

MODELLI DI CONOSCENZA INCERTA

Nicola Fanizzi

Ingegneria della Conoscenza

CdL in Informatica • *Dipartimento di Informatica*

Università degli studi di Bari Aldo Moro

Probabilità

Semantica della Probabilità

Distribuzioni: Caso Discreto

Distribuzioni: Caso Continuo

Definizione Assiomatica della Probabilità

Proprietà delle Distribuzioni

Probabilità Condizionata

Chain Rule

Teorema di Bayes

Valori Attesi

Informazione / Entropia

Entropia e Information Gain

Indipendenza

Indipendenza Condizionata

Indipendenza Incondizionata

Belief Network

Osservazioni e Query

Costruire Belief Network

Inferenza Probabilistica

Inferenza Esatta

Inferenza Approssimata

Modelli Probabilistici Sequenziali

Catene di Markov

Modelli di Markov Nascosti

Belief Network Dinamica

Simulazione Stocastica

Campionare una Variabile

Forward Sampling su BN

Markov Chain Monte Carlo

Gibbs Sampling

PROBABILITÀ

Nel mondo reale: decisioni in base a informazioni *incomplete*

- difficile capire l'*esatto* stato del mondo
 - un medico non conosce le esatte condizioni interne del paziente
 - un docente non sa con precisione quanto sia stato compreso dai discenti
 - un robot non conosce quanto sia successo in una stanza lasciata da poco
- per prendere decisioni un KBS deve usare *tutta* la conoscenza che ha

Conoscenza limitata → **incertezza**

- Si richiameranno concetti legati a
 - *probabilità*
 - assunzioni di *indipendenza* sulla rappresentazione del mondo
 - modalità di *ragionamento* su tali rappresentazioni

Premessa 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 da:
 - *possibilità* di incidente
 - vantaggi / svantaggi:
 - *utilità* in caso d'incidente
 - *scomodità* dell'uso
 - importanza della *mobilità*

INCERTEZZA

Incerteza *epistemologica* concerne le credenze sullo stato del mondo





- ad es. “*persona molto alta*” offre solo una *vaga* conoscenza sull'esatto valore della sua altezza (imprecisione)
- l'incerteza *ontologica*, invece, riguarda il mondo in sé

Ragionare in presenza di *incerteza*

- problema studiato in *Teoria della probabilità* e in *Teoria delle decisioni*
 - calcolo dell'*azzardo*: incerteza sulle conseguenze di decisioni da prendere
 - non sempre prendere la migliore decisione possibile porta al successo

PROBABILITÀ

Misura del *credito* (belief) **bayesiana** o **soggettiva**
ossia “basata sulle conoscenze del soggetto” (e non arbitraria)

- prospettiva diversa da quella *oggettiva* / *frequentista*
 - ad es. A, B e C e gioco con un dado
 - A effettua un lancio, ottenendo 
ma dice a B solo che il risultato è *pari*,
mentre a C non viene detto nulla
 - diversa conoscenza → diverse probabilità soggettive per uno stesso evento:
 - per A: $P_A(\text{)} = 1$ certa/o del risultato osservato
 - per B: $P_B(\text{)} = \frac{1}{3}$ se disposto a credere ad A
 - per C: $P_C(\text{)} = \frac{1}{6}$ massima incertezza
 - credenze che riguardano un lancio specifico non uno generico

TEORIA DELLA PROBABILITÀ

Studio di come la conoscenza impatti sulle credenze:

- **credibilità** di una proposizione α misurata da un valore in $[0, 1]$ (*convenzionalmente*)
 - probabilità **nulla**: si crede che α sia del tutto falsa
 - nessuna nuova evidenza potrà cambiarla
 - probabilità pari a **1**: α assolutamente vera
 - probabilità in $]0, 1[$: incertezza sulla credibilità della sua verità
 - non significa che α sia solo parzialmente vera, bensì che la sua verità sia sconosciuta

Nozioni base:

- **variabile:** *aleatoria* / *casuale*
 - denotata con iniziale maiuscola e dotata di *dominio* di valori
 - v. *booleana* ha dominio $\{true, false\}$
 - per brevità $Happy = true$ denotato in minuscolo *happy*, analogamente *fire* per $Fire = true$
 - v. *discreta* dominio finito o almeno enumerabile
- **mondo:** funzione che associa a ogni variabile un valore
 - viceversa una variabile è una funzione dai mondi al suo dominio
 - ad es., 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 disequaglianza tra variabili e valori/variabili
 - es. $A = true, X < 7, Y > Z$
- **proposizione:** ottiene applicando a prop. primitive i connettivi logici

PROBABILITÀ DELLE PROPOSIZIONI

Dato l'insieme di *mondi possibili* Ω

- *finito* (assumendo un numero finito di variabili)

Misura di probabilità $P : \Omega \rightarrow \mathbb{R}_+$ tale che

$$\sum_{w \in \Omega} P(w) = 1$$

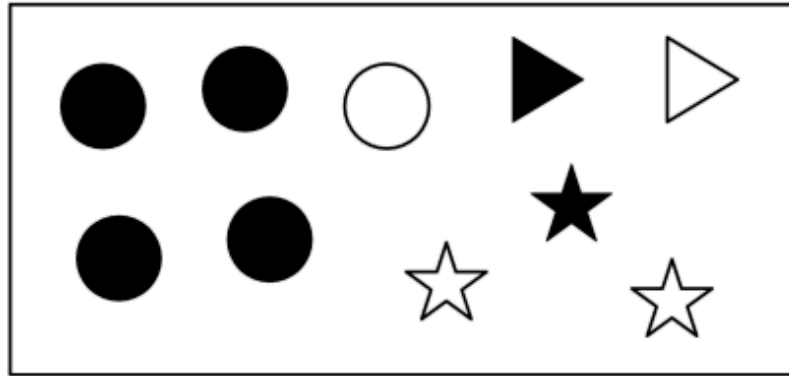
- *1* valore *convenzionale*

Da cui, data una proposizione α :

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

- misura coerente con la probabilità dei mondi

Esempio — Si considerino i 10 mondi in figura



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

DISTRIBUZIONI: CASO DISCRETO

Caso semplice: X variabile discreta

- **distribuzione di probabilità di X** funzione $P(X) : \text{dom}(X) \rightarrow \mathbb{R}$
 - $P(x)$ probabilità della proposizione $X = x$, con $x \in \text{dom}(X)$

Estensione al caso di un *insieme di variabili* $\{X_1, \dots, X_n\}$

- **distribuzione di probabilità congiunta $P(X_1, \dots, X_n)$** definita sui mondi di Ω
 - probabilità di w (assegnazione totale), ossia della proposizione che definisce w
 - ad es., $P(X, Y)$ distribuzione su X e Y :
dati $x \in \text{dom}(X)$ e $y \in \text{dom}(Y)$

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

- proposizione $X = x \wedge Y = y$

DISTRIBUZIONI: CASO CONTINUO

- dominio *infinito* di una variabile
- numero *infinito* di variabili

misura μ funzione definita sugli *insiemi di mondi*, a valori reali non-negativi che soddisfa le proprietà:

1. $\mu(S_1 \cup S_2) = \mu(S_1) + \mu(S_2)$ se $S_1 \cap S_2 = \emptyset$
2. $\mu(\Omega) = 1$ dove Ω è l'insieme di tutti i mondi
 - μ definita non necessariamente su tutti gli insiemi di mondi, ma solo su quelli descritti da formule logiche

probabilità della proposizione α :

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

- la probabilità di un insieme di mondi può non essere nulla anche se lo sono le misure di tutti i suoi mondi

DENSITÀ

Funzione di densità di probabilità $p : \mathbb{R} \rightarrow \mathbb{R}_+$ con integrale **unitario**

- probabilità che la variabile continua assuma un valore tra a e b :

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

- per qualsiasi formula su **intervalli** (anche con $<$)
 - possibile che, per qualsiasi a , $P(X = a) = P(a \leq X \leq a) = 0$

Distribuzioni: densità / probabilità

- **parametrica**: formula con un numero finito di parametri prefissati
- **non parametrica**: non fissa il numero dei parametri, anche tantissimi

Altro metodo: discretizzazione in un numero finito di mondi

- ad es., altezze limitate approssimate al centimetro

Misura $P : \text{Proposizioni} \rightarrow \mathbb{R}$ che soddisfa gli *assiomi della probabilità*:

1. $0 \leq P(\alpha)$ per ogni proposizione α
 - credito mai negativo
2. $P(\tau) = 1$ se τ è una tautologia
 - vera in tutti i mondi possibili
3. $P(\alpha \vee \beta) = P(\alpha) + P(\beta)$ se α e β si contraddicono reciprocamente
 - mutuamente esclusive, i.e. $\neg(\alpha \wedge \beta)$ tautologia: mai entrambe vere

Una **misura** che rispetti tali proprietà rientra nella **Teoria della Probabilità**

- **frequenze empiriche**: proposizioni su **proporzioni** di esempi d'un dataset
 - obbediscono agli assiomi → seguono le regole della probabilità
 - ma esistono misure non riconducibili a frequenze empiriche

Assiomatizzazione del significato della probabilità

- **coerente**: la probabilità definita in termini dei mondi possibili segue gli assiomi
- **completa**: un sistema che obbedisce agli assiomi ha una semantica probabilistica

.....

Proposizione — Se il numero di variabili discrete è finito,
gli assiomi sono completi e coerenti rispetto alla semantica definita

dim. mondi mutuamente esclusivi → dalle loro probabilità, usando gli assiomi,
si può ricavare quella di qualunque proposizione [1]

PROPRIETÀ DELLE DISTRIBUZIONI

Proposizione — Date le proposizioni α e β e la var. V :

1. $P(\neg\alpha) = 1 - P(\alpha)$
2. Se $\alpha \leftrightarrow \beta$ allora $P(\alpha) = P(\beta)$
3. $P(\alpha) = P(\alpha \wedge \beta) + P(\alpha \wedge \neg\beta)$
4. $P(\alpha) = \sum_{d \in \text{dom}(V)} P(\alpha \wedge V = d)$
5. $P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$

dim. si veda [1]

negazione
equivalenza, equiprobabilità
ragionamento per casi
marginalizzazione
disgiunzione

Disponibilità di fatti nuovi → credibilità da *aggiornare*

Probabilità condizionata $P(h \mid e)$ di h *data* e :
misura del credito di h (*ipotesi*), assumendo la verità di e (*evidenza*)

- $P(h \mid e)$ 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
- $P(h)$ probabilità **a priori** di h
 - corrisponde a $P(h \mid \text{true})$, 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 consegna

- evidenza raccolta via via dai sensori
- se sono rumorosi, il robot può sbagliarsi sull'idea del mondo
 - pur essendo consapevole dell'informazione ricevuta

SEMANTICA DELLA PROBABILITÀ CONDIZIONATA

Si definisce prima sui mondi e poi sulle proposizioni

- selezionando i mondi possibili in base all'evidenza (come per la conseguenza logica)

probabilità condizionata $P(w \mid e)$ del *mondo* w data l'evidenza e :

$$P(w \mid e) = \begin{cases} 0 & e \text{ falsa in } w \\ c \cdot P(w) & e \text{ vera in } w \end{cases}$$

- ogni mondo in cui e sia falsa ha probabilità condizionata nulla
 - e fa *scartare* mondi incompatibili
- per gli altri mondi, probabilità normalizzate
 - c *costante di normalizzazione* (dipende da e)

Perché $P(w \mid e)$ sia una misura di probabilità, dato e :

$$\begin{aligned} 1 &= \sum_{w \in \Omega} P(w \mid e) \\ &\stackrel{\text{per casi}}{=} \sum_{w: e \text{ vera in } w} P(w \mid e) + \sum_{w: e \text{ falsa in } w} P(w \mid e) \\ &= \sum_{w: e \text{ vera in } w} c \cdot P(w) + 0 \\ &\stackrel{\text{def. } P(e)}{=} c \cdot P(e) \end{aligned}$$

$$\rightarrow c = 1/P(e)$$

- pertanto $P(w \mid e)$ definibile solo se $P(e) > 0$
 - difatti $P(e) = 0 \rightarrow e$ impossibile

Probabilità condizionata della *proposizione* h data e :

- sommando le probabilità condizionate dei mondi in cui h è vera

$$\begin{aligned} P(h \mid e) &= \sum_{w: h \text{ vera in } w} P(w \mid e) \\ &\stackrel{\text{per casi}}{=} \sum_{w: h \wedge e \text{ vera in } w} P(w \mid e) + \sum_{w: h \wedge \neg e \text{ vera in } w} P(w \mid e) \\ &= \sum_{w: h \wedge e \text{ vera in } w} \frac{1}{P(e)} \cdot P(w) + 0 \\ &= \frac{P(h \wedge e)}{P(e)} \end{aligned}$$

Altrove questa viene data come definizione di $P(h \mid e)$

- qui ricavata da una definizione più semplice

Esempio — Si considerino i mondi della figura precedente, assumendo che le loro probabilità siano pari a 0.1

- data l'evidenza $Filled = false$,
solo 4 mondi avranno una probabilità a posteriori non nulla
 - $P(Shape = circle \mid Filled = false) = 0.25$
 - $P(Shape = star \mid Filled = false) = 0.5$

Distribuzione di probabilità condizionata in funzione delle variabili, essendo X e Y (insiemi di) variabili:

- dati $x \in \text{dom}(X)$ e $y \in \text{dom}(Y)$

$$P(X = x \mid Y = y)$$

probabilità condizionata delle proposizioni corrispondenti

Cfr. nel testo *Background Knowledge and Observation*

- modello di BK in termini di modello probabilistico
- le osservazioni formano l'evidenza sulla quale operare il condizionamento

CHAIN RULE

Decomposizione di una congiunta

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

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

- la parte destra si annulla se è nullo uno dei prodotti
 - anche in caso alcuni fossero indefiniti

Osservazione

Le stesse probabilità condizionate sono misure di probabilità

→ ogni risultato valido per le distribuzioni non condizionate vale anche per quelle condizionate:

- mantenendo fissa l'evidenza
 - ad es., dalle proprietà viste in precedenza (caso 5.):

$$P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$$

quindi anche

$$P(\alpha \vee \beta \mid k) = P(\alpha \mid k) + P(\beta \mid k) - P(\alpha \wedge \beta \mid k)$$

TEOREMA DI BAYES

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

Proposizione (*Regola di Bayes*¹)

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

essendo $P(e) \neq 0$

- $P(h \mid e)$ probabilità **a posteriori**
- $P(e \mid h)$ **verosimiglianza** o *likelihood*
- $P(h)$ probabilità **a priori** di h

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

essendo $P(e \mid k) \neq 0$

- partendo da $P(h \mid k)$, con k già osservata o conoscenza di fondo *implicita*
- nuova osservazione $e \rightarrow$ nuova misura $P(h \mid e \wedge k)$

USO DELLA REGOLA DI BAYES

- Si noti che: $P(h \mid e) \propto P(e \mid h) \cdot P(h)$
 - usata per confrontare diverse h_i su gli stessi dati e
- denominatore costante (non dipende da h) trascurabile nei confronti ma determinabile ragionando *per casi*:
 - dato \mathcal{H} insieme di tutte possibili ipotesi (mutuamente esclusive) per le proprietà delle distribuzioni:

$$P(e) = \sum_{h \in \mathcal{H}} P(e \wedge h) = \sum_{h \in \mathcal{H}} P(e \mid h) \cdot P(h)$$

- i.e. somma dei numeratori per le diverse ipotesi
 - difficile da calcolare se \mathcal{H} grande
- 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:

si osservano sintomi (S) e si vorrebbero determinare potenziali malattie (M)

- servirebbe calcolare $P(M | S)$
 - difficile perché dipende dal *contesto*
 - es. alcune malattie sono più frequenti negli ospedali
- tipicamente più facile determinare $P(S | M)$
 - relazione che lega i sintomi alle malattie, meno dipendente dal contesto
- probabilità correlate attraverso Bayes
 - dove $P(M)$ rappresenta l'influenza del contesto

Esempio ⚡ — Diagnostica dell'impianto luci:

- si vuole sapere se l'**interruttore** s_1 funzioni
- non può saperlo l'elettricista che ha realizzato l'impianto, ma potrebbe essere 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 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):

$$\begin{aligned} P(\textit{fire} \mid \textit{alarm}) &= \frac{P(\textit{alarm} \mid \textit{fire}) \cdot P(\textit{fire})}{P(\textit{alarm})} = \\ &= \frac{P(\textit{alarm} \mid \textit{fire}) \cdot P(\textit{fire})}{P(\textit{alarm} \mid \textit{fire}) \cdot P(\textit{fire}) + P(\textit{alarm} \mid \neg \textit{fire}) \cdot P(\neg \textit{fire})} \end{aligned}$$

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

PROBABILITÀ CONDIZIONATA E IMPLICAZIONE

$$P(f \leftarrow e) \stackrel{?}{=} P(f | e)$$

Non necessariamente:

- $e \rightarrow f \equiv \neg e \vee f$ quindi $P(e \rightarrow f) = P(\neg e \vee f)$
misura delle interpretazioni (mondi) per cui e falsa o f vera
 - e.g., insieme degli **uccelli** piccolo rispetto all'insieme degli **animali**
 - $P(\neg flies | bird)$: **bassa**
 - gli uccelli che non volano sono rari
 - $P(\neg flies \leftarrow bird) = P(\neg flies \vee \neg bird)$: **alta**
 - proporzione dominata da animali che non sono uccelli
 - per lo stesso motivo anche $P(bird \rightarrow flies)$ **alta**

Medie ponderate su tutti i mondi possibili dei valori di funzioni numeriche

Data f , funzione sui mondi, **valore atteso** di f rispetto a P :

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

- caso speciale: α proposizione e $f(\alpha) = \begin{cases} 1 & \alpha \text{ vera} \\ 0 & \alpha \text{ falsa} \end{cases}$ (cfr. interpretazione π)

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

- esempi: f funzione che restituisca
 - il valore di una variabile
 - il numero di bit usati per descrivere un mondo
 - una misura di gradimento

Esempio — Domotica: $\mathcal{E}_P(\textit{number_of_broken_switches})$
numero atteso di deviatori rotti secondo la distribuzione P

- numero **medio**, nel lungo periodo, di deviatori non funzionanti secondo P
- con 3 deviatori, ognuno con probabilità **0.7** di essere rotto, 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

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:

$$\mathcal{E}(\textit{number_of_broken_switches} \mid \neg \textit{lit}(l_1))$$

- mediando i numeri di interruttori rotti su tutti i mondi in cui l_1 è spenta

.....

Variabile booleana su $\{0, 1\}$: **valore atteso** = 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

L'*informazione* si misura in *bit*

X variabile aleatoria:

- $x \in \text{dom}(X)$ identificabile con un codice di $\lceil -\log_2 P(x) \rceil$ bit
- numero atteso di bit per trasmettere un x :

$$H(X) = \sum_{x \in \text{dom}(X)} -P(X = x) \cdot \log_2 P(X = x)$$

contenuto informativo o entropia di X

- **NB** H funzione della variabile X (non dei suoi valori) $\rightarrow H(X)$ numero
 - come $P(X)$ che restituisce un numero per ogni valore di X

ENTROPIA E INFORMATION GAIN

Entropia di X data l'osservazione $Y = y$:

$$H(X \mid Y = y) = \sum_x -P(X = x \mid Y = y) \cdot \log_2 P(X = x \mid Y = y)$$

Entropia condizionata di X data Y ,
valore atteso mediando sul valore di Y osservato:

$$H(X \mid Y) = \sum_y P(Y = y) \cdot H(X \mid Y = y) = \sum_y P(Y = y) \cdot \sum_x -P(X = x \mid Y = y) \cdot \log_2 P(X = x \mid Y = y)$$

Information Gain (IG) per un *test* che determina il valore di Y :

$$H(X) - H(X \mid Y)$$

numero di bit (≥ 0) risparmiati nel descrivere X , noto il valore di Y

Esempio — Ruota della fortuna con numeri 1,2,...,8, equiprobabili

- S risultato di un giro: $H(S) = -\sum_{i=1}^8 \frac{1}{8} \cdot \log_2 \frac{1}{8} = 3$ bit
- Dato un sensore G che indica se il risultato superi 6: $G = true$ sse $S > 6$
 - $H(S \mid G) = -\left(0.25 \log_2 \frac{1}{2} + 0.75 \log_2 \frac{1}{6}\right) = 2.19$
 - si può costruire un codice che usa 219 bit per predire 100 risultati
 - IG di G : $3 - 2.19 = 0.81$ bit
- Sensore di parità E (even): $E = true$ sse S è pari
 - $H(S \mid E) = -\left(0.5 \log_2 \frac{1}{4} + 0.5 \log_2 \frac{1}{4}\right) = 2$
 - IG di E : 1 bit

Utilità:

- **diagnosi**: scelta del test che massimizzi l'informazione ricavabile
- apprendimento di **alberi di decisione**:
scelta del partizionamento degli esempi che porta al maggior IG
 - elementi da distinguere: diversi valori nel concetto obiettivo
 - probabilità ottenute dalle proporzioni di ogni valore nella parte di training set nel nodo
- apprendimento **Bayesiano**:
 - decisione sul miglior modello rispetto ai dati osservati

INDIPENDENZA

Per definire distribuzioni su molte variabili serve molta conoscenza

- assiomi della probabilità, deboli:
pochi vincoli sulle probabilità condizionate da sfruttare
 - ad es., con n variabili binarie, $2^n - 1$ probabilità da conoscere per una completa distribuzione dalla quale derivare quelle condizionate
 - serve un DB enorme

Per limitare il quantitativo d'informazione richiesta,
spesso si assume che una variabile dipenda direttamente *solo* da alcune altre

INDIPENDENZA CONDIZIONATA

Assunzioni tali che si richiedano:

- meno dati per specificare un modello
- strutture più semplici → ragionamento 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:
 - $P(h \mid f \wedge e) = 1$ se f implica h
 - $P(h \mid f \wedge e) = 0$ se f implica $\neg h$

Tipo comune di *conoscenza qualitativa* spesso disponibile

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

- cioè f irrilevante per la probabilità di h data e

Estendendo l'idea a un dato insieme di variabili Z_s :

- X *condizionatamente* indipendente da Y dato Z_s sse

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

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

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

- dato un valore per ogni variabile in Z_s ,
conoscere il valore di Y non influenza la credibilità di un dato valore per X
- indipendenza condizionata
 - *facile* da accertare (spesso) e utile nell'*inferenza*
 - *raro* invece disporre di una tabella delle probabilità dei mondi
e determinare 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 da studio e intelligenza
 - quindi, nota *Answers*, *Intelligent* dipenderebbe da *Works_hard*
 - casi di risposte approfondite fornite senza aver studiato molto fanno aumentare il credito nell'intelligenza
- il voto (*Grade*) dovrebbe dipendere dalle risposte, non da intelligenza o studio profuso
 - → *Grade* indipendente da *Intelligent* data *Answers*
- in mancanza delle risposte, *Intelligent* influenzerebbe *Grade*
 - in generale: studenti più intelligenti forniscono risposte migliori
 - → *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. $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

INDIPENDENZA INCONDIZIONATA

X e Y *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
- proprietà raramente determinabile per via numerica da dati disponibili

BELIEF NETWORK

Indipendenza condizionata → rappresentazione concisa dei domini:

- data X , solo alcune variabili influenzano *direttamente* il suo valore
 - insieme di variabili che la influenzano localmente: *Markov blanket*
 - località che va sfruttata
 - date queste variabili, X condizionatamente indipendente dalle altre
- *indipendenza* condizionata
 - *ordinamento* delle variabili
 - *grafo orientato*

Rete Bayesiana: modello della (in)dipendenza condizionata fra variabili

- definito da un ***ordinamento totale*** sull'insieme delle sue variabili
 - ad es. X_1, \dots, X_n :
- ***distribuzione*** di probabilità ***congiunta*** decomposta in termini di probabilità condizionate, tramite ***chain rule***:

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

i.e., in termini di variabili e distribuzioni:

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

Genitori di una variabile X_i :

insieme minimale di **predecessori** di X_i nell'ordinamento totale,
 $parents(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$, tale che gli altri predecessori di X_i siano
condizionatamente indipendenti da X_i dato $parents(X_i)$

- X_i dipende dai genitori, ma è indipendente dagli altri predecessori:

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

- più insiemi minimali soddisfano la condizione → scelta casuale
 - alcuni dei predecessori dipendono deterministicamente dagli altri
- con la **chain rule**, **fattorizzazione** della **congiunta**:

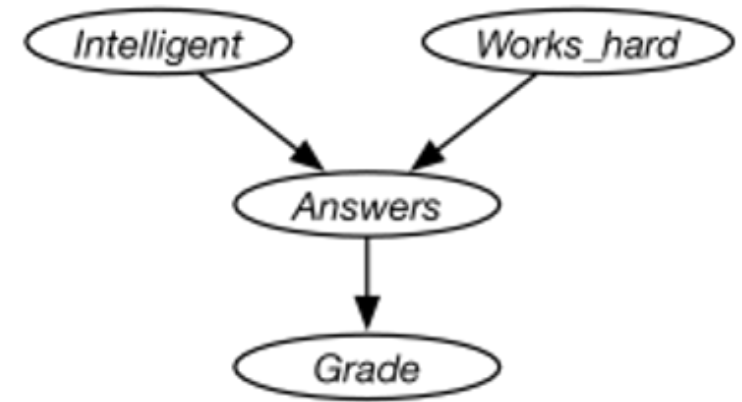
$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid parents(X_i))$$

Belief Network [BN] o rete bayesiana: grafo aciclico orientato (DAG) con:

- una variabile X per nodo (con un suo dominio) con
 - un arco verso X dai nodi relativi a ognuno dei genitori $parents(X)$
 - una distribuzione condizionata $P(X \mid parents(X))$
 - probabilità a priori, per quelle senza genitori
- **aciclica** per costruzione:
 - la fattorizzazione dipende dall'ordinamento
 - genitori scelti tra i predecessori
 - diversi ordinamenti \rightarrow diverse BN
 - alcuni ordinamenti portano a reti con meno archi
- **indipendenza** condizionata: 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* dell'esempio precedente

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



- **fattorizzazione della distribuzione congiunta:**

$$P(\textit{Intelligent}, \textit{Works_hard}, \textit{Answers}, \textit{Grade}) =$$

$$P(\textit{Intelligent}) \cdot P(\textit{Works_hard})$$

$$\cdot P(\textit{Answers} \mid \textit{Intelligent}, \textit{Works_hard})$$

$$\cdot P(\textit{Grade} \mid \textit{Answers})$$

- **si considerano domini semplici**

- **ad es.**, $\textit{Answers} = \{\textit{insightful}, \textit{clear}, \textit{superficial}, \textit{vacuous}\}$

- o anche il testo stesso delle risposte

Data una BN che specifica una distribuzione congiunta, problema di *inferenza probabilistica* più comune:

- calcolo della *distribuzione a posteriori* di una o più **variabili di query** data un'evidenza
 - congiunzione di assegnazioni di valori a variabili

.....

Esempio — Prima delle osservazioni, $P(\textit{Intelligent})$ fornita dalla BN; dopo *inferenza* per determinare $P(\textit{Grade})$

- query: osservato il voto **A** (un valore di *Grade*), distribuzione a posteriori $P(\textit{Intelligent} \mid \textit{Grade} = \textit{A})$
- query: osservato pure $\textit{Works_hard} = \textit{false}$, distribuzione a posteriori $P(\textit{Intelligent} \mid \textit{Grade} = \textit{A} \wedge \textit{Works_hard} = \textit{false})$
- *Intelligent* e *Works_hard* indipendenti quando non siano date osservazioni ma, se si conosce il voto, diventano *dipendenti*
 - ciò spiega perché ci si vanta se si hanno buoni voti senza studiare molto: accresce la probabilità di risultare intelligenti

Per modellare un dominio attraverso una BN:

- ***Variabili rilevanti***
 - cosa può essere osservato del dominio
 - feature osservata → variabile
 - per potere poi condizionare sulle osservazioni
 - informazione di interesse data la probabilità a posteriori
 - feature → variabile per possibili query
 - variabili ***nascoste*** o **latenti**: né osservate né di query
 - per tener conto di dipendenze
 - utili semplificare il modello:
meno probabilità condizionate da specificare

- **Valori del dominio**: decidere il livello di dettaglio del ragionamento utile a rispondere alle query
 - per ogni variabile, specifica del significato di ogni valore
 - in ottemperanza del **principio di chiarezza**:
onniscenza del sistema riguardo il valore di ogni variabile
 - cosa deve accadere perché una variabile (non latente) assuma un dato valore
 - va documentato il significato delle variabili e dei loro valori
 - tranne per quelle latenti che vanno **apprese** dai dati
- **Relazioni** tra le variabili
 - archi da considerare per definire *parent*
- **Dipendenza** della distribuzione di una variabile **dai genitori**
 - tabelle delle distribuzioni condizionate → **CPT**

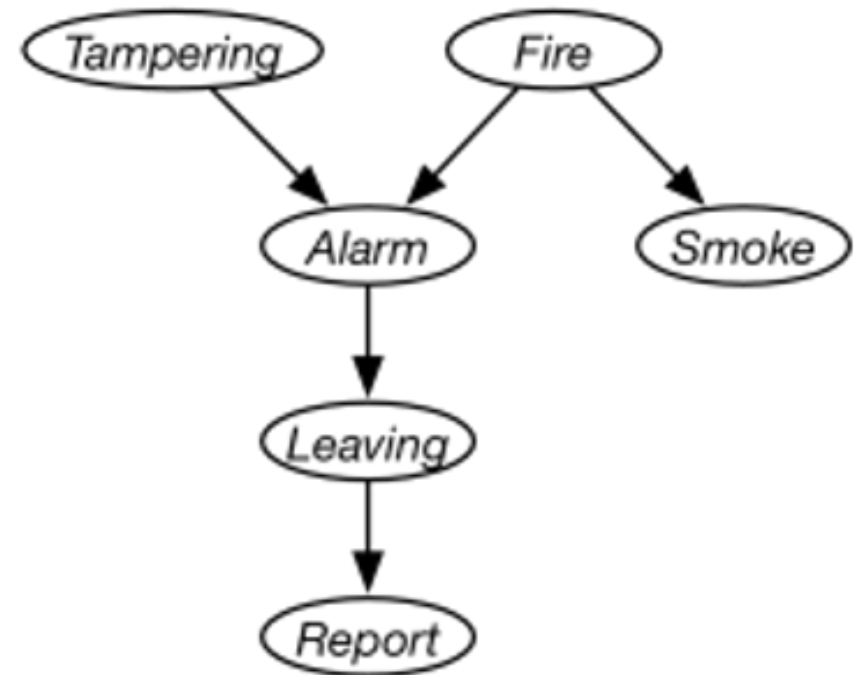
Esempio — Diagnostica di allarme anti-incendio che tiene anche conto di: manomissioni, rumore dei sensori e info contraddittorie sulla situazione

- *Report* segnala eventuali evacuazioni in corso, variabile *rumorosa*:
 - a volte sbaglia perché non è in corso alcuna evacuazione (*falso positivo*)
 - altre volte non segnala che è in corso un'evacuazione (*falso negativo*)
 - l'evacuazione dipende dall'*allarme* scattato
 - *manomissione* o *incendio* possono influenzare l'allarme
 - la segnalazione di *fumo* dipende solo dalla presenza dell'incendio
- *Variabili* booleane (elenco ordinato):
 - *Tampering* vera sse c'è stata la manomissione dell'allarme
 - *Fire* vera sse c'è un incendio
 - *Alarm* vera sse si sente l'allarme
 - *Smoke* vera sse c'è fumo
 - *Leaving* vera sse molta gente sta lasciando in massa l'edificio
 - *Report* vera sse si riferisce di gente che abbandona l'edificio

(..cont.)

- Relazioni di *indipendenza condizionata*:

- *Fire* indipendente da *Tampering*
 - (senza condizioni)
- *Alarm* dipende da *Fire* e *Tampering*
 - nessuna assunzione di indipendenza per *Alarm* rispetto ai predecessori
- *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*
 - data *Alarm*, è indipendente dalle altre variabili
- *Report* dipende direttamente solo da *Leaving*



(..cont.)

- **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:

- indipendenza di *Alarm* da *Smoke* dato *Fire*
- dipendenza di *Smoke* da *Fire*

- **Domini**

- variabili booleane: variante in minuscolo per denotare l'assegnazione di *true*, altrimenti si fa precedere l'op. di negazione \neg

- ad esempio:

Tampering = *true* si abbrevia con *tampering*

Tampering = *false* abbreviato con \neg *tampering*

(..cont.)

- **Probabilità condizionate (CPT)** usate negli esempi che seguiranno:

- $P(\text{tampering}) = 0.02$
- $P(\text{fire}) = 0.01$
- $P(\text{alarm} \mid \text{fire} \wedge \text{tampering}) = 0.5$
- $P(\text{alarm} \mid \text{fire} \wedge \neg \text{tampering}) = 0.99$
- $P(\text{alarm} \mid \neg \text{fire} \wedge \text{tampering}) = 0.85$
- $P(\text{alarm} \mid \neg \text{fire} \wedge \neg \text{tampering}) = 0.0001$
- $P(\text{smoke} \mid \text{fire}) = 0.9$
- $P(\text{smoke} \mid \neg \text{fire}) = 0.01$
- $P(\text{leaving} \mid \text{alarm}) = 0.88$
- $P(\text{leaving} \mid \neg \text{alarm}) = 0.001$
- $P(\text{report} \mid \text{leaving}) = 0.75$
- $P(\text{report} \mid \neg \text{leaving}) = 0.01$

- probabilità delle assegnazioni negative, complementari:

es. $P(\neg \text{fire}) = 0.99$

(..cont.)

- Probabilità a priori prima di ogni osservazione (evidenza):

- $P(\textit{tampering}) = 0.02$

- $P(\textit{report}) = 0.028$

- $P(\textit{leaving}) = ?$

$$P(\textit{fire}) = 0.01$$

$$P(\textit{smoke}) = 0.0189$$

- Avendo osservato l'arrivo di un **rapporto** (i.e. data *report*):

- $P(\textit{tampering} \mid \textit{report}) = 0.399$

- $P(\textit{fire} \mid \textit{report}) = 0.2305$

- $P(\textit{smoke} \mid \textit{report}) = 0.215$

- probabilità di *tampering* e *fire* aumentate a causa di *report*
quindi anche quella di *smoke* che dipende da *fire*

- Avendo osservato (solo) del **fumo** (i.e. data *smoke*):

- $P(\textit{tampering} \mid \textit{smoke}) = 0.02$

- probabilità di manomissione non influenzata dall'osservazione di fumo

- $P(\textit{fire} \mid \textit{smoke}) = 0.476$

- $P(\textit{report} \mid \textit{smoke}) = 0.320$

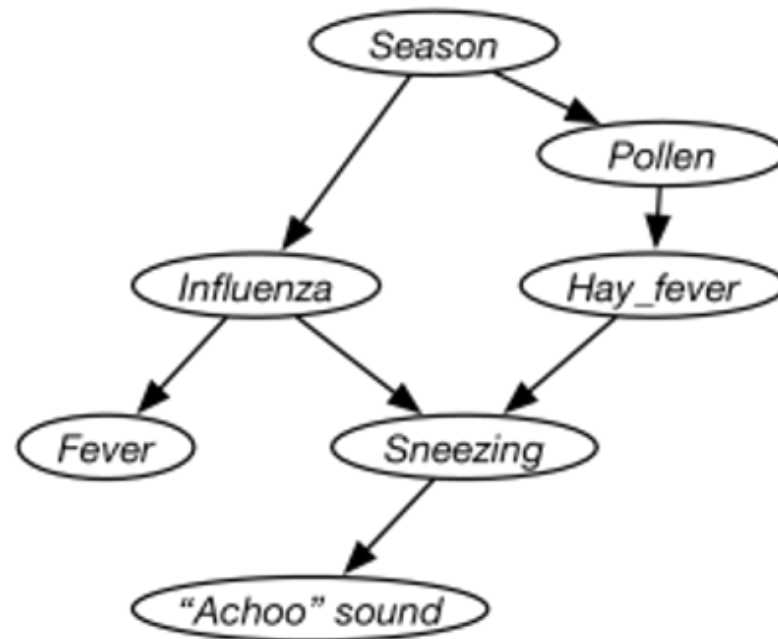
- si noti come entrambe aumentano

(..cont.)

- Avendo osservato *report* e *smoke*:
 - $P(\text{tampering} \mid \text{report} \wedge \text{smoke}) = 0.0284$
 - incendio ancora più probabile
 - $P(\text{fire} \mid \text{report} \wedge \text{smoke}) = 0.964$
 - invece, 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
- Invece, avendo osservato *report* ma non *smoke*:
 - $P(\text{tampering} \mid \text{report} \wedge \neg \text{smoke}) = 0.501$
 - $P(\text{fire} \mid \text{report} \wedge \neg \text{smoke}) = 0.0294$
 - osservato *report*, *fire* diventa molto meno verosimile quindi la probabilità di *tampering* aumenta per poter spiegare quanto osservato

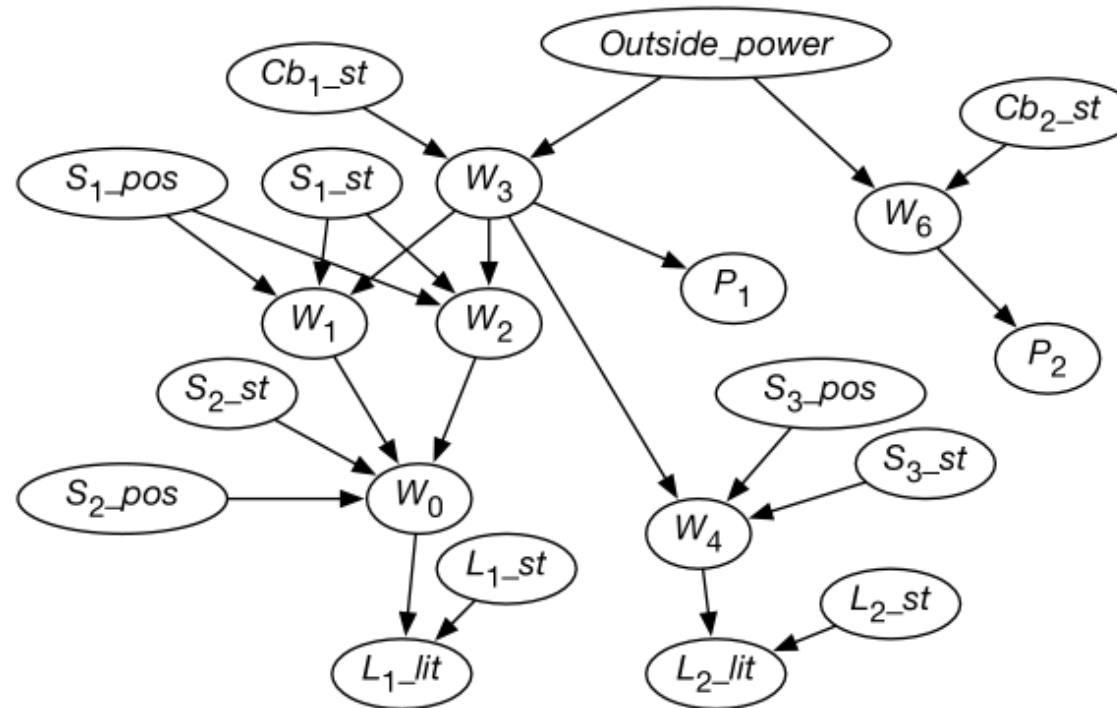
Esempio \sphericalangle – Diagnosi dello *starnuto*

- potrebbe dipendere dall'influenza o da febbre da fieno
 - 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



(..cont.)

- Per ogni cavo (wire) w_i , c'è una variabile W_i con dominio $\{live, dead\}$ per indicare se vi passi corrente (*live*) o meno (*dead*)
- *Outside_power* (stesso dominio) indica la disponibilità di energia dall'esterno
- Per ogni deviatore s_i , S_i_pos indica la posizione, dominio $\{up, down\}$
- Per ogni deviatore s_i , S_i_st indica lo stato, dominio $\{ok, upside_down, short, intermittent, broken\}$
- Per ogni salvavita cb_i , Cb_i_st ha il dominio $\{on, off\}$, dove *on* indica che l'energia elettrica può fluire attraverso cb_i e *off* che non può farlo
- Per ogni luce l_i , L_i_st con dominio $\{ok, intermittent, broken\}$ denota uno dei suoi tre stati

(..cont.)

Ordinamento causale naturale

- L_1_lit : l'accensione di l_1 dipende solo dalla presenza di energia nel cavo w_0 e dal buon funzionamento di l_1
 - altre variabili irrilevanti
 - posizione interruttore s_1
 - accensione di l_2
 - ...
 - genitori di L_1_lit : W_0 e L_1_st
- W_0 : sapendo che sono alimentati (*live*) i cavi w_1 e w_2 , la posizione di s_2 e le condizioni del deviatore, i valori delle altre variabili (diverse da L_1_lit) non influenzano W_0
 - genitori di W_0 : S_2_pos , S_2_st , W_1 e W_2

(..cont.)

- W_1 : probabilità da specificare

- $P(W_1 = live \mid S_{1_pos} = up \wedge S_{1_st} = ok \wedge W_3 = live)$
- $P(W_1 = live \mid S_{1_pos} = up \wedge S_{1_st} = ok \wedge W_3 = dead)$
- $P(W_1 = live \mid S_{1_pos} = up \wedge S_{1_st} = upside_down \wedge W_3 = live)$
- \vdots
- $P(W_1 = live \mid S_{1_pos} = down \wedge S_{1_st} = broken \wedge W_3 = dead)$
 - 2 valori per S_{1_pos} , 5 per S_{1_ok} e 2 per W_3 , quindi $2 \times 5 \times 2 = 20$ casi per la specifica della probabilità condizionata quando $W_1 = live$
 - può essere assegnata arbitrariamente
 - ha senso sfruttare conoscenza che vincoli i valori
 - valori per $W_1 = dead$ calcolati in base ai valori per $W_1 = live$

- S_{1_st} non ha genitori

- basta una distribuzione a priori che specifichi (almeno 4 del)le seguenti probabilità: $P(S_{1_st} = ok)$, $P(S_{1_st} = short)$, $P(S_{1_st} = upside_down)$, $P(S_{1_st} = intermittent)$ e $P(S_{1_st} = broken)$

- Analogamente per le altre variabili

(..cont.)

Uso della BN:

- Condizionando sul sapere che interruttori e salvavita funzionino e sui valori dell'energia in ingresso e le posizioni degli interruttori, si simula come dovrebbe funzionare l'illuminazione
- Dati i valori dell'energia dall'esterno e della posizione degli interruttori, si sa inferire la probabilità di ogni risultato
 - ad es. dell'accensione di l_1
- Dati i valori degli interruttori e sull'accensione delle luci, si può inferire la probabilità a posteriori che ogni interruttore o salvavita sia un particolare stato
- Date alcune osservazioni, si può determinare la posizione più probabile degli interruttori
- Date alcune posizioni degli interruttori, alcuni output e valori intermedi si può determinare la probabilità di ogni altra variabile nella rete

(..cont.)

Assunzioni di indipendenza insite nel modello:

- Il DAG specifica che i vari componenti si rompano in maniera indipendente
 - per modellare ***dipendenze*** sulla rottura si possono ***aggiungere*** archi (o variabili)
 - ad es., se alcune luci non si rompono in modo indipendente forse vengono dallo stesso stock: si può aggiungere una variabile per modellare lo stock e indicare se fosse buono o cattivo rendendola genitore di L_i_st per ogni L_i dello stock
 - le luci si romperebbero ora in maniera dipendente
 - osservata una luce rotta, la probabilità che lo stock sia difettoso potrebbe aumentare, rendendo più verosimile che altre luci dello stock siano rotte
 - se non si è sicuri sulla provenienza delle luci da uno stesso stock si possono aggiungere variabili per tener conto anche di tale situazione
- La BN fornisce una specifica dell'indipendenza che permette di modellare in modo ***naturale*** le dipendenze

(..cont.)

Il modello implica che non vi sia alcuna possibilità di corti nei cavi o che l'edificio sia cablato diversamente:

- ad es., implica che w_0 non può essere mandato in corto su w_4 in quanto w_0 riceve energia da w_4
 - si possono aggiungere altre dipendenze per modellare ogni altro tipo di corto
- in alternativa si aggiunge un nodo che indichi che il *modello* è *appropriato*
 - e archi da tale nodo verso ogni variabile che rappresenti la presenza di energia in un cavo o luce
 - quando il modello è *appropriato*, si possono usare le probabilità viste in un esempio precedente
 - quando è *inappropriato*, si può, ad es., specificare che ogni cavo e luce funzioni in modo casuale
 - quando ci sono osservazioni strane rispetto al modello originario – impossibili o estremamente inverosimili – la probabilità che il modello sia inappropriato aumenterà

Esercizio — Simulare l'uso di una di queste BN o una a piacere attraverso

- gli **strumenti** contenuti nel sito del libro di testo
- nella sezione **Link** di questo blocco

INFERENZA PROBABILISTICA

Problema: calcolare distribuzioni

- **compito tipico:** calcolare la **distribuzione a posteriori** di una o più variabili di query data dell'evidenza
- problema complesso: NP-hard² soluzioni approssimate
 - o #NP (NP-sharp) per il calcolo della probabilità a posteriori o a priori di una variabile

Approcci principali all'**inferenza probabilistica** con le BN:

- inferenza esatta
- inferenza approssimata

INFERENZA ESATTA

Probabilità calcolate *precisamente*

1. *Enumerazione* dei mondi consistenti (i.e. coerenti) con l'evidenza
2. Sfruttando la struttura della rete:
 - *algoritmo di eliminazione delle variabili*, basato sulla *programmazione dinamica*, sfrutta l'indipendenza condizionata

In [1] (per eventuali progetti sull'argomento): ↵

- **Eliminazione di Variabili per Belief Network**
 - simile all'algoritmo per CSP
- **Rappresentazione di Probabilità Condizionate e Fattori**

INFERENZA APPROSSIMATA

Metodi per stimare le probabilità:

- forniscono $[l, u]$ **limiti garantiti** di variazione per la probabilità esatta p
 - un algoritmo *anytime* garantisce che l e u si avvicinino col passare del tempo (o con più spazio a disposizione)
- producono **limiti probabilistici** sull'errore, garantendo:
 - errore (basso) per una certa percentuale di volte
 - stime di probabilità convergenti verso quella esatta
 - velocità di convergenza → es. *simulazione stocastica*
- **inferenza variazionale** dà (quasi sempre) buone *approssimazioni*:
 - si sceglie una *classe* di rappresentazioni più *semplici* (complessità)
 - e.g. delle reti non connesse (senza archi)
 - nella classe si trova il modello più vicino al problema originario
 - distribuzione a posteriori, *facile* da calcolare, vicina a quella cercata
 - problema di minimizzazione dell'errore, seguito da uno di inferenza

MODELLI PROBABILISTICI SEQUENZIALI

BN speciali con una struttura che *si ripete*

- per ragionare sul *tempo* o su altri tipi di *sequenze*
 - es. parole in una frase
- numero *illimitato* di variabili casuali

Cenni su:

- Catene di Markov
- Modelli Nascosti di Markov
- Reti Bayesiane Dinamiche

Catena di Markov (MC) — BN con le variabili in una *sequenza*: ogni variabile dipende direttamente solo dalla precedente

- usate per rappresentare sequenze di (valori o) *stati* (spazio finito | enumerabile)
 - ad es. sequenze di stati in un sistema dinamico o di parole in un testo
 - **PageRank** si basa su questo modello (cfr. specchietto nel testo)
- **stage**: punto/posto nella sequenza



(frammento di) MC come BN: può estendersi indefinitamente

TEMPO E INDIPENDENZA: MEMORIA

Sequenze spesso intese *nel tempo*:

BN → dipendenza dal solo genitore:
assunzione di Markov (*memorylessness*)

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

- S_t rappresenta lo **stato** al tempo t
 - intuitivamente, S_t porta con sé tutte le informazioni *storiche* che possono influenzare gli stati futuri

*“il futuro è condizionatamente indipendente
dal passato dato il presente”*

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

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

- definibile attraverso da due sole probabilità (condizionate):
 - $P(S_0)$ che specifica le condizioni iniziali
 - $P(S_{i+1} \mid S_i)$ che specifica le *dinamiche*, le stesse per ogni $i \geq 0$
 - dinamica del mondo che non cambia nel tempo
- con pochi parametri si specifica una struttura *infinita*
- è possibile poi porre domande su punti arbitrari nel futuro o nel passato

Per determinare la distribuzione dello stato S_i , si può usare l'**eliminazione di variabili** per sommare su tutte le precedenti (successive irrilevanti)

- per calcolare $P(S_i | S_k)$:
 - se $i > k$, basta considerare solo le variabili tra S_i e S_k
 - se $i < k$, solo quelle di indice inferiore a k

Distribuzione stazionaria di una MC:

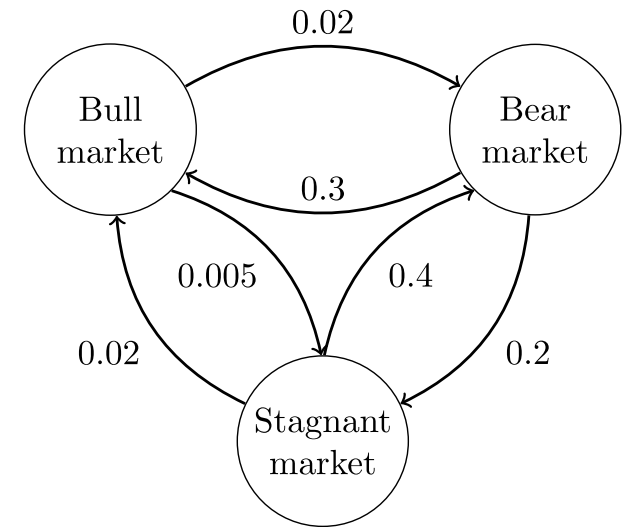
se vale una volta varrà anche per la successiva

- in tal caso, per ogni stato s : $P(S_{i+1} = s) = P(S_i = s)$
quindi:

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

Esempio — Mercato finanziario

- spazio stati: {Bull, Bear, Stagnant}
- probabilità delle transizioni: diagramma →
 - anche attraverso matrice
- distribuzione stazionaria:
 $\pi = \langle 0.885, 0.071, 0.044 \rangle$



da [Wikipedia](#)

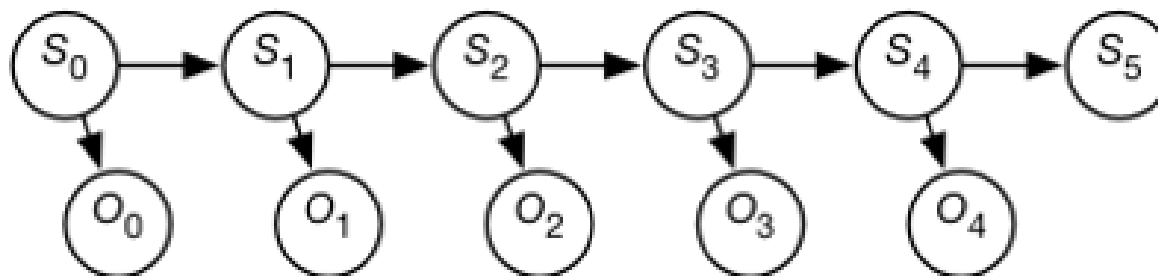
MC: PROPRIETÀ

- MC **ergodica**: per ogni coppia di stati s_1 e s_2 nel dominio di S_i , probabilità non nulla di raggiungere s_2 da s_1
- MC **periodica**: *periodo* p sottomultiplo delle differenze tra visite di uno stesso stato altrimenti **aperiodica** ($p = 1$)
 - ad es., MC con spazio-stati $\{0, \dots, 9\}$ e che ogni volta sceglie con probabilità 0.5 se aggiungere 1 o 9 (modulo 10) *periodica* con $p = 2$:
 - partendo da uno stato pari a tempo 0, sarà in stati pari in istanti pari e dispari altrimenti
- 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

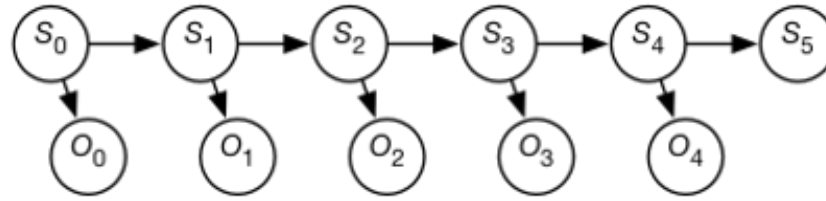
Esercizio — Simulare una MC attraverso il [playground di Setosa.io](#) visti gli esempi [\[MCVE\]](#).

Modello di Markov Nascosto o *Hidden Markov Model* (HMM)

catena di Markov aumentata con l'aggiunta di nodi per le *osservazioni*:



- oltre alle transizioni di stato, per ogni istante di tempo t osservazione O_t (variabile) che dipende da S_t (e da t)
 - *dominio*: insieme delle possibili osservazioni
 - o. *parziali*: stati diversi mappati su una stessa osservazione
 - o. *rumorose*: in uno stesso stato, diverse osservazioni in momenti diversi, casualmente



HMM estendibile indefinitamente

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* comprende le seguenti distribuzioni:

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

Belief Network Dinamica (DBN):

BN con una struttura regolare ripetuta nel tempo (discreto)

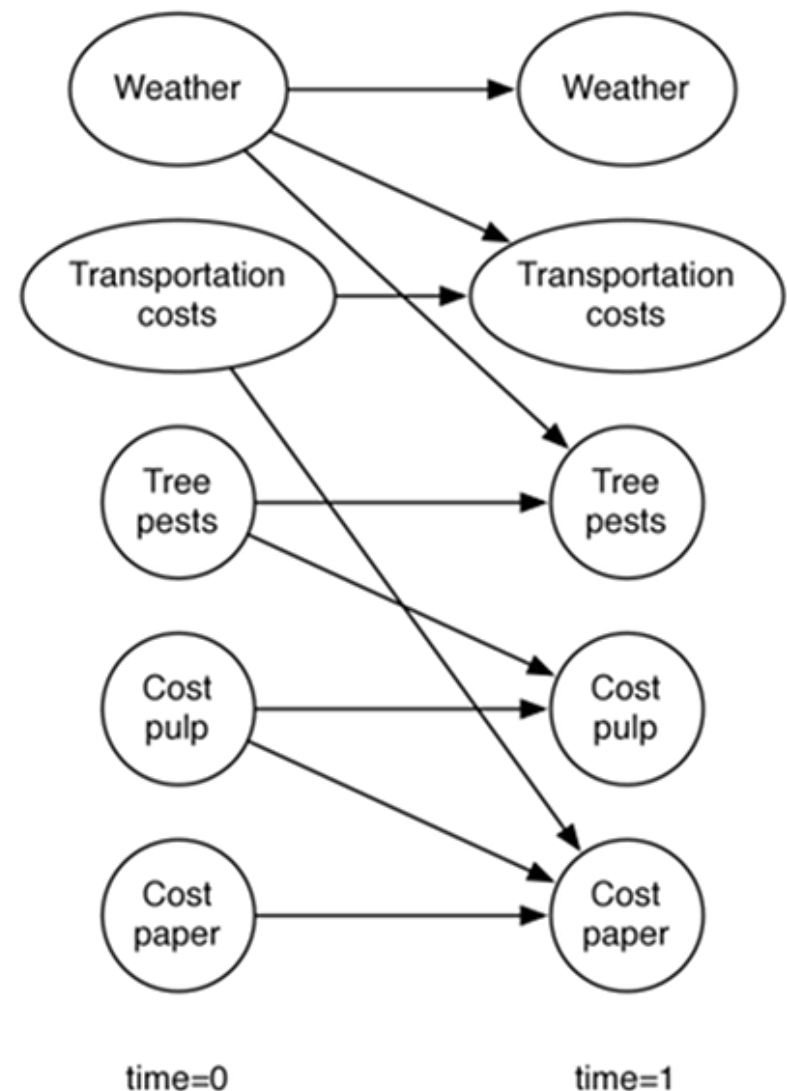
- Come HMM, ma stati/osservazioni rappresentati attraverso *feature* (anziché da una sola var.):
 - F feature $\rightarrow F_t$ variabile per il valore di F al tempo t

Assunzioni per le DBN:

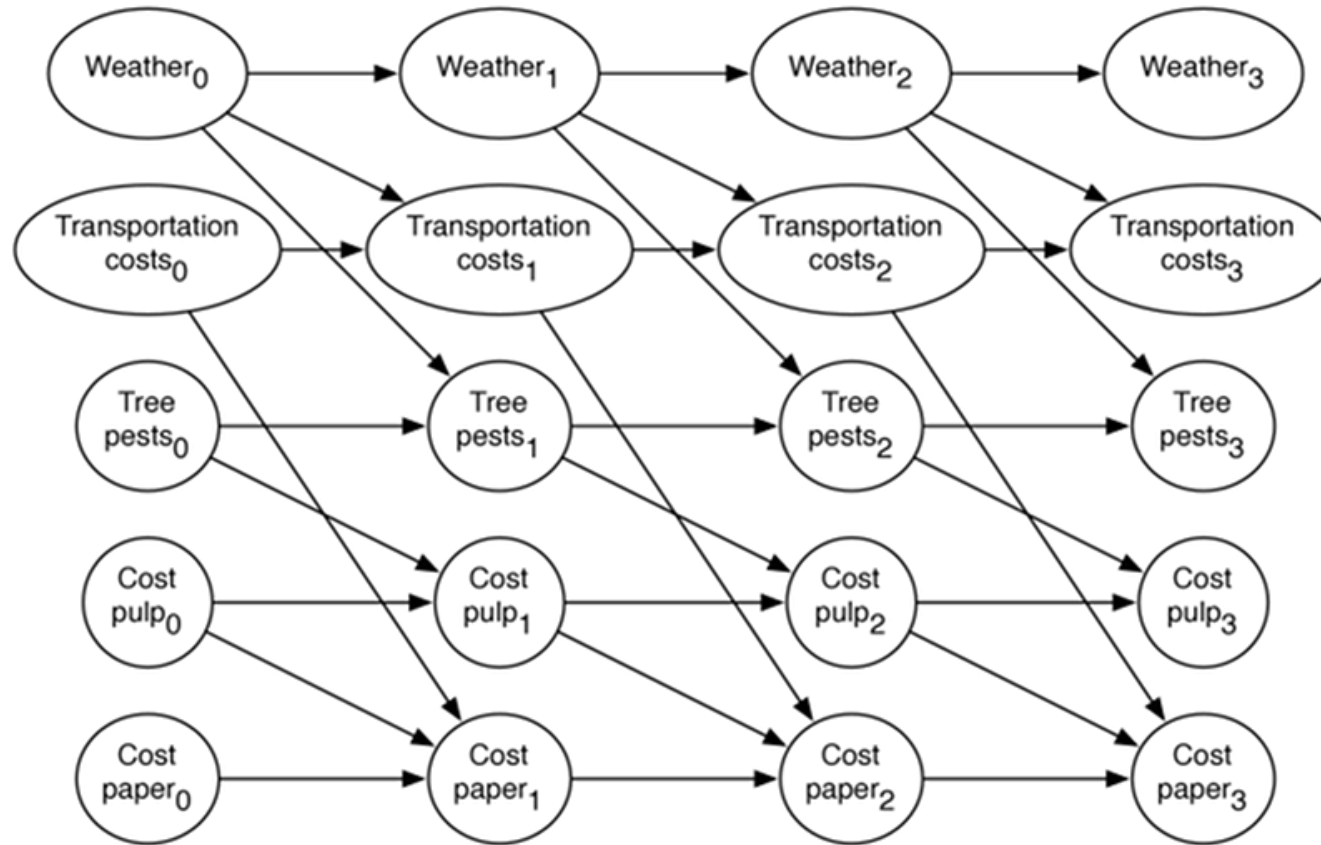
- insieme di feature fissato nel tempo
- per ogni $t > 0$, genitori di F_t variabili relative a t o $t - 1$ (grafo aciclico)
 - la struttura non dipende dal valore di t
(tranne per $t = 0$, caso speciale)
- per ogni $t > 0$ stessa distribuzione condizionata, i.e. dipendenza dai genitori: **modello stazionario**

DBN rappresentabile come BN a due passi con 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
 - grafo risultante aciclico
- probabilità associate alla rete:
 $P(F_0 \mid parents(F_0))$ 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 $P(F_i \mid \text{parents}(F_i))$, per $i > 1$, avrà la stessa struttura e le stesse probabilità condizionate di $P(F_1 \mid \text{parents}(F_1))$

SIMULAZIONE STOCASTICA

Spesso problemi troppo complessi per un'efficiente inferenza esatta

→ *inferenza approssimata*

- si ricorre a metodi basati sulla generazione di *campioni* casuali della distribuzione (a posteriori) specificata dalla rete

Simulazione stocastica insieme di campioni mappati *su* e *da* probabilità

- *inferenza*: si va dalle probabilità ai campioni e viceversa
 - ad es., probabilità $P(a) = 0.25$ indica che su N campioni, per circa un quarto di essi a risulterà vera

Problemi tipici:

1. *generare* i campioni
2. *inferire* probabilità dai campioni
3. *incorporare* le osservazioni

Metodi Monte Carlo

Basi per metodi che usano il campionamento per calcolare la distribuzione a posteriori di una variabile in una BN:

- *forward sampling*
 - *rejection sampling* ✗
 - *importance sampling* ✗
 - *particle filtering* ✗
- *Markov Chain Monte Carlo* (MCMC)
 - *Gibbs sampling*

si veda anche [3]

Generazione campioni dalla distribuzione di una sola variabile X :

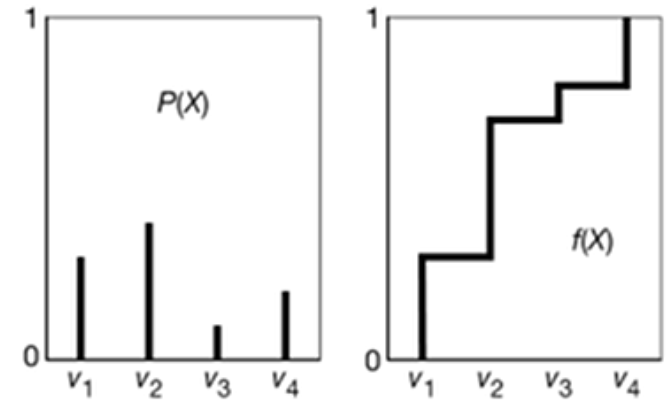
- ordinare i valori del dominio
- definire la *distribuzione cumulativa*³ in funzione di x :

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

- per generare un *campione casuale* v per X :
 - generare y da una distribuzione uniforme su $[0, 1]$
 - ricavare $v \in \text{dom}(X)$ che abbia y come immagine nella cumulativa:
 - i.e. tale che $f(v) = y$, ossia $v = f^{-1}(y)$

Esempio — Sia X con dominio $\{v_1, v_2, v_3, v_4\}$ ordinato ($v_1 < v_2 < v_3 < v_4$) e distribuzione $P(X)$:

- $P(X = v_1) = 0.3, P(X = v_2) = 0.4,$
 $P(X = v_3) = 0.1, P(X = v_4) = 0.2$
- definita la cumulativa $f(v_i) = P(X \leq v_i)$:
 - 30% del codominio di f ha v_1 come contro-immagine: possibilità di essere campionato, i.e. se $y \in [0, 0.3]$
 - analogamente, 40% per v_2 , ecc.



Cumulativa da istogramma

Stima delle probabilità da un insieme di campioni attraverso la *media campionaria*:

- data una proposizione α ,
 s proporzione di campioni con α *vera* rispetto al totale n
- per la *legge dei grandi numeri*:⁴
 s si avvicina asintoticamente alla probabilità esatta p al crescere di n

Stima dell'errore ϵ attraverso la *disuguaglianza di Hoeffding*:

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

- da cui si può ricavare n che garantisce una stima *probabilmente approssimativamente corretta* della probabilità

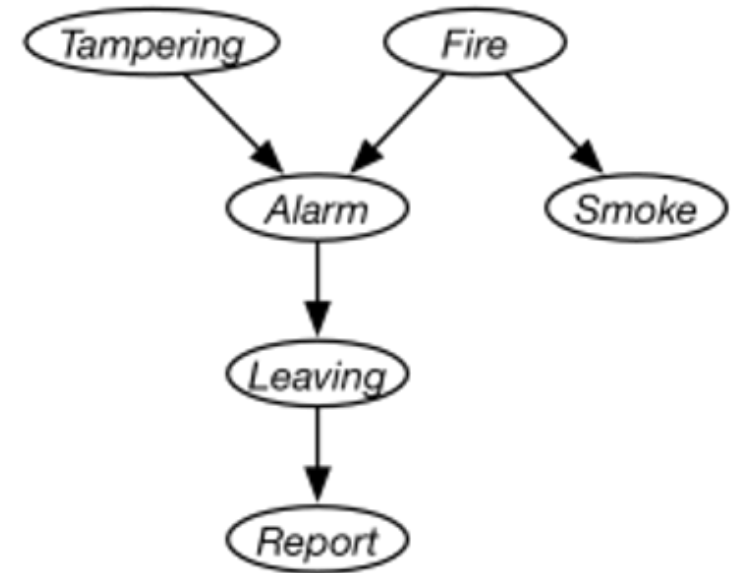
Campionare le variabili di una BN
per *stimare* le *probabilità* a priori di qualsiasi variabile

- Dato un ordinamento X_1, \dots, X_n delle variabili, tale che ognuna sia preceduta dai genitori, per estrarre un campione/tupla (della distribuzione congiunta):
 - X_1 si campiona usando la cumulativa
 - per i da 2 a n :
 - data X_i , valori della n -pla per i genitori già disponibili
 - si campiona un valore per X_i dalla distribuzione di X_i , dati i valori già assegnati ai suoi genitori

Stima della distribuzione di una variabile di query considerando la proporzione di campioni assegnati a ogni valore della variabile

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

- ripetere: costruzione di un campione
 1. si campiona *Tampering* con la cumulativa
 - ad es. *Tampering* = *false*
 2. si campiona *Fire* analogamente
 - ad es. *Fire* = *true*
 3. si campiona *Alarm*, usando $P(\textit{Alarm} \mid \textit{Tampering} = \textit{false}, \textit{Fire} = \textit{true})$
 - ad es. *Alarm* = *true*
 4. si campiona *Smoke* usando $P(\textit{Smoke} \mid \textit{Fire} = \textit{true})$
 5. ecc.
- fino a ottenere il numero desiderato di campioni ($\{s_1, \dots, s_n\}$)



campione	<i>Tampering</i>	<i>Fire</i>	<i>Alarm</i>	<i>Smoke</i>	<i>Leaving</i>	<i>Report</i>
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

Es. stima della probabilità di *Report* = *true*:

- proporzione dei campioni con *Report* vera

Metodo che prescinde dall'ordine di campionamento

Distribuzione **stazionaria** di una catena di Markov: distribuzione delle sue variabili non modificata dalla funzione di transizione della catena

- se la catena mescola sufficientemente, c'è un'unica distribuzione stazionaria
- approssimabile facendo **girare** il modello sufficientemente a lungo

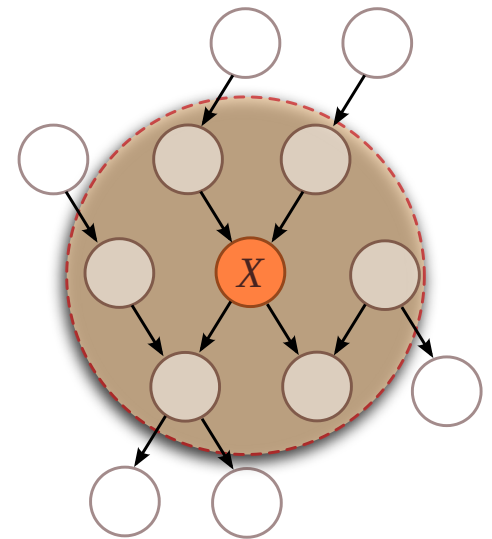
Metodi **Markov Chain Monte Carlo** (MCMC) per generare campioni (assegnazioni complete):

- si costruisce una **catena di campioni** con la distribuzione desiderata come sua (sola) distribuzione stazionaria
- quindi si campiona dalla catena
 - campioni distribuiti secondo la distribuzione-obiettivo
- **riscaldamento/burn-in**: si possono scartare i primi campioni, probabilmente lontani dalla distribuzione stazionaria

GIBBS SAMPLING

Crea una catena di campioni da una BN con variabili osservabili fissate ai valori osservati (evidenza) e campionando le altre

- si genera un nuovo campione dal precedente, modificando i valori non fissati
 - anche per una sola variabile alla volta [3]
 - variabili campionate in base alle loro distribuzioni condizionate ai valori correnti delle altre
- una X in una BN dipende solo dai valori delle variabili nel suo **Markov blanket** $mb(X)$ con:
 - i suoi genitori
 - i suoi figli
 - gli altri genitori dei suoi figli



procedure Gibbs_sampling($B, e, Q, n, burn_in$):

Input

B : belief network

e : evidenza; assegnazione di valori ad alcune variabili

Q : variabile 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

Osservazioni

- Si avvicina alla distribuzione reale se NON ci sono probabilità *nulle*
- La *velocità* dipende da quella del rimescolamento delle probabilità (da quanto spazio delle probabilità viene esplorato) che a sua volta dipende da quanto esse siano *estreme*
 - funziona bene per probabilità non estreme

Esempio — BN $A \rightarrow B \rightarrow C$ con A, B, C booleane

- CPT:
 - $P(a) = 0.5$
 - $P(b \mid a) = 0.99$
 - $P(b \mid \neg a) = 0.01 \rightarrow P(\neg b \mid \neg a) = 0.99$
 - $P(c \mid b) = 0.99$
 - $P(c \mid \neg b) = 0.01 \rightarrow P(\neg c \mid \neg b) = 0.99$
- nessuna osservazione e variabile di query: C

Due assegnazioni con tutte le variabili con lo stesso valore (vero/falso)

- stessa probabilità e più probabili di ogni altra assegnazione

Il Gibbs sampling porta *rapidamente* verso una di tali assegnazioni

- mentre richiederebbe molto tempo per arrivare alle altre
 - scelte molto improbabili
 - sostituendo **0.99** e **0.01** con numeri ancor più vicini a **1** e **0** convergenza più lenta

Esercizio — Provare la demo

RIFERIMENTI

- [1] D. Poole, A. Mackworth: *Artificial Intelligence: Foundations of Computational Agents*. Cambridge University Press. 2nd Ed. [Ch.8]
- [2] D. Poole, A. Mackworth, R. Goebel: *Computational Intelligence: A Logical Approach*. Oxford University Press
- [3] S. J. Russell, P. Norvig: *Artificial Intelligence* Pearson. 4th Ed. [ch.14] - cfr. anche ed. Italiana (cap. 12-13)
- [4] J. Sowa: *Knowledge Representation: Logical, Philosophical, and Computational Foundations* Brooks Cole/Cengage
- [5] J. Pearl: *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann (1988)
- [6] J. Pearl: *Causality: models, reasoning and inference*. 2nd edition, Cambridge University Press (2009)
- [7] J. Y. Halpern: *Reasoning about uncertainty*. MIT Press (2003)
- [8] D. Koller, N. Friedman: *Probabilistic graphical models: principles and techniques*. MIT Press (2009)
- [9] D. MacKay: *Information theory, inference, and learning algorithms*. Cambridge University Press (2003)

LINK

[Informazione] Si veda [sito](#) di [9] oppure [Wikipedia](#)

[BN] [it.Wikipedia](#) e [en.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](#)

[MCMC] su [Wikipedia](#) con puntatori a implementazioni

NOTE

¹ Thomas Bayes su [it.Wikipedia](#), [en.Wikipedia](#).

² **#NP** ("sharp-NP") classe di complessità del calcolo delle probabilità a priori o a posteriori di una variabile (richiede conteggi, non solo decisioni).

³ Funzione di ripartizione o cumulativa: [Wikipedia](#).

⁴ 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 ; cfr. [Wikipedia](#).

[◀] consigliata la lettura

[versione] 5/12/2022, 19:06:50

Figure tratte da [1] salvo diversa indicazione

formatted by [Markdeep 1.14](#) 