendenze modellate con le CPD non impongono la direzione da utilizzare nelle inferenze per rispondere a query. Per ricavare le probabilità a posteriori necessarie a rispondere alle query, anche in verso opposto alle dipendenze della BN, si possono usare i teoremi della Teoria della probabilità, in particolare quello di Bayes.

4.2 Costruzione di Belief Network

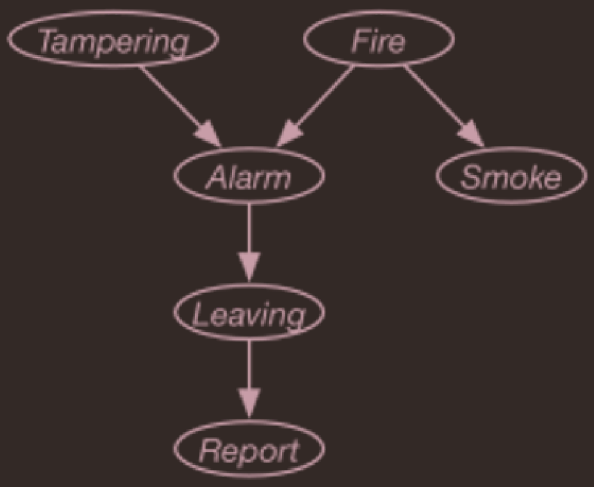
Per modellare un dominio attraverso una BN occorre:

- individuare le *variabili rilevanti*: cosa può essere osservato del dominio, ogni feature osservata costituirà una *variabile osservata* utile a potere poi condizionare le query sulla base dei valori assegnati; le feature che costituiranno le *variabili di query* ossia le informazioni di cui interessi determinare la probabilità a posteriori; potrebbero essere individuate anche altre variabili *nascoste* o **latenti**, né osservate né di query, utili a considerare dipendenze indirette e a semplificare il modello, perché vi saranno meno probabilità condizionate da specificare.
- decidere i *domini di valori*, ossia il livello di dettaglio del ragionamento che porterà a rispondere alle query. Per ogni variabile, si specifica il significato di ogni valore in ottemperanza al **principio di chiarezza** che richiede l'onniscenza del sistema riguardo il valore di ogni variabile, ossia sappia cosa deve accadere perché una variabile (non latente) assuma un dato valore; va quindi documentato il significato delle variabili e dei loro valori (escludendo quelle latenti i cui valori potranno in seguito essere *appresi* dai dati).
- specificare le *relazioni* di dipendenza diretta tra le variabili, ossia gli archi da considerare per definire la funzione $parent$.
- definire le CPD per ognuna variabile in dipendenza *dai genitori*, solitamente in forma di tabelle (*CPT*).

Esempio — Diagnostica di allarme anti-incendio che tiene anche conto di manomissioni, rumore dei sensori di rilevamento (fumo, uscita persone) che portino a informazioni contraddittorie sulla situazione. In particolare, *Report* segnala che è in corso un'evacuazione ma è una variabile *rumorosa* perché a volte non è in corso un'evacuazione segnalata (*falso positivo*) o altre volte potrebbe non segnalare un'evacuazione davvero in corso (*falso negativo*); l'evacuazione dipende dal fatto che sia scattato l'*allarme*; una *manomissione* o un *incendio* possono influenzare l'attivazione dell'allarme: la segnalazione di *fumo* dipende solo dalla rilevazione dell'incendio.

Modello risultante:

- *variabili* booleane (elenco ordinato)
  - *Tampering* vera sse c'è stata la *manomissione* dell'impianto;
  - *Fire* vera sse c'è un *incendio*;
  - *Alarm* vera sse si sente l'*allarme*;
  - *Smoke* vera sse c'è *fumo*;
  - *Leaving* vera sse la gente *lascia* in massa l'edificio;
  - *Report* vera sse si invia un *avviso* di abbandono dell'edificio in corso;
- relazioni di *indipendenza condizionata*:
  - *Fire* indipendente da *Tampering* (senza condizioni);
  - *Alarm* dipende solo da *Fire* e *Tampering*, senza altre dipendenze dalle precedenti;
  - *Smoke* dipende solo da *Fire* ed è indipendente da *Tampering* e *Alarm* data *Fire*;
  - *Leaving* dipende solo da *Alarm* e non direttamente da *Fire*, *Tampering* o *Smoke*, ossia, data *Alarm*, è indipendente dalle altre variabili;
  - *Report* dipende direttamente solo da *Leaving*;

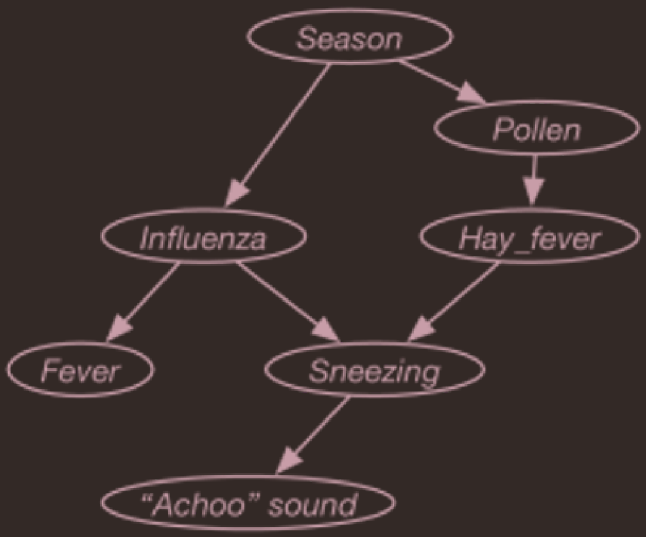


- *fattorizzazione* risultante:
  $P(Tampering, Fire, Alarm, Smoke, Leaving, Report) =$
  $P(Tampering) \cdot P(Fire)$
  $\cdot P(Alarm \mid Tampering, Fire) \cdot P(Smoke \mid Fire)$
  $\cdot P(Leaving \mid Alarm) \cdot P(Report \mid Leaving)$
  NB l'allarme non è sensibile al fumo ma al calore sprigionato dal fuoco: da qui l'indipendenza di *Alarm* da *Smoke* dato *Fire* e la dipendenza di *Smoke* da *Fire*.
- *domini*
  - variabili booleane: si userà il nome in minuscolo per denotare l'assegnazione di *true*, preceduto dalla negazione $\neg$ nel caso opposto, ad esempio: $Tampering = true$ si abbrevia con $tampering$, e $Tampering = false$ con $\neg tampering$

- specifica delle *probabilità condizionate* (CPD) usate negli esempi che seguiranno
  - $P(tampering) = 0.02$;
  - $P(fire) = 0.01$;
  - $P(alarm \mid fire \wedge tampering) = 0.5$;
  - $P(alarm \mid fire \wedge \neg tampering) = 0.99$;
  - $P(alarm \mid \neg fire \wedge tampering) = 0.85$;
  - $P(alarm \mid \neg fire \wedge \neg tampering) = 0.0001$;
  - $P(smoke \mid fire) = 0.9$;
  - $P(smoke \mid \neg fire) = 0.01$;
  - $P(leaving \mid alarm) = 0.88$;
  - $P(leaving \mid \neg alarm) = 0.001$;
  - $P(report \mid leaving) = 0.75$;
  - $P(report \mid \neg leaving) = 0.01$.
- le probabilità delle assegnazioni negative si possono ottenere per complemento, ad esempio:
  - $P(\neg smoke \mid fire) = 1 - P(smoke \mid fire) = 0.1$
  - $P(\neg smoke \mid \neg fire) = 1 - P(smoke \mid \neg fire) = 0.99$;
- probabilità a priori prima di ogni osservazione (evidenza):
  - $P(tampering) = 0.02$;
  - $P(fire) = 0.01$;
  - $P(smoke) = P(smoke \wedge fire) + P(smoke \wedge \neg fire)$
    $= P(smoke|fire)P(fire) + P(smoke|\neg fire)P(\neg fire) = 0.9 \cdot 0.01 + 0.01 \cdot 0.99$
    $= 0.009 + 0.0099 = 0.0189$;
  - $P(report) = 0.028$.
- avendo osservato l'arrivo di un *rapporto* (ossia data l'evidenza $report$) si ha che:
  - $P(tampering \mid report) = 0.399$;
  - $P(fire \mid report) = 0.2305$;
  - $P(smoke \mid report) = 0.215$
    si noti che le probabilità di $tampering$ e $fire$ sono aumentate a causa di $report$ quindi anche quella di $smoke$ che dipende da $fire$.
- avendo osservato (solo) del *fumo* (i.e. $smoke$):
  - $P(tampering \mid smoke) = 0.02$,
    ossia la probabilità di manomissione non è influenzata dall'osservazione di fumo;
  - $P(fire \mid smoke) = 0.476$ e $P(report \mid smoke) = 0.320$,
    si noti come entrambe aumentino dopo aver osservato del fumo;
- avendo osservato $report$ e $smoke$:
  - $P(tampering \mid report \wedge smoke) = 0.0284$,
    il caso di incendio diventa ancora più probabile;
  - $P(fire \mid report \wedge smoke) = 0.964$
    quindi se è pervenuto un rapporto, la presenza di fumo rende la manomissione meno probabile, essendo $report$ *spiegabile* da $fire$ che adesso è molto più probabile;
- ancora, avendo osservato $report$ ma non $smoke$ si ha che
  - $P(tampering \mid report \wedge \neg smoke) = 0.501$;
  - $P(fire \mid report \wedge \neg smoke) = 0.0294$,
    quindi osservato $report$, $fire$ diventa molto meno verosimile e la probabilità di $tampering$ aumenta per poter spiegare quanto osservato.

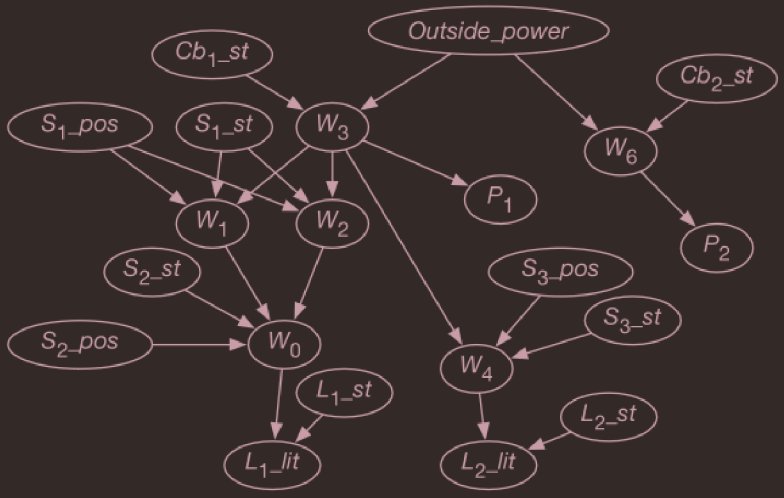
Esempio — Diagnosi dello *starnuto*: potrebbe dipendere dall'influenza o da febbre da fieno, variabili non indipendenti ma correlate alla stagione:

- la febbre da fieno con il quantitativo di pollini nella stagione;
- la febbre dipende direttamente dall'influenza;
- lo starnuto viene avvertito dal *suono*.



Esempio — Impianto elettrico.

Variabili per: accensione delle luci, posizioni degli deviatori, lo stato di guasto di luci e deviatori, il passaggio di corrente nei cavi. BN in figura:



4.3 Probabilità Condizionate e Fattori

Un **fattore** è una funzione definita su un insieme di variabili. Il suo **ambito** / *scope* è costituito da tali variabili e data un'assegnazione a ciascuna di esse, il fattore calcolerà un valore numerico (una probabilità).

Nel caso della *probabilità condizionata* si prenderà in considerazione il fattore $P(Y \mid X_1, \ldots, X_k)$, funzione da $Y, X_1, \ldots, X_k$ in $\mathbb{R}_+$ con il vincolo:

$$\forall x_1 \ldots \forall x_k \sum_{y \in dom(Y)} P(Y = y \mid X_1 = x_1, \ldots, X_k = x_k) = 1$$

Normalmente una distribuzione condizionata si rappresenta con una CPT. Questo è possibile con insiemi finiti di variabili e domini di valori.

Una possibile rappresentazione per una tabella multidimensionale per la CPD $P(Y = y \mid X_1 = v_1, \ldots, x_k = v_k)$ nella notazione Python per array e dizionari potrebbe essere la seguente: `p[v_1]...[v_k][y]`.

Esempio — Robot con azioni possibili indicate dalla variabile $Action$, $dom(Action) = \{go\_out, get\_coffee\}$.

Si considera lo stato di *bagnato* (indicato dalla variabile Wet) causato dalla pioggia (indicato da $Rain$) nel contesto di un'uscita da un edificio ($Action = go\_out$) o per colpa del trasporto $Action = get\_coffee$ d'una tazza piena ($Full$) di caffè: Wet sarà indipendente da $Rain$ quando $Action = get\_coffee$, ma dipendente da $Rain$ se $Action = go\_out$; Wet sarà indipendente da $Full$ dato $Action = go\_out$, ma dipendente da $Full$ dato $Action = get\_coffee$.

Una CPT corrispondente può essere:

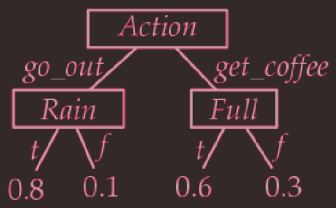
$Action$	$Rain$	$Full$	$P(Wet = true \mid Action, Rain, Full)$
go_out	false	false	0.1
go_out	false	true	0.1
go_out	true	false	0.8
go_out	true	true	0.8
get_coffee	false	false	0.3
get_coffee	false	true	0.6
get_coffee	true	false	0.3
get_coffee	true	true	0.6

Nel seguito si vedranno diverse possibili rappresentazioni per CPD/fattori.

4.3.1 Alberi di Decisione

In caso di CPT molto grandi ci saranno troppe probabilità da determinare. A tale scopo conviene sfruttare la *struttura* della rete. Considerando l'**indipendenza specifica per il contesto**, si sfrutta l'indipendenza condizionata al fine di definire un albero di decisione:

Esempio — La CPD dell'esempio precedente può essere riassunta attraverso l'albero in figura:



4.3.2 Sistema Deterministico con Input Rumorosi

Si può descrivere un fattore anche attraverso un sistema deterministico rappresentato da una formula logica ovvero anche da un programma avente come *input* variabili genitrici e input stocastici, dette **variabili rumore** o **esogene**, fra loro indipendenti (senza condizionamento), mentre le **variabili endogene**, quelle interne alla BN, potranno essere specificate deterministicamente in funzione di quelle esogene e delle genitrici.

Esempio — Nel caso precedente relativo al robot, si può ridefinire il modello attraverso la *formula*:

$$\begin{aligned} wet \leftrightarrow\ & (\ (go\_out \wedge rain \wedge n_0) \vee \\ & (go\_out \wedge \neg rain \wedge n_1) \vee \\ & (\neg go\_out \wedge full \wedge n_2) \vee \\ & (\neg go\_out \wedge \neg full \wedge n_3)\ ) \end{aligned}$$

dove le n_i sono variabili-rumore indipendenti, con $P(n_0) = 0.8$, $P(n_1) = 0.1$, $P(n_2) = 0.6$, $P(n_3) = 0.3$, ad esempio se $go\_out = true$ e $rain = false$, allora $wet = true$ quando $n_1 = true$, il che accade con probabilità 0.1.

Modello come *programma*:

```
if go_out:
   if rain:
      wet := flip(0.8)
   else:
      wet := flip(0.1)
else:
   if full:
      wet := flip(0.6)
   else:
      wet := flip(0.3)

flip(x) = (random() < x)
```

dove: `random()` restituisce un numero casuale nell'intervallo $[0, 1)$ e `flip(x)` crea una nuova variabile aleatoria indipendente, vera con probabilità x, ossia una n_i nella formula logica.

Come sistema deterministico si usa:

- il completamento di Clark di un programma logico nella **programmazione logica probabilistica** (*probabilistic logic programming*);
- un programma nella **programmazione probabilistica** (*probabilistic programming*) [PP];
- una formula logica nel **conteggio pesato del modello** (*weighted model counting*).

4.3.3 Noisy–Or

Si possono modellare i fattori come casi in cui qualcosa si avvera a causa di qualcos'altro: ad esempio si può avvertire la presenza di un sintomo a causa di una malattia.

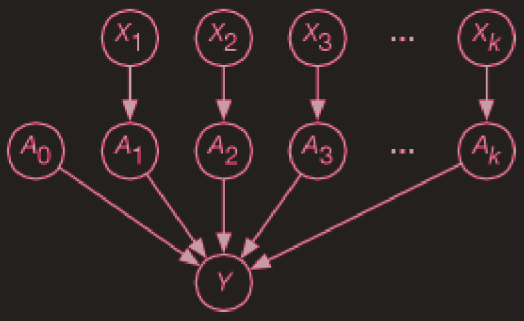
Il **noisy-or** è un modello di probabilità condizionata in cui una variabile figlia è vera se si *attiva* almeno una delle variabili genitrici (ciascuna con una data probabilità). Sia Y la figlia definita come *disgiunzione* delle attivazioni delle genitrici $X_1, \ldots, X_k$ (booleane). I *parametri* $w_0, \ldots, w_k$ costituiscono un sistema deterministico con input rumorosi, in cui:

$$Y \equiv n_0 \vee (n_1 \wedge x_1) \vee \ldots \vee (n_k \wedge x_k)$$

con n_i variabili di rumore indipendenti, dove $P(n_i) = w_i$ e ogni x_i per indicare che $X_i = true$.

La BN per $P(Y \mid X_1, X_2, \ldots, X_k)$ può essere definita introducendo le variabili booleane $A_0, A_1, \ldots, A_k$ con:

- $P(A_i = true \mid X_i = true) = w_i$ e $P(A_i = true \mid X_i = false) = 0$;
- $P(A_0 = true) = w_0$: si osservi che $P(Y \mid A_0, A_1, \ldots, A_k) = 1$ se almeno una delle A_i è vera (nulla se sono tutte false) quindi w_0 è la probabilità di Y quando tutte le X_i sono false.



La probabilità di Y aumenta in base alla numerosità di X_i vere.

Esempio — robot visto in precedenza

Si considerano le probabilità che il robot sia bagnato a causa della pioggia, se piove, a causa del caffè, se prende il caffè e che ciò possa essere dovuto ad altre cause

Si avrà una disgiunzione di cause, con:

- $P(wet\_from\_rain \mid rain) = 0.3$;
- $P(wet\_from\_coffee \mid coffee) = 0.2$;
- $P(wet\_for\_other\_reasons) = 0.1$ termine bias.

4.3.4 Modelli Log–lineari e Regressione Logistica

Un modello log-lineare specifica probabilità, non-normalizzate, tramite un prodotto di termini che andrà normalizzato per inferire probabilità. I termini sono tutti positivi quindi si può scrivere il prodotto come esponenziale di una somma, più facile da trattare.

Date h ed e booleane:

$$P(h \mid e) = \frac{P(h \wedge e)}{P(e)} = \frac{P(h \wedge e)}{P(h \wedge e) + P(\neg h \wedge e)} = \frac{1}{1 + \frac{P(\neg h \wedge e)}{P(h \wedge e)}}$$
$$= \frac{1}{1 + \exp\left(-\log \frac{P(h \wedge e)}{P(\neg h \wedge e)}\right)} = \text{sigmoid}(\log \text{odds}(h \mid e))$$

ricordando che $\text{sigmoid}(x) = 1/(1 + \exp(-x))$ e i rapporti fra probabilità detti **odds condizionati** sono definiti nel seguente modo:

$$\text{odds}(h \mid e) = \frac{P(h \wedge e)}{P(\neg h \wedge e)} = \frac{P(h \mid e)}{P(\neg h \mid e)} = \frac{P(h \mid e)}{1 - P(h \mid e)} = \frac{P(e \mid h)}{P(e \mid \neg h)} \cdot \frac{P(h)}{P(\neg h)}$$

Qui $\frac{P(e|h)}{P(e|\neg h)}$ si chiama anche **likelihood ratio** e $\frac{P(h)}{P(\neg h)} = \frac{P(h)}{1-P(h)}$ sono detti **prior odds**. Se $\frac{P(e|h)}{P(e|\neg h)}$ si può scrivere come prodotto allora il suo logaritmo sarà una somma.

Si può considerare anche un modello della probabilità condizionata $P(Y \mid X_1, \ldots, X_k)$ basato sulla *regressione logistica* definito come segue:

$$P(Y = true \mid X_1, \ldots, X_k) = \text{sigmoid}\left(\sum_i w_i \cdot X_i\right)$$

Assumendo un input fittizio aggiuntivo $X_0 = 1$ con termine noto w_0, esso corrisponde a una decomposizione della probabilità condizionata, con una likelihood ratio pari a un prodotto di termini, uno per ogni X_i. Si noti che $P(Y \mid X_1 = 0, \ldots, X_k = 0) = \text{sigmoid}(w_0)$ quindi w_0 corrisponde alla probabilità del caso in cui tutti i suoi genitori sono nulli mentre ogni w_i specifica un valore da aggiungere al cambiare di X_i: $P(Y \mid X_1 = 0, \ldots, X_i = 1, \ldots, X_k = 0) = \text{sigmoid}(w_0 + w_i)$. Assumendo l'esclusiva influenza sul figlio di ciascun genitore, gli odds sono decomponibili in prodotti di termini dipendenti da una sola variabile.

Esempio — (Robot) si supponga che:

$$P(wet \mid Rain, Coffee, Kids, Coat) = \text{sigmoid}(-1.0 + 2.0 \cdot Rain + 1.0 \cdot Coffee + 0.5 \cdot Kids - 1.5 \cdot Coat)$$

Si possono calcolare le seguenti probabilità:

- $P(wet \mid \neg rain \wedge \neg coffee \wedge \neg kids \wedge \neg coat) = \text{sigmoid}(-1.0) = 0.27$
- $P(wet \mid rain \wedge \neg coffee \wedge \neg kids \wedge \neg coat) = \text{sigmoid}(1.0) = 0.73$
- $P(wet \mid rain \wedge \neg coffee \wedge \neg kids \wedge coat) = \text{sigmoid}(-0.5) = 0.38$

Servono meno parametri dei 2^4=16 richiesti da una CPT, ma più assunzioni di indipendenza. Tale modello è simile al *Noisy-or* ma in quest'ultimo una variabile è vera a causa di *uno* dei genitori, mentre nella regressione logistica le influenze sul figlio dei genitori si *sommano*. Si veda lo specchietto in [PM23].

5 Inferenza Probabilistica

Scopo dell'inferenza è quello di rispondere a query su un dato modello probabilistico essenzialmente attraverso il calcolo di distribuzioni (possibilmente in modo efficiente). Un tipico task è quello di calcolare la **distribuzione a posteriori** di una o più variabili di query data l'evidenza disponibile. Il problema risulta complesso. Si può dimostrare che sia NP-hard[3] (nel caso in cui si richiedano soluzioni entro limiti dati) e #NP (sharp NP) per il calcolo della probabilità a posteriori o a priori di una variabile. Ciò nel caso pessimo, ma in genere si può sfruttare la struttura per rendere il compito più agevole.

Gli *approcci* principali all'*inferenza probabilistica* con le BN si caratterizzano in termini di inferenza *esatta* ed *approssimata*.

Inferenza Esatta

Tale forma di inferenza richiede che le probabilità vengano calcolate *precisamente*. Fondamentalmente gli approcci relativi a tale forma si dividono in due modalità principali:

1. *Enumerazione* dei mondi consistenti (i.e. coerenti) con l'evidenza;
2. Sfruttamento della struttura della rete: ad esempio l'*algoritmo di eliminazione delle variabili*, basato sulla *programmazione dinamica*, sfrutta le relazioni di indipendenza condizionata.

Si vedano in [PM23] (per eventuali progetti sull'argomento):

- Eliminazione di Variabili per BN, simile all'algoritmo per CSP;
- altri algoritmi.

Inferenza Approssimata

Esistono diversi metodi per stimare le probabilità in base al tipo di approssimazione:

- metodi che forniscono **limiti garantiti** $[l, u]$ di variazione entro i quali ricada la probabilità esatta p, ad esempio un algoritmo *anytime* garantisce che l e u tendano ad avvicinarsi reciprocamente col passare del tempo (o con più spazio a disposizione);

- metodi che forniscono **limiti probabilistici** sull'errore garantendo un errore contenuto (ad esempio 0.1) in un'alta percentuale di casi (ad esempio 95%), ovvero stime di probabilità che convergono nel tempo verso quella esatta e una certa velocità di convergenza, come ad esempio gli algoritmi di *simulazione stocastica* descritti nel seguito;
- metodi di **inferenza variazionale** capaci di fornire generalmente buone *approssimazioni*: scelta una *classe* di rappresentazioni più *semplici* (in termini di complessità), ad esempio BN non connesse (senza archi), si trova in tale classe il modello *più vicino* al problema originario ossia una distribuzione a posteriori, *facile* da calcolare, più vicina a quella cercata in modo da trasformare il problema in uno di di minimizzazione dell'errore, seguito da quello di inferenza vera e propria.

6 Modelli Probabilistici Sequenziali

Modelli Probabilistici Sequenziali

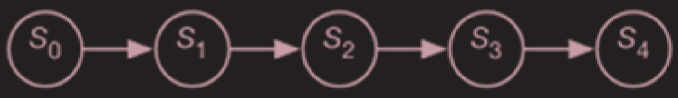
Si considerino BN speciali con una struttura che *si ripete*. Essi rendono possibile il ragionamento su *sequenze* non solo *temporali* ma anche nello spazio come, ad esempio, le parole in una frase. È possibile estendere le strutture anche a un numero *illimitato* di variabili casuali.

Nel seguito si richiameranno brevemente alcuni modelli fra i più usati.

6.1 Catene di Markov

Una **catena di Markov** (MC) è una BN con nodi/variabili disposti in una *sequenza*, quindi ogni variabile dipenderà direttamente solo dalla precedente.

Tramite queste strutture è possibile rappresentare sequenze di (valori o) *stati* (tratti da uno spazio finito o almeno enumerabile). Ad esempio si possono considerare sequenze di stati in un sistema dinamico o di parole in un testo (PageRank si basa su questo modello; cfr. specchietto nel testo [PM23]). Ogni *posizione* nella sequenza sarà indicato anche come **stage**.



(frammento di) MC come BN: può estendersi indefinitamente

Le sequenze sono spesso intese come successioni *nel tempo*. Essendo BN in cui ogni nodo, tranne il primo, dipende dal solo genitore, si dice che viene fatta una **assunzione di Markov** (mancanza di memoria, o *memorylessness*) che può essere descritta dalla seguente proprietà della distribuzione

$$P(S_{i+1} \mid S_0, \ldots, S_i) = P(S_{i+1} \mid S_i)$$

dove S_t rappresenta lo **stato** al tempo t: intuitivamente, S_t porta con sé tutte le informazioni *storiche* che possono influenzare gli stati futuri. Si dice anche che:

> *"il futuro è condizionatamente indipendente dal passato dato il presente"*

Una MC è un **modello stazionario** (i.e. *omogeneo nel tempo*) se tutte le variabili condividono un unico dominio e stesse probabilità di transizione per ogni stage:

$$\forall i \geq 0 : P(S_{i+1} \mid S_i) = P(S_1 \mid S_0)$$

il che la rende definibile attraverso da due sole distribuzioni:

- $P(S_0)$ che specifica le condizioni iniziali;
- $P(S_{i+1} \mid S_i)$ che specifica le *dinamiche*, le stesse per ogni $i \geq 0$ (la dinamica del mondo non cambia nel tempo).

Quindi con pochi parametri si specifica una struttura potenzialmente *non limitata*. Sarà poi possibile porre domande su punti arbitrari nel futuro o nel passato.

Per determinare le distribuzioni:

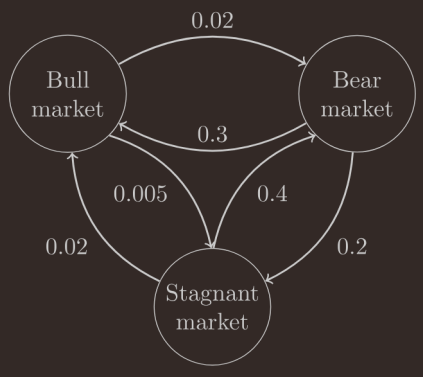
- per $P(S_i)$: si può usare l'*eliminazione di variabili* sommando su tutte le precedenti (successive irrilevanti);
- per $P(S_i \mid S_k)$: se $i > k$, basta considerare solo le variabili tra S_i e S_k, altrimenti solo quelle di indice inferiore a k.

La **distribuzione** di una MC si dice **stazionaria** se ciò che vale una volta varrà anche per la successiva: P stazionaria se per ogni stato s si ha che $P(S_{i+1} = s) = P(S_i = s)$, per cui:

$$P(S_i = s) = \sum_{s_i} P(S_{i+1} = s \mid S_i) \cdot P(S_i)$$

Esempio — Catene generate da un *processo markoviano* relativo al mercato finanziario

- spazio stati: $\{\text{Bull}, \text{Bear}, \text{Stagnant}\}$;
- probabilità delle transizioni nel diagramma che segue (ovvero anche attraverso matrice):



da Wikipedia

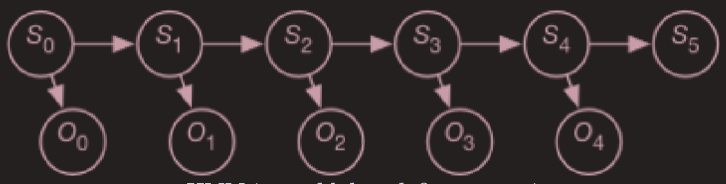
distribuzione stazionaria: $\pi = \langle 0.885, 0.071, 0.044 \rangle$.

Proprietà delle MC

- MC **ergodica**: per ogni (s_1, s_2) probabilità non nulla di raggiungere s_2 da s_1;
- MC **periodica**: $\exists p$, *periodo*, tale che la differenza tra (due) passaggi per uno stesso stato divisibile per p, altrimenti **aperiodica** ($p = 1$), ad esempio, una MC con uno stato al giorno tenendo conto anche del giorno della settimana uno stato il cui giorno sia lunedì sarà seguito da uno che cade di martedì... sarà *periodica* con $p = 7$;
- Una MC ergodica e aperiodica ha una sola **distribuzione** *stazionaria* detta di **equilibrio** raggiungibile a partire da qualunque stato: per ogni distribuzione su S_0, la distribuzione su S_i si avvicinerà sempre di più a quella di equilibrio al crescere di i.

6.2 Modelli di Markov Nascosti

Un **modello di Markov nascosto** o *Hidden Markov Model* (HMM) è costituito da una catena di Markov estesa con l'aggiunta di nodi per le *osservazioni* per ciascuno stage, come in figura:



HMM (estendibile indefinitamente)

Oltre alle transizioni di stato, per ogni istante di tempo t c'è l'osservazione O_t (variabile) che dipende da S_t (e da t) con *dominio* costituito dall'insieme delle possibili osservazioni; tali osservazioni sono *parziali*, cioè stati diversi possono essere mappati su una stessa osservazione e *rumorose*, ossia in uno stesso stato ma in momenti diversi si possono avere in maniera casuale diverse osservazioni.

Sono necessarie anche le seguenti assunzioni di base:

- lo stato al tempo $t+1$ dipende direttamente solo dallo stato al tempo t per $t \geq 0$, come nelle catene di Markov;
- l'osservazione a tempo t dipende direttamente solo dallo stato al tempo t.

Un HMM *stazionario* è quindi definito attraverso le seguenti distribuzioni:

- $P(S_0)$ specifica delle condizioni iniziali;
- $P(S_{t+1} \mid S_t)$ specifica della dinamica;
- $P(O_t \mid S_t)$ specifica del modello del sensore.

Sono possibili diverse inferenze su tali modelli, come ad esempio ricostruire la sequenza di stati data una sequenza di osservazioni (ovvero la sua probabilità).

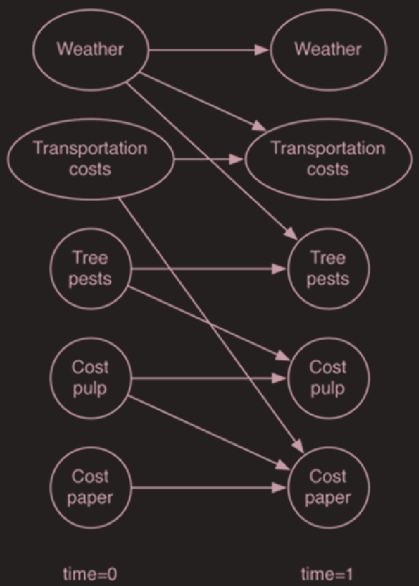
6.3 Belief Network Dinamica

Una **Belief Network Dinamica** (DBN) è una BN con una struttura regolare ripetuta nel tempo (discreto). Essa è simile a un HMM, ma stati/osservazioni sono rappresentati attraverso *feature* (anziché da una sola variabile): data una F feature con F_t si denoterà la variabile per il valore di F al tempo t.

Le assunzioni necessarie per definire una DBN sono:

- l'insieme di feature fissato nel tempo;
- per ogni $t > 0$, i genitori di F_t sono variabili relative a t o $t-1$ (grafo aciclico), quindi la struttura non dipende dal valore di t (tranne che per $t = 0$, caso speciale);
- per ogni $t > 0$ si ha la stessa distribuzione condizionata, ossia la dipendenza dai genitori, quindi il modello risulta *stazionario*.

Una DBN può essere rappresentata come una **BN a due passi** contenenti le variabili dei primi due istanti (0 e 1)



- per ogni F ci sono due variabili F_0 e F_1;

- $parents(F_0)$ può includere solo variabili per l'istante 0 e il grafo risultante sarà aciclico;
- le probabilità associate alla rete sono $P(F_0 \mid parents(F_0))$ e $P(F_1 \mid parents(F_1))$.

La BN a due passi può essere *dispiegata* (*unfolded*) in una BN replicando la struttura nei momenti successivi:

Dispiegamento della DBN precedente

In tale BN $P(F_i \mid parents(F_i))$, per $i > 1$, avrà la stessa struttura e le stesse probabilità condizionate di $P(F_1 \mid parents(F_1))$.

7 Simulazione Stocastica

Spesso i problemi di inferenza sono troppo complessi per un'efficiente inferenza esatta: ci sono molte variabili e molte dipendenze in gioco. In tali casi, conviene ricorrere all'*inferenza approssimata* con metodi basati sulla generazione di *campioni* casuali della distribuzione (a posteriori) specificata dalla rete.

Nella **simulazione stocastica** si costruisce un insieme di campioni mappati *su* e *da* probabilità. Per l'*inferenza* si va dalle probabilità ai campioni e viceversa. Ad esempio, la probabilità $P(a) = 0.25$ indica che su N campioni, per circa un quarto di essi la variabile booleana A risulterà vera.

I problemi tipicamente correlati con la simulazione stocastica sono i seguenti:

1. come *generare* i campioni;
2. come *inferire* probabilità dai campioni;
3. come *incorporare* nel meccanismo le osservazioni a disposizione.

I metodi **Monte Carlo** forniscono le basi per soluzioni che usano il campionamento per calcolare la distribuzione a posteriori di una variabile in una BN. Nel seguito si considereranno *forward sampling* e *Gibbs sampling*. Per altri metodi di sampling e approfondimenti si vedano [RN20, PM23].

7.1 Campionamento di una Variabile

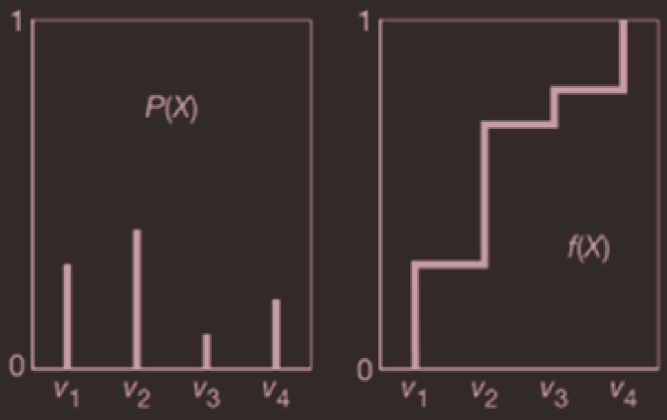
Per la *generazione di campioni* dalla distribuzione di una sola variabile X occorre ordinare i valori del dominio per definire la *distribuzione cumulativa*[4] f in funzione di x come segue:

$$f(x) = P(X \leq x)$$

Quindi, per ottenere un *campione casuale* v per X, si genera y da una distribuzione uniforme su $[0, 1]$ (numeri pseudo-casuali) e si ricava il valore $v \in dom(X)$ che abbia y come immagine nella cumulativa, ossia tale che $f(v) = y$, ovvero $v = f^{-1}(y)$.

Esempio — Si consideri la variabile X con dominio $\{v_1, v_2, v_3, v_4\}$ ordinato (i.e. $v_1 < v_2 < v_3 < v_4$) e distribuzione $P(X)$ definita come segue:

- $P(X = v_1) = 0.3$, $P(X = v_2) = 0.4$, $P(X = v_3) = 0.1$, $P(X = v_4) = 0.2$



Cumulativa da istogramma

Definita la cumulativa $f(v_i) = P(X \leq v_i)$:

- 30% del codominio di f ha v_1 come contro-immagine, ossia possibilità di essere campionato quando risulta $y \in [0, 0.3]$; analogamente, 40% per v_2, ecc. ...

Avendo a disposizione un insieme di campioni, per la *stima delle probabilità* da tale insieme basta ricorrere alla *media campionaria*: data una proposizione α sulle variabili, sia s la proporzione di campioni (assegnazioni) per i quali α è *vera* rispetto al totale n. Per la *legge dei grandi numeri*[5], s si avvicina asintoticamente alla probabilità esatta $p = P(\alpha)$ al crescere di n.

Per una *stima dell'errore* ϵ si può utilizzare la *disuguaglianza di Hoeffding*:

$$P(|p - s| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

da cui si può ricavare il numero di campioni n che garantisce una stima *probabilmente approssimativamente corretta* della probabilità.

7.2 Forward Sampling su BN

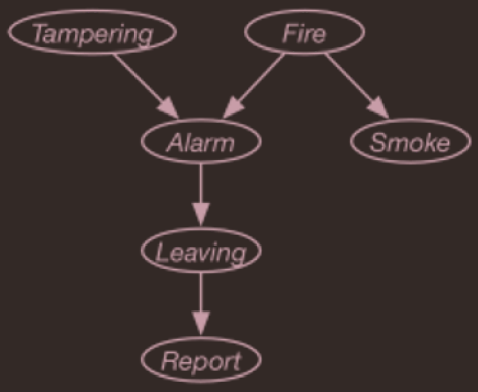
Campionando tutte le variabili di una BN ogni campione sarà generato in proporzione alla sua probabilità. Ciò consente di *stimare* le *probabilità* a priori di qualsiasi variabile.

Dato un ordinamento $X_1, \ldots, X_n$ delle variabili, tale che ognuna sia preceduta dai propri genitori, per estrarre un campione/tupla dalla distribuzione congiunta:

- si parte campionando X_1 usando la cumulativa;
- per ogni i da 2 a n:
  - si campiona un valore per X_i da $P(X_i|\text{parents}(X_i))$, distribuzione condizionata di X_i (una CPD della BN), dati i valori già campionati per i genitori.

La stima della distribuzione di una variabile di query si ottiene considerando la proporzione di campioni assegnati a ogni valore di tale variabile.

Esempio — Per la BN vista in precedenza, ordinamento come in tabella



Si deve ripetere la costruzione di un campione, con un valore per variabile, nell'ordine:

1. campionando $Tampering$ con la cumulativa si ottiene, ad esempio, $Tampering = false$;
2. campionando $Fire$ analogamente, si ha ad esempio $Fire = true$;
3. campionando $Alarm$ da $P(Alarm \mid Tampering = false, Fire = true)$ si potrebbe avere, ad esempio $Alarm = true$;
4. poi si campiona $Smoke$ usando $P(Smoke \mid Fire = true)$;
5. ecc...

Ciò va ripetuto fino a ottenere il numero desiderato di campioni/tuple $\{s_1, \ldots, s_n\}$.

campione	$Tampering$	$Fire$	$Alarm$	$Smoke$	$Leaving$	$Report$
s_1	false	true	true	true	false	false
s_2	false	false	false	false	false	false
s_3	false	true	true	true	true	true
s_4	false	false	false	false	false	true
s_5	false	false	false	false	false	false
s_6	false	false	false	false	false	false
s_7	true	false	false	true	true	true
s_8	true	false	false	false	false	true
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_{1000}	true	false	true	true	false	false

Ad esempio, per la stima di $P(Report = true)$ andrà considerata la proporzione dei campioni per i quali $Report$ è vera sui 1000 campioni.

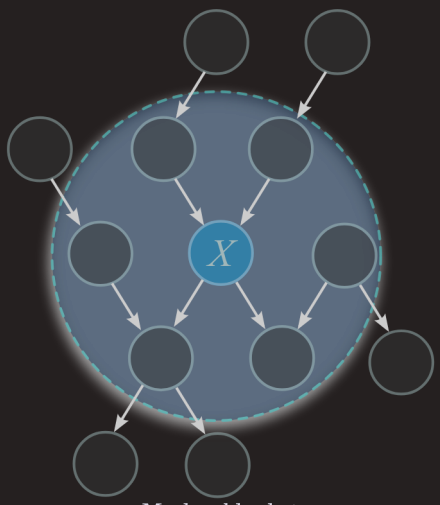
7.3 Markov Chain Monte Carlo

Si considera ora un metodo che prescinde dall'ordine di campionamento.

La distribuzione *stazionaria* di una catena di Markov è la distribuzione delle sue variabili che non cambia in base alla funzione di transizione della catena: se la catena mescola sufficientemente, c'è un'unica distribuzione stazionaria. Tale distribuzione può essere approssimata facendo *girare* il meccanismo di campionamento sufficientemente a lungo.

Nei metodi detti **Markov Chain Monte Carlo** (MCMC) per generare campioni da una distribuzione si costruisce una *catena di Markov* di campioni avente, a lungo andare, la distribuzione desiderata come sua (sola) distribuzione stazionaria, quindi si campiona da tale catena perché i suoi campioni saranno distribuiti secondo la distribuzione-obiettivo. Può occorrere una fase iniziale di *riscaldamento/burn-in*: andrebbero scartati i primi campioni, probabilmente lontani dalla distribuzione stazionaria, che sarà ottenuta più avanti.

In particolare nel **Gibbs sampling** si crea una catena di campioni da una BN, fissando i valori per le variabili osservate (condizionamento) e campionando le altre. Si genera un nuovo campione a partire dal precedente, modificando i valori per le variabili non fissate. Si può anche considerare una sola variabile alla volta [RN20]: la singola variabile sarà campionata in base alla distribuzione condizionata ai valori correnti delle altre variabili. Si noti che una variabile X in una BN dipende solo dai valori delle variabili nel suo **Markov blanket** $mb(X)$ contenente i suoi genitori, i suoi figli e gli altri genitori dei suoi figli, come si vede in figura:



Markov blanket

L'algoritmo può essere formalizzato come segue:

procedure Gibbs_sampling($B, e, Q, n, burn\_in$):
  Input
    B: belief network
    e: evidenza; assegnazione di valori ad alcune variabili
    Q: variabile di query
    n: numero di campioni da generare
    $burn\_in$: numero di campioni iniziali da scartare
  Output
    distribuzione a posteriori su Q
  Local
    $sample[]$, array in cui $sample[var] \in dom(var)$
    real $counts[k]$, array inizializzato a 0 per ogni $k \in dom(Q)$
  Inizializzare $sample[X] \leftarrow e[X]$ se X osservata,
    altrimenti con un valore causale da $dom(X)$
  repeat $burn\_in$ volte
    for each variabile X non osservata **do**
      $sample[X] \leftarrow$ campione casuale da $P(X \mid mb(X))$
  repeat n volte
    for each variabile X non osservata **do**
      $sample[X] \leftarrow$ campione da $P(X \mid mb(X))$
    $v \leftarrow sample[Q]$
    $counts[v] \leftarrow counts[v] + 1$
  return $counts / \sum_v counts[v]$ // normalizzazione dell'array

Esempio — Data una BN precedente: per campionare $P(Tampering|Smoke \wedge \neg report)$, si costruiscono i campioni:

campione	*Tampering*	*Fire*	*Alarm*	*Smoke*	*Leaving*	*Report*
s_1	**true**	true	false	true	false	false
s_2	**true**	true	false	true	false	false
s_3	**true**	true	false	true	false	false
s_4	**true**	true	true	true	false	false
s_5	**true**	true	true	true	true	false
s_6	**false**	true	true	true	true	false
…						
s_{1000}	**false**	true	true	true	true	false

- *Smoke* e *Report* sono fissate a *true* e false, rispettivamente;
- si genera s_1 casualmente e si seleziona *Tampering*;
- *Fire* e *Alarm* formano il Markov blanket per *Tampering* quindi si estrae un campione casuale per $P(Tampering \mid fire \wedge \neg alarm)$; supponendo sia *true* si ottiene così s_1;
- dato s_2, si estrae un valore casuale da $P(Fire \mid tampering \wedge \neg alarm \wedge smoke)$; si supponga sia *true* per ottenere quindi s_3; poi si campiona un valore da P(Alarm | tampering ∧ fire ∧ ¬leaving), ad es. *true*;
- alla fine la stima della probabilità di *tampering* sarà la proporzione dei casi *true* nei campioni, generati dopo il periodo di burn-in.

La distribuzione dei campioni si avvicina alla distribuzione congiunta reale se non ci sono probabilità *nulle*. La *velocità* dipende da quella del rimescolamento delle probabilità (ovvero da quanta parte dello spazio delle probabilità viene esplorato) che a sua volta dipende da quanto esse risultino *estreme*. In genere il metodo funziona bene per probabilità non estreme.

Esempio — Data una BN con la semplice struttura $A \to B \to C$ con A, B, C tutte variabili booleane e le seguenti CPT:

- $P(a) = 0.5$
- $P(b \mid a) = 0.99$
- $P(b \mid \neg a) = 0.01$ per cui $P(\neg b \mid \neg a) = 0.99$
- $P(c \mid b) = 0.99$
- $P(c \mid \neg b) = 0.01$ per cui $P(\neg c \mid \neg b) = 0.99$

Nessuna osservazione e variabile di query: $P(C)$

Per le due assegnazioni in cui tutte le variabili abbiano lo stesso valore (vero/falso), si troverebbe la stessa probabilità e risulterebbero più probabili di ogni altra assegnazione. Il Gibbs sampling porterebbe *rapidamente* verso una di tali assegnazioni, mentre richiederebbe molto tempo per arrivare alle altre, in quanto molto improbabili da selezionare. Con CPT con probabilità ancor più estreme (vicine a 1 e 0) la convergenza sarebbe ancor più lenta.

Riferimenti Bibliografici

[**PM23**] D. Poole, A. Mackworth: *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press. 3a ed. 2023

[**RN20**] S.J. Russell, P. Norvig: *Artificial Intelligence*. Pearson. 4th Ed. 2020

Link

[**Informazione**] Si veda sito oppure Wikipedia
[**BayesianNets**] Tutorial, BN Editor e altro materiale probabilistic.net
[**BayesServer**] Simulatore e API BayesServer
[**BayANet**] Simulatore @ Manchester University
[**StochasticSimulation**] su Wikipedia
[**CatenaMarkoviana**] en.Wikipedia
[**MCVE**] Visual Explanation e playground @ `setosa.io`
[**MonteCarlo**] Applet per l'integrazione approssimata
[**PP**] en.Wikipedia

Note

[1] Thomas Bayes su it.Wikipedia, en.Wikipedia.

[2] Esistono anche modelli grafici *non orientati* come i **Markov random field**. Per approfondimenti vedere lo specchietto in [PM23].

[3] *#NP* ("sharp-NP") classe di complessità del calcolo delle probabilità a priori o a posteriori di una variabile (richiede conteggi, non solo decisioni).

[4] Funzione di ripartizione o cumulativa: Wikipedia.

[5] Legge dei grandi numeri o di Bernoulli: $\bar{X}_n$ dei campioni converge asintoticamente a μ della distribuzione con cui sono generati i campioni X_i.