Laurea Magistrale in Informatica A.A. 2020/2021
Università degli Studi di Milano-Bicocca

Appunti Modelli Probabilistici per le Decisioni

Marta Pelusi
@Marta629

1. Regola di Bayes

$$\mathrm{P}(\text{causa}|\text{effetto}) = \frac{\mathrm{P}(\text{effetto}|\text{causa})}{\mathrm{P}(\text{effetto})} = \frac{\mathrm{P}(\text{effetto}|\text{causa})\mathrm{P}(\text{causa})}{\sum_{\mathrm{h}\in\text{causa}} \mathrm{P}(\text{effetto}|\mathrm{h})\mathrm{P}(\mathrm{h})}$$

Si può calcolare il termine superiore e poi normalizzare (dividere per la somma dei termini al numeratore di tutte le ipotesi)

$$\mathrm{P}(\text{causa} =< \mathrm{v}, \mathrm{f} > |\text{effetto} = \mathrm{v}) = \alpha$$
$$< \mathrm{P}(\text{effetto}|\text{causa})\mathrm{P}(\text{causa}), \mathrm{P}(\text{causa}|\neg\text{effetto})\mathrm{P}(\neg\text{effetto}) >$$

Significato del teorema di Bayes – apprendimento dall'esperienza:

$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \alpha\mathrm{P}(\mathrm{B}|\mathrm{A})\mathrm{P}(\mathrm{A})$$

- P(A) probabilità a priori
- P(A|B) probabilità a posteriori
- P(B|A) verosimiglianza
- α fattore di normalizzazione

1.1 Catena di probabilità condizionali

P(A,B) è detta probabilità congiunta e le si può applicare la regola del prodotto:

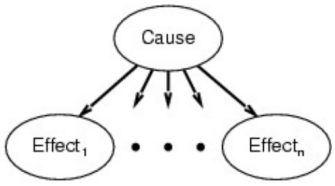
$$\mathrm{P}(\mathrm{A}, \mathrm{B}) = \mathrm{P}(\mathrm{A}|\mathrm{B})\mathrm{P}(\mathrm{B})$$

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

$$\mathrm{P}(\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}) = \mathrm{P}(\mathrm{A}|\mathrm{B}, \mathrm{C}, \mathrm{D})\mathrm{P}(\mathrm{B}|\mathrm{C}, \mathrm{D})\mathrm{P}(\mathrm{C}|\mathrm{D})\mathrm{P}(\mathrm{D})$$

1.2 Reti Bayesiane

Le reti Bayesiane permettono di rappresentare le assunzioni di indipendenza condizionale tra variabili in modo chiaro ed efficiente utilizzando la rappresentazione basata su grafi diretti aciclici.



1.3 Indipendenza condizionale

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

$$\mathrm{P}(\mathrm{A}|\mathrm{B},\mathrm{C}) = \mathrm{P}(\mathrm{A}|\mathrm{C})$$

Ovvero A è condizionalmente indipendente da B dato C se la conoscenza di B non porta a nessuna ulteriore variazione della probabilità di A rispetto a quella apportata dall'avversarsi di C. Così per quanto riguarda A, se conosciamo C possiamo ignorare B.
Dall'indipendenza condizionale di A e B dato C otteniamo:

$$\mathrm{P}(\mathrm{A},\mathrm{B}|\mathrm{C}) = \mathrm{P}(\mathrm{A}|\mathrm{C})\mathrm{P}(\mathrm{B}|\mathrm{C}) = \mathrm{P}(\mathrm{B},\mathrm{A}|\mathrm{C})$$

Se C è un insieme vuoto otteniamo:

$$P(A,B) = P(A)P(B)$$

2. Inferenza

La probabilità a priori o incondizionata traduce la nostra convinzione rispetto alla verità della proposizione in assenza di evidenza. La distribuzione di probabilità fornisce valori per tutti i possibili assegnamenti. La distribuzione di probabilità congiunta per un dato insieme di variabili aleatorie fornisce il valore di probabilità associato ad ogni evento atomico costituito da realizzazione congiunte. Essa descrive completamente il dominio caratterizzato tramite variabili aleatorie.

È possibile scrivere in forma generale la regola di marginalizzazione per due insiemi di variabili Y e Z:

$$P(Y) = \sum_{z} P(Y,z)$$

In alternativa è possibile usare la probabilità condizionale, ovvero applicando la regola nota col nome di regola del condizionamento:

$$P(Y) = \sum_{z} P(Y|z)P(z)$$

Indichiamo con X la variabile oggetto della query e sia data la realizzazione congiunta e (evidenza) per un sottoinsieme E di variabili dette variabili con evidenza o evidenziate. Si indichi con Y l'insieme delle restanti variabili sulle quali non è disponibile evidenza, che viene detto insieme delle variabili senza evidenza. L'intero insieme delle variabili che caratterizzano il problema viene ottenuto tramite l'operazione di unione:

$$\{X\} \cup E \cup Y$$

La distribuzione marginale a posteriori per la variabile X viene ottenuta applicando un procedimento di somma (marginalizzazione) rispetto all'insieme delle variabili senza evidenza Y, ovvero

$$P(X|E = e) = \alpha P(X, E = e) = \alpha \sum_{y} P(X, E = e, Y = y)$$

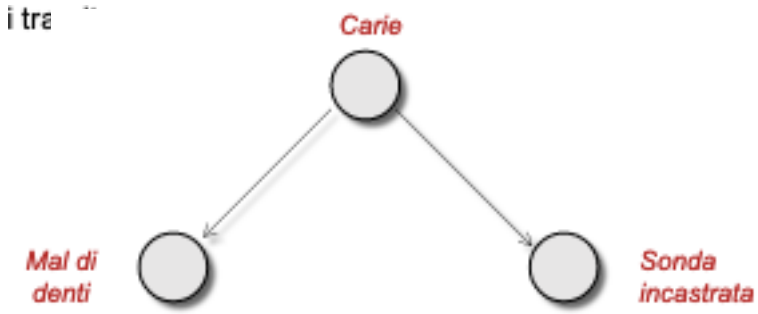
L'indipendenza condizionata consente di limitare fortemente la complessità del modello.
Ricordiamo la regola di Bayes che tipicamente viene impiegata per calcolare la probabilità della causa data la conoscenza dello stato degli effetti:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y)$$

La seguente uguaglianza

$$P(m, s, c) = \alpha P(m, s|c)P(c) = \alpha P(m|c)P(s|c)P(c)$$

può essere rappresentata in termini grafici tramite il seguente grafico:



3. Reti Bayesiane

I concetti di indipendenza e indipendenza condizionata tra variabili offrono la possibilità di memorizzare e trattare in modo efficiente distribuzioni congiunte di probabilità con dimensione elevata.

Una rete Bayesiana è un grafo orientato aciclico in cui i nodi sono annotati con una trasformazione quantitativa (tabelle di probabilità condizionata CPT) e i cui archi definiscono dipendenza e indipendenza condizionale tra le variabili rappresentate dai nodi. Il grafo orientato è costituito da:

- Nodo, associato ad una variabile, relazione 1-1 tra nodo e variabile (aleatoria)
- Arco orientato, che collega due nodi e traduce di norma una relazione di causalità diretta

Le variabili possono essere continue o discrete.

Di norma diremo che X è causa diretta di Y e che X è genitore di Y, quindi Y è figlio di X:

$$parents(Y) = X, \qquad parents(X) = \emptyset$$

3.1 Semantica delle reti Bayesiane

La semantica delle reti Bayesiane può essere presentata e compresa in base alle seguenti chiavi di lettura:

- La rete rappresenta una distribuzione congiunta di probabilità
- La rete codifica un insieme di relazioni di indipendenza condizionale

Ogni rete Bayesiana costituisce una descrizione completa del dominio che rappresenta e pertanto ogni elemento della distribuzione di probabilità congiunta può essere calcolato a partire dall'informazione contenuta nella rete. Un generico elemento della distribuzione di probabilità congiunta è associato ad una realizzazione congiunta delle variabili (nodi) presenti nella rete:

$$P(X_1 = x_1 \wedge \dots \wedge X_n = x_n)$$

Ogni elemento della distribuzione congiunta di probabilità può essere calcolato sfruttando la seguente formula di fattorizzazione della distribuzione congiunta di probabilità:

$$P(x_1, \dots x_n) = \Pi_{i=1}^{n} P\big(x_i \big| parents(X_i)\big)$$

La regola della probabilità congiunta la possiamo scrivere come:

$$\begin{aligned} P(x_1, \dots, x_n) &= P(x_n|x_{n-1}, \dots, x_1)P(x_{n-1}|x_{n-2}, \dots, x_1) \cdot \dots \cdot P(x_2|x_1)P(x_1) \\ &= \Pi_{i=1}^{n} P(x_i|x_{n-1}, \dots x_1) \end{aligned}$$

Questa uguaglianza è vera per ogni insieme di variabili aleatorie ed è nota con il termine di chain rule. Confrontando la chain rule con la formula di fattorizzazione è possibile verificare che la specificazione della distribuzione di probabilità congiunta è equivalente all'asserzione generale che, per ogni variabile X_i:

$$P(x_i|x_{i-1}, \dots, x_1) = P\big(x_i \big| parents(X_i)\big)$$

A patto che $parents(x_i) \subseteq \{x_{i-1}, \dots, x_1\}$

Una rete Bayesiana rappresenta correttamente un dominio solo a condizione che ogni nodo risulti condizionalmente indipendente dai suoi predecessori per un dato

ordinamento, dati i suoi genitori. Pertanto per costruire una rete Bayesiana che abbia la corretta struttura del dominio da modellare è necessario scegliere, per ogni nodo, i nodi genitore in modo che tale proprietà risulti verificata. Intuitivamente, l'insieme dei genitori per ogni nodo X_i, ovvero tutti i nodi per influenzeranno direttamente il nodo X_i, devono poter essere scelti tra $X_{i-1}, \dots, X_1$.

Procedura per la costruzione incrementale della componente topologica di una rete Bayesiana:

1. Selezionare un insieme di variabili (aleatore) $\{X_1, \dots, X_n\}$ da utilizzare per descrivere il dominio da modellare
2. Scegliere un ordinamento delle variabili $\{X_{(1)}, \dots, X_{(n)}\}$
3. Inizializzare il numero di nodi aggiunti alla rete ad uno (i=1)
4. Selezionare la variabile $X_{(i)}$ e aggiungere il nodo corrispondente nella rete
    a. Porre $parents(X_{(i)})$ uguale all'insieme minimale di nodi, attualmente appartenenti alla rete $\{X_{(1)}, \dots, X_{(i-1)}\}$ che soddisfa la proprietà di indipendenza condizionale

$$P(X_{(i)} | X_{(i-1)}, \dots, X_{(1)}) = P\left(X_{(i)} \middle| parents(X_{(i)})\right)$$

    b. Computare la CPT per la variabile $X_{(i)}$
5. Incrementare il numero di nodi aggiunti alla rete (i=i+1) se si sono aggiunte tutte le variabili alla rete (i>n) allora la procedura termina, in caso contrario tornare al passo 4

3.2 Relazioni di indipendenza condizionale nelle reti Bayesiane

Abbiamo proposto una semantica numerica per le reti Bayesiane in termini di rappresentazione di una distribuzione congiunta di probabilità

$$P(x_1, \dots, x_n) = \Pi_{i=1}^{n} P(x_i | parents(X_i))$$

L'utilizzo di tale semantica ha consentito di derivare un metodo per la costruzione di reti Bayesiane che ha posto in evidenza come ogni nodo risulti essere condizionalmente indipendente dai suoi predecessori data la conoscenza dello stato dei suoi nodi genitore:

$$P(X\_((i)) \mid X\_((1)), \dots, X\_((i-1))) = P(X\_i | parents(X\_((i))))$$

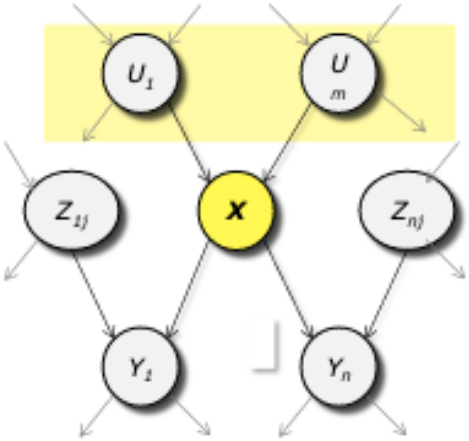
Dato un ordinamento di variabili $\{X_{(1)}, \dots, X_{(n)}\}$ per ogni variabile (nodo) X_i i suoi predecessori o non discendenti sono i nodi $\{X_{(1)}, \dots, X_{(i-1)}\}$, mentre i suoi successori o discendenti sono i nodi $\{X_{(i+1)}, \dots, X_{(n)}\}$.

È possibile procedere anche secondo la direzione opposta, ovvero partire da una semantica topologica che specifica le relazioni di indipendenza condizionale

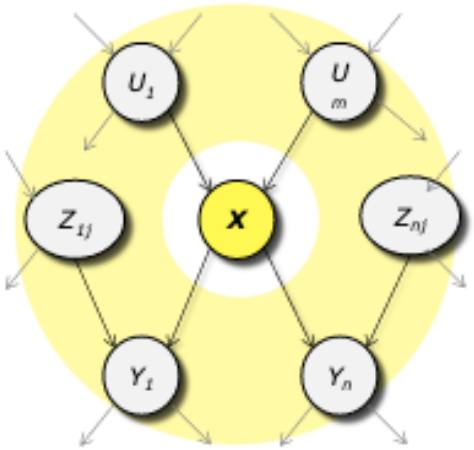
codificate dalla componente strutturale della rete Bayesiana e giungere a ricavarne una semantica numerica.

La semantica topologica viene fornita per mezzo di una delle due specificazioni equivalenti:

4. Un nodo è condizionalmente indipendente dai suoi predecessori o non-discendenti dati i suoi genitori
5. Un nodo è condizionalmente indipendente da tutti i nodi restanti della rete, data la conoscenza dello stato dei suoi genitori, dei suoi figli e dei genitori dei suoi figli, ovvero l'insieme di nodi che è noto col nome di Markov Blanket



Il nodo **X** è *condizionalmente indipendente dai* suoi *non-discendenti* (Z_{ij}) *data la conoscenza* dello *stato* dei suoi *genitori* (U_i)



Il nodo **X** è *condizionalmente indipendente* da tutti i *nodi restanti* della rete *data* la *conoscenza* dello *stato* del suo *Markov Blanket* (U_i, Z_{ij},Y_j).

4. d-separazione delle variabili

Def: Due variabili X e Z sono d-separate da un insieme E di variabili con evidenza (osservazioni) se e solo se ogni cammino non orientato da X a Z è "bloccato".

Def: Un cammino è "bloccato" se e solo se vale almeno una delle seguenti condizioni:

- Lungo il cammino esiste una variabile $V \in E$ tale che ( ) $\leftarrow (V) \rightarrow$ ( )
- Lungo il cammino esiste una variabile $V \in E$ tale che ( ) $\rightarrow (V) \rightarrow$ ( )
- Lungo il cammino esiste una variabile $V \notin E$ (e nessuno dei suoi discendenti appartiene all'insieme E) tale che ( ) $\rightarrow (V) \leftarrow$ ( )

Teorema: se in una rete Bayesiana un insieme e di variabile con evidenza d-separa X e Z, allora X e Z sono indipendenti.

4.1 Procedura di moralizzazione

Procedura per verificare se due variabili sono d-separate (solo C.S.), ovvero
- Se le variabili risultano d-separate tramite la moralizzazione => sicuramente sono d-separate
- Se le variabili risultano non essere d-separate tramite la moralizzazione => non possiamo escludere che siano d-separate (non so se sono veramente non d-separate, quindi potrebbero essere lo stesso d-separate)

Riponde alla domanda: "sono A, B indipendenti date D, F?", ovvero $P(A|BDF) = P(A|DF)$?

1. Disegnare il grafo ancestrale (grafo degli antenati) di tutte le variabili menzionate nell'espressione
2. Moralizzo il grafo combinando i matrimoni tra genitori, ovvero per ogni coppia i variabili che hanno in comune un figlio si traccia un arco (anche se sono più di 2 genitori)
3. Disorientazione del grafo togliendo le frecce (diventa un grafo non orientato)
4. Cancellare tutte le variabili che sono osservare e tutti gli archi relativi a queste variabili
5. Leggere la risposta:
    a. Variabili sconnesse => variabili d-separate => variabili indipendenti
    b. Variabili connesse => non è garantito che queste variabili siano indipendenti
    c. Nessuna variabile => variabili d-separate => variabili indipendenti

5. Rappresentazione efficiente delle distribuzioni condizionali

Può esistere una distribuzione canonica che si adatta a rappresentare diversi pattern standard; in questo caso è possibile, una volta identificato il pattern, specificare un numero limitato di parametri per compilare le CPT. L'esempio più semplice di tali pattern è rappresentato dai nodi deterministici. Un nodo deterministico è caratterizzato dal fatto che il valore che esso assume è completamente determinato tramite il valore assunto dai suoi genitori; non vi è alcuna incertezza.
La relazione potrebbe essere:
- Logica (disgiunzione, congiunzione, …)
- Numerica (minimo, massimo, differenza, …)

Relazioni incerte possono essere caratterizzate per mezzo delle cosiddette Noisy Logical Relationships. L'esempio standard è offerto dal modello Noisy-OR che generalizza la funzione logica OR.
Il modello Noisy-OR consente di introdurre incertezza circa la capacità di ogni nodo genitore di causare il valore vero (v) della variabile figlio (la relazione di

causalità tra genitore e figlio potrebbe essere inibita. Questo modello effettua le seguenti ipotesi:

- Tutte le possibili cause sono note (eventualmente si può aggiungere leak node)
- L'inibizione di un genitore è indipendente dall'inibizione degli altri genitori per il nodo considerato

La discretizzazione costituisce una buona soluzione anche se spesso porta ad una considerevole perdita di accuratezza e a CPTs di grandi dimensioni. La soluzione alternativa consiste nel definire famiglie standard di funzioni di densità di probabilità che vengono descritte tramite un numero finito di parametri. Ad esempio la funzione di densità di probabilità Gaussiana univariata $N(\mu, \sigma^2)$ è completamente specificata tramite i seguenti due parametri μ (media) e σ^2 (varianza).

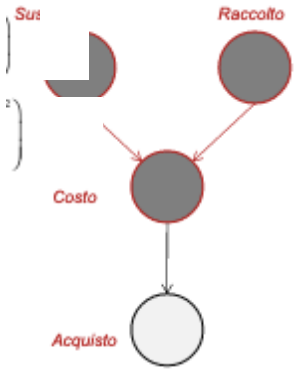
Una rete che contiene nodi discreti e continui viene detta rete Bayesiana ibrida. Per specificare una rete Bayesiana ibrida dobbiamo definire due nuovi tipi di distribuzione:

- La distribuzione condizionale di una variabile continua dati i genitori discreti o continui
- La distribuzione condizionale di una variabile discreta dati i genitori continui

La scelta più comune è costituita dalla distribuzione gaussiana lineare in base alla quale il nodo figlio (variabile aleatoria ad esso associata) è distribuito secondo una distribuzione gaussiana, la cui media $\boldsymbol{\mu}$ varia linearmente con il valore del genitore e la cui deviazione standard $\boldsymbol{\sigma}$ è fissata.
La distribuzione gaussiana lineare gode di alcune proprietà:

- Una rete che contiene solo nodi continui con distribuzione gaussiana lineare ha una distribuzione di probabilità congiunta gaussiana multivariata
- Una rete che contiene nodi continui con distribuzione gaussiana lineare e nodi discreti, nessuno dei quali sia figlio di nodi continui, definisce una distribuzione gaussiana condizionale per ogni assegnamento di valori per le variabili discrete; la distribuzione sulle variabili continue è gaussiana multivariata

Passiamo ora a considerare la distribuzione associata a variabili discrete con genitori continui. Considerando per esempio il nodo acquisto del modello, appare ragionevole che il cliente acquisti se il prezzo (costo) è basso, mentre non acquisti se il prezzo è alto. È altresì ragionevole ipotizzare che la probabilità di acquisto vari in modo morbido in una regione di costo intermedio. In altre parole: la distribuzione condizionale per la variabile acquisto sarà definita tramite una funzione a soglia soft.



Un modo di ottenere soglie soft è tramite l'impiego dell'integrale della distribuzione normale standard:

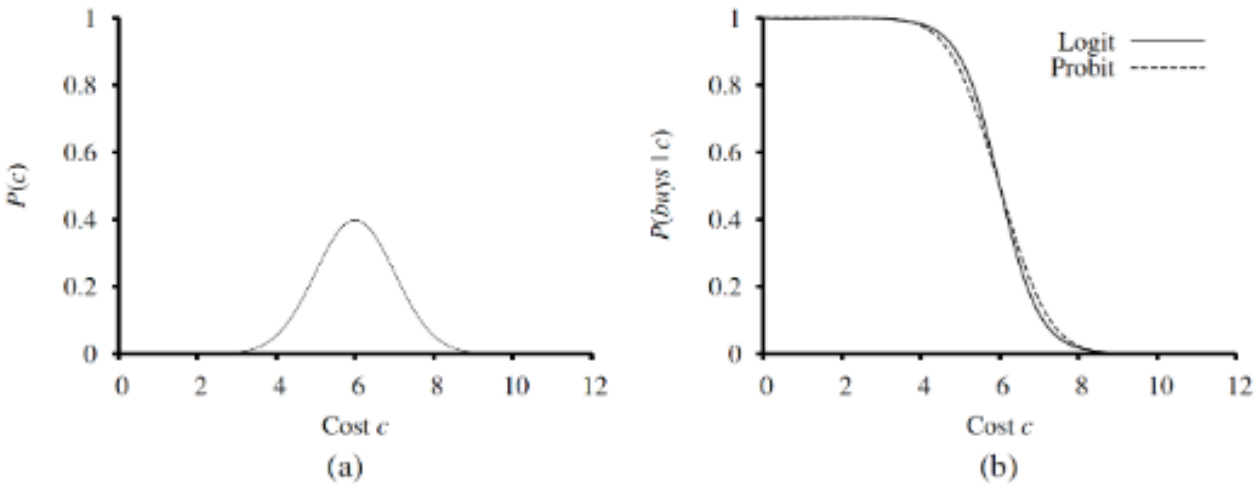
$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$$

Allora la probabilità di acquisto dato costo potrebbe essere definita tramite la distribuzione Probit, come segue:

$$P(\text{acquisto}|\text{costo} = \text{c}) = \phi\left(\frac{-c + \mu}{\sigma}\right)$$

Il che significa che la soglia di costo si verifica intorno al valore μ, lo spessore della regione soglia è proporzionale al valore σ, e la possibilità di acquistare diminuisce all'aumentare del costo. Un'alternativa è offerta dalla distribuzione Logit, che utilizza la funzione sigmoid per ottenere una soglia soft:

$$P(\text{acquisto}|\text{costo} = \text{c}) = \frac{1}{1 + \exp\left(-\frac{2(-c + \mu)}{\sigma}\right)}$$



6. Inferenza nelle reti Bayesiane

Il compito primario di ogni sistema inferenziale probabilistico consiste nel computare la distribuzione a posteriori per un determinato insieme di variabili query, quando si sia osservato un determinato evento, ovvero un assegnamento congiunto di valori ad un insieme di variabili evidenziate (con evidenza).

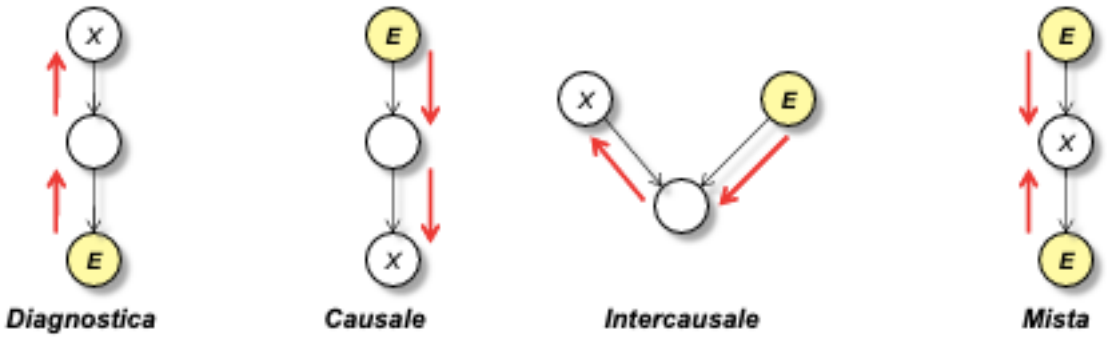
Notazione:
- X variabile query
- E insieme delle variabili con evidenza $\{E_1, \dots, E_m\}$
- E specifico evento (assegnamento congiunto di valori alle variabili evidenziate)
- Y insieme delle variabili non evidenziate $\{Y_1, \dots, Y_l\}$

L'insieme completo delle variabili è pertanto

$$X = \{X\} \cup E \cup Y$$

Una tipica query richiede di computare la distribuzione a posteriori $P(X|E = e)$.

Dato un modello di rete Bayesiana è possibile effettuare le seguenti tipologie di inferenza:



Ricapitolando, ogni distribuzione condizionale può essere ottenuta tramite un procedimento di somma di opportuni elementi dell'intera distribuzione di probabilità congiunta (marginalizzazione). Nello specifico è possibile rispondere ad una query del tipo

$$P(X|E = e)$$

Sfruttando la seguente equazione:

$$P(X|E = e) = \alpha P(X, E = e) = \alpha \sum_{y} P(X, E = e, Y = y)$$

Una rete Bayesiana offre una rappresentazione completa dell'intera distribuzione di probabilità congiunta ed, in base alla regola di fattorizzazione

$$P(x_1, \dots, x_n) = \Pi_{i=1}^{n} P\left(x_i \middle| parents(X_i)\right)$$

Mostra come i termini $P(X = x, E = e, Y = y)$ nella distribuzione congiunta possano essere scritti sottoforma di prodotti di probabilità condizionali della rete.

6.1 Inferenza per enumerazione

L'algoritmo di enumerazione, che consiste nella valutazione ricorsiva dell'albero (costruito sulla base delle CPT della rete Bayesiana per rispondere ad una determinata query), può essere migliorato sostanzialmente tramite l'eliminazione delle ripetizioni di computazione all'interno dell'albero. L'idea consiste nell'eseguire ogni computazione una sola volta e memorizzarne il risultato per utilizzi successivi.
Sono disponibili diversi algoritmi che sfruttano il principio di riduzione e il più semplice è l'algoritmo di eliminazione variabili. Tale algoritmo si basa sul principio che, data una variabile oggetto della query, ogni variabile che non sia predecessore di tale variabile oppure che non sia predecessore di una variabile con evidenza risulta essere irrilevante per la query. Pertanto l'algoritmo di eliminazione variabili, per ogni data query, procede rimuovendo dall'albero tutte le variabili non predecessore

- Della variabile oggetto della query
- Di una variabile con evidenza

E valutando successivamente l'albero risultante per risolvere la query.

6.2 Generazione numeri pseudo-casuali

Def: una sequenza di numeri casuali è una sequenza di realizzazioni di variabili aleatorie indipendenti e identicamente distribuite. Una sequenza di numeri pseudo-casuali è una sequenza di numeri che "sembrano" impredicibili, da cui non si riesce ad estrarre alcuna regolarità.

Proprietà statistiche dei numeri casuali:

- Indipendenza statistica
- Uniformità della distribuzione
- Riproducibilità della sequenza di valori
- Non ripetitività su un prefissato periodo

Generazione di una sequenza di numeri con 10 cifre:

Si parte da un numero, lo si eleva al quadrato e si prendono le 10 cifre centrali, etc…
Nota: la sequenza non è veramente casuale perché ogni numero è determinato dal precedente (ma "sembra" casuale!).

6.2.1 MCM – Moltiplicative Congruential Method

$$x_n = ax_{n-1}, \qquad 0 \leq x_n \leq m$$

- x_0 valore iniziale (seme o seed)
- a, m interi positivi opportunamente scelti (m dovrebbe essere scelto tale che il numero di valori generati prima di una ripetizione sia molto grande; solitamente è scelto come un numero primo molto grande, ad esempio $m = 2^{31} - 1, a = 7^5$)

6.2.2 LCM – Linear Congruential Method

$$x_n = (ax_{n-1} + c) \bmod m$$

- x_0 valore iniziale (seme o seed)
- a, c, m opportunamente scelti con $a, c \geq 0, m > x_0, c, a$

Per "numero random" (o numero casuale) si intende una variabile aleatoria distribuita in modo uniforme tra 0 e 1, quindi si generano numeri interi per poi trasformarli in numeri reali:

$$u_n = \frac{x_n}{m}$$

Questo metodo presenta le seguenti caratteristiche negative:

- È ciclico con periodo circa pari a m
- I numeri generati sono discretizzati, infatti u_k può assumere solo i valori $0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}$, ma in realtà potrebbe assumere qualsiasi valore nell'intervallo [0.5/m, 0.6/m] con probabilità 0.1/m, ma la probabilità che questo avvenga è 0

Scegliendo m abbastanza grande si può ridurre sia il fenomeno della periodicità sia il fatto di generare numeri razionali. Inoltre non è necessario ai fini della simulazione che vengano generati tutti i numeri tra [0,1] (sarebbero infiniti), ma è sufficiente che quanti più numeri possibili all'interno dell'intervallo abbiano la stessa probabilità di essere generati.

6.2.3 Generazione distribuzione generica

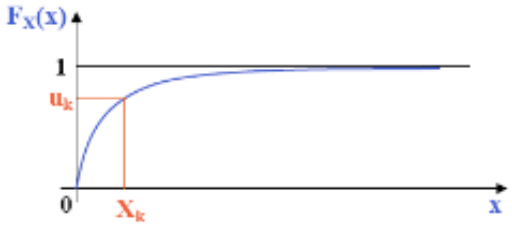
a partire da una sequenza di numeri random U(0,1) opportunamente generati, i metodi per la generazione di variabili aleatorie con distribuzione generica sono:

- Tecnica di trasformazione inversa
- Metodo di accettazione/rifiuto
- Metodo di composizione

6.2.4 Trasformazione inversa

si vuole generare una variabile aleatoria X con funzione di densità di probabilità $f_x(x)$:

1. Si calcola la funzione di distribuzione di probabilità o funzione cumulativa di probabilità
$$F_X(x) = \int_{-\infty}^{x} f_X(\tau)d\tau$$
Tale funzione (qualora sia possibile calcolarla in forma chiusa) è continua, monotona crescente ed è sempre compresa tra 0 e 1 per definizione di $F_X(x) = P(X \leq x)$
2. Si pone $u = F_X(x)$ con u numero random, $u \sim U(0,1)$
3. Si risolve $F = F_X^{-t}(u)$ e la variabile aleatoria X è distribuita secondo $f_X(x)$, $X \sim f_X(x)$



Per applicare quanto detto, vediamo com'è possibile ottenere una variabile aleatoria esponenziale partendo da una variabile aleatoria uniforme U: supponiamo di voler costruire una successione di numeri pseudocasuali come osservazioni della distribuzione esponenziale, ovvero con funzione di distribuzione

$$F(x) = 1 - e^{-\lambda x}$$

Ponendo $U = F(x) = 1 - e^{-\lambda x}$ si ricava $1 - U = e^{-\lambda x}$ da cui si ricava $\ln(1 - U) = \ln\left(e^{-\lambda x}\right) \Rightarrow \ln(1 - U) = -\lambda x \Rightarrow x = -\frac{\ln(1-U)}{\lambda}$, ovvero

$$F^{(-1)}(U) = -\frac{\ln(1 - U)}{\lambda}$$

Quindi se U è una variabile aleatoria uniformemente distribuita in [0,1], allora

$$x = F^{-1}(U) = -\frac{\ln(1 - U)}{\lambda}$$

È una variabile aleatoria con distribuzione esponenziale con media $\frac{1}{\lambda}$.

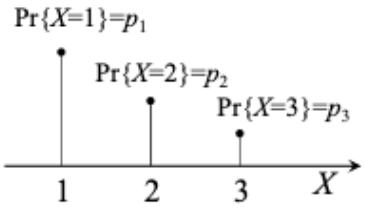
Se una variabile aleatoria U ha distribuzione uniforme in [0,1], allora anche 1-U ha distribuzione uniforme in [0,1] e quindi si può sostituire nell'argomento del logaritmo (1-U) con U. tuttavia, questo cambiamento potrebbe indurre un cambiamento nella correlazione delle variabili X generate.

Il metodo della trasformazione inversa può essere esteso ed utilizzato anche nel caso di distribuzioni discrete, ovvero quando si assume che la variabile X sia una variabile aleatoria discreta. Supponiamo, quindi, che X assuma valori $x_1, x_2, \ldots$ e

supponiamo che essi siano ordinati, ovvero $x_1 < x_2 <\ldots$. Data una variabile U uniformemente distribuita in [0,1] si definisce la variabile X nel seguente modo:

$$X(U) = \max\{x\_i \,|U \in [F(x\_i - 1) - F(x\_i)\}$$

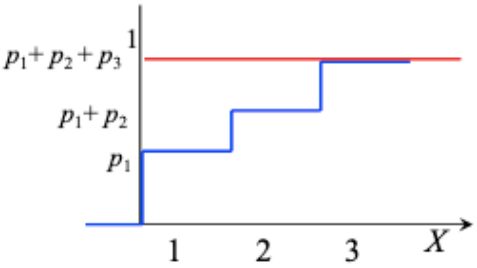
- Funzione di densità



$$\sum_{i=-\infty}^{\infty} \Pr\{X = i\} = 1$$

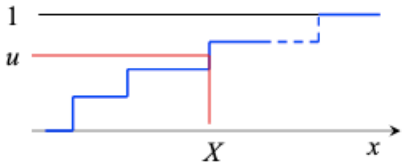
- Distribuzione Cumulativa

$$F_X(x) = \Pr\{X \le x\} = \sum_{i=-\infty}^{x} \Pr\{X = i\}$$



- In generale

$$\Pr\{X = x_j\} = p_j,\ j = 0,1,\ldots,N, \quad \sum_{j=0}^{N} p_j = 1$$



$$X = \begin{cases} x_0 & \text{if } u \le p_0 \\ x_1 & \text{if } p_0 \le u \le p_0 + p_1 \\ \vdots & \vdots \\ x_j & \text{if } \sum_{i=0}^{j-1} p_i \le u \le \sum_{i=0}^{j} p_i \\ \vdots & \vdots \end{cases}$$

6.2.5 Metodo acceptance-rejection

Il metodo della trasformazione inversa è basato sul calcolo della funzione inversa F^{-1} che non sempre può essere calcolata in modo o comunque non in modo efficiente. Per questo motivo sono stati sviluppati altri metodi, tra i quali il metodo detto acceptance-rejection.
Supponiamo di conoscere la densità di probabilità della variabile aleatoria X che intendiamo generare e che $f_X(x)$ sia definita su un intervallo finito [a, b] e la sua immagine sia definita sul codominio [0,c]. In pratica la funzione $f_X(x)$ è tutta contenuta all'interno del rettangolo [a,b]x[0,c].

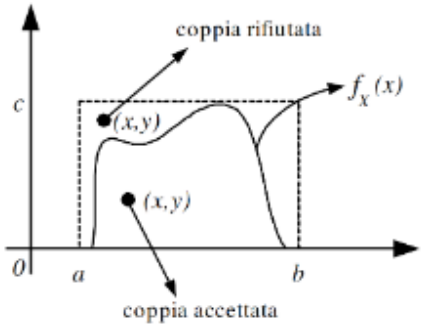
Generiamo due sequenze pseudo-casuali uniformi tra [0,1]; U_1, U_2.
Successivamente deriviamo altre due seguente numeriche uniformi secondo la seguente regola:

$$\begin{cases} X = a + (b-a)U_1 \\ Y = cU_2 \end{cases}$$

Ad ogni coppia di valori (U_1,U_2) corrisponderà una coppia (X, Y) appartenente al rettangolo [a,b]x[0,c].

Se la coppia (x,y) cade all'interno dell'area della funzione $\boldsymbol{f_X(x)}$, viene accettata e sarà successivamente utilizzata per creare la sequenza pseudo-casuale desiderata, altrimenti viene scartata. In questo ultimo caso la procedura viene ripetuta fino a trovare una nuova coppia contenuta nell'area di $f_X(x)$.
La sequenza di valori X così ottenuta è una sequenza pseudo-casuale che segue la legge di distribuzione $f_X(x)$ (infatti abbiamo scelto solo valori che cadono nella sua area).

Questo metodo è molto efficiente quando l'area di $f_X(x)$ copie quasi tutto il rettangolo [a,b]x[0,c]:

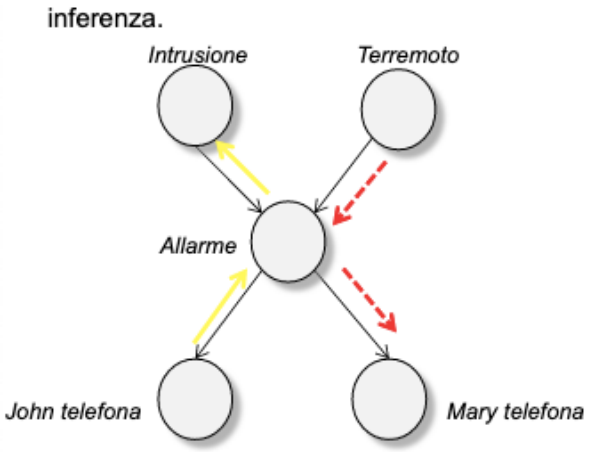


Procedimento:
si vuole generare una variabile aleatoria X con funzione di densità di probabilità $f_X(x)$ su un intervallo [a,b].

1. Si genera un'istanza di una variabile R distribuita in modo uniforme nell'intervallo [a,b] (U(a,b))
2. Si accetta tale valore con probabilità pari a $\frac{f_X(R)}{\max f_X(x)}$, lo si rifiuta con probabilità $1 - \frac{f_X(R)}{\max f_X(x)}$

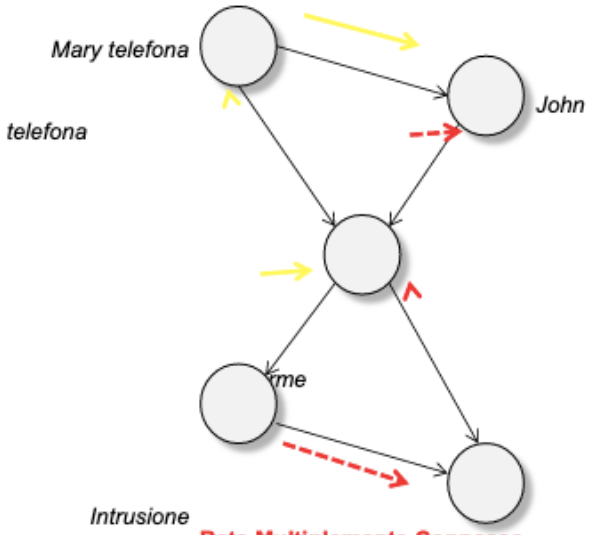
7. Inferenza approssimata

La struttura della rete è fondamentale per valutare la complessità di un algoritmo di inferenza. L'inferenza in reti Bayesiane con struttura generale è NP-hard.



Rete Singolarmente Connessa - Polytree

Al più un cammino non orientato tra ogni coppia di nodi. Complessità spazio e tempo di inferenza esatta lineare nella dimensione della rete (numero di elementi delle CPTs). Se il numero di genitori di ogni nodo è limitato da una costante, la complessità è lineare nel numero di nodi.



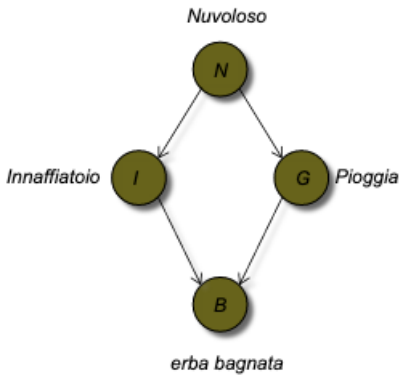
Rete Multiplamente Connessa

Più di un cammino non orientato per almeno una coppia di nodi. Complessità spazio e tempo di inferenza esatta nel caso peggiore è esponenziale anche sotto l'ipotesi che il numero di genitori sia limitato superiormente da una costante.

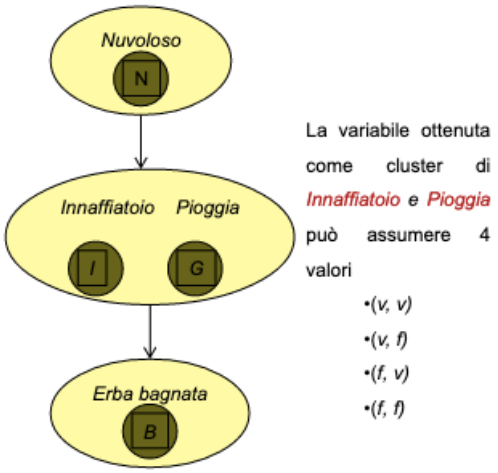
L'algoritmo di eliminazione variabili è efficiente per rispondere a query univariate.

In una rete polytree computare la probabilità a posteriori per tutte le variabili richiede di effettuare O(n) query che costano O(n), quindi in totale $O(n^2)$. l'impiego di un algoritmo appartenente alla classe che va sotto il nome di classe degli algoritmi di clustering consente si ridurre la complessità a O(n). Il principio sul quale si basa la classe degli algoritmi di clustering è rappresentato dal fatto di unire i nodi di una rete formando clusters di variabili in modo tale che la rete di clusters risultante sia un polytree al quale applicare un algoritmo di inferenza efficiente.

Il modello di rete Bayesiana a sinistra può essere convertito nel polytree riportato a destra:



Rete Multiplamente Connessa



Rete Singolarmente Connessa - Polytree

Data l'intrattabilità dell'inferenza esatta nel caso di reti multiplamente connesse (NP-hard) diviene essenziale prendere in considerazione algoritmi di inferenza approssimata. Nello specifico presenteremo di seguito algoritmo approssimati appartenenti alla categoria che va sotto il nome di Monte Carlo, la cui accuratezza dipende dal numero di campioni generati:

- Direct Sampling
- Markov Chain Sampling

7.1 Direct sampling

L'elemento base di ogni algoritmo di campionamento è la capacità di generare campioni da una specificata distribuzione di probabilità. Ad esempio, l'esito del lancio di una moneta bilanciata è una variabile casuale che può assumere due valori {testa, croce} con distribuzione a priori <0.5,0.5>. Campionare questa distribuzione equivale al lancio della moneta: il 50% dei lanci mostreranno esito testa, mentre il restante 50% mostreranno l'esito croce.

La situazione più semplice di campionamento di una rete Bayesiana è quella nella quale nessuna variabile della rete ha evidenza associata. L'idea è di campionare ogni variabile della rete seguendo l'ordine topologico. L'ordine topologico è stabilito a partire dalle variabili con meno genitori fino alle variabili con più genitori.

L'algoritmo campiona-priori genera campioni dalla distribuzione di probabilità congiunta a priori per il modello di rete Bayesiana specificata. Indichiamo tramite

$$S_{CP}(x_1, \dots, x_n)$$

La probabilità che uno specifico evento sia generato tramite l'algoritmo campiona-priori. Dalla natura del procedimento di campionamento abbiamo che

$$S_{CP}(x_1, \dots, x_n) = \Pi_{i=1}^{n} P\left(x_i \middle| parents(X_i)\right)$$

In quanto ogni campione dipende solo dal valore assunto dai nodi genitore.
Questa espressione ci è familiare in quanto rappresenta la probabilità dell'evento considerato secondo la distribuzione di probabilità congiunta descritta tramite il modello di rete Bayesiana (fattorizzazione). Pertanto possiamo scrivere

$$S_{CP}(x_1, \dots, x_n) = P(x_1, \dots, x_n)$$

In ogni algoritmo di campionamento le risposte vengono computate tramite conteggio dei campioni generati. Supponiamo vengano generati N campioni e sia

$$N_{CP}(x_1, \dots, x_n)$$

La frequenza riscontrata per l'evento $(x_1, \dots, x_n)$ negli N campioni.
Ci si attende che la frequenza converga al suo valore atteso in accordo alla distribuzione di probabilità dalla quale vengono estratti gli N campioni:

$$\lim_{N\to\infty} \frac{N_{CP}(x_1, \dots, x_n)}{N} = S_{CP}(x_1, \dots, x_n) = P(x_1, \dots, x_n)$$

Verrà usato il simbolo ≈ per indicare che la probabilità stimata diviene esatta nel limite in cui N tende ad infinito. Una tale stima viene detta consistente.
Ad esempio è possibile produrre una stima consistente della probabilità di ogni evento parzialmente specificato $x_1, \dots, x_m,\ m \leq n$ come segue

$$P(x_1, \dots, x_m) \approx \frac{N_{CP}(x_1, \dots, x_m)}{N}$$

La probabilità di un evento parzialmente specificato viene stimata come la frazione dei casi compatibili con l'evento parzialmente specificato sul numero di tutti i casi generati tramite campionamento.

L'algoritmo campiona-priori consente di ottenere una stima della probabilità di ogni evento nel caso in cui si consideri un modello di rete Bayesiana dove nessun nodo abbia evidenza.
Nel caso in cui si sia interessati a computare una probabilità condizionata del tipo

$$P(X|E = e)$$

È possibile utilizzare l'algoritmo di campionamento con rigetto (rejection sampling) che utilizza l'algoritmo campiona-priori per generare campioni dalla distribuzione di probabilità congiunta rappresentata dalla rete Bayesiana e successivamente rifiuta tutti quei campioni generati che non sono conformi dal punto di vista dell'evidenza e.
La stima della probabilità a posteriori viene ottenuta contando la frequenza per X=x nell'insieme dei campioni non rigettati. Indichiamo con

$$\hat{P}(X|E = e)$$

La stima della distribuzione di probabilità a posteriori computata dall'algoritmo di campionamento con rigetto. In base all'algoritmo avremo

$$\hat{P}(X|E=e) = \alpha N_{CP}(X, E=e) = \frac{N_{CP}(X, E=e)}{N_{CP}(E=e)}$$

In base alla relazione

$$P(x_1, \dots, x_n) = \frac{N_{CP}(x_1, \dots, x_n)}{N}$$

La relazione precedente diventa

$$\hat{P}(X|E=e) = \frac{P(X, E=e)}{P(E=e)} = P(X|E=e)$$

Quindi l'algoritmo di campionamento con rigetto produce stime consistenti della vera distribuzione di probabilità. La stima converge al valore corretto man mano che il numero di campioni non rigettati aumenta; la deviazione standard dell'errore è proporzionale a $\frac{1}{\sqrt{N_{valid}}}$, dove N_{valid} è il numero di campioni non rigettati.
Purtroppo l'algoritmo di campionamento con rigetto rifiuta troppo campioni ed il tasso di rigetto cresce esponenzialmente con il numero di variabili per le quali è disponibile evidenza. Questa particolarità rende inutilizzabile l'algoritmo in questione nel caso di modelli reali di reti Bayesiane.

Un'alternativa all'algoritmo di campionamento con rigetto è offerta dall'algoritmo di likelihood weighting che evita di generare campioni che dovranno essere successivamente rigettati. L'algoritmo di likelihood weighting fissa il valore dei nodi (variabili) per i quali è disponibile evidenza E=e, in accordo con l'evidenza medesima, e successivamente genera campioni solo per i nodi restanti X e Y. L'algoritmo di likelihood weighting inoltre pesa in modo differente i vari eventi e il peso di ogni evento è il likelihood che l'evento associa all'evidenza.

Intuitivamente gli eventi per i quali l'evidenza disponibile è inverosimile devono pesare in misura minore nel processo di stima delle probabilità a posteriori. Senza fornire ulteriori dettagli possiamo dire quanto segue:

- Il meccanismo di likelihood weighting utilizza tutti i campioni generati (più efficiente del campionamento con rigetto)
- All'aumentare del numero di nodi con evidenza la performance degrada in quanto molti dei campioni estratti avranno un peso infinitesimale; la stima sarà interamente dipendente da pochissimi campioni con peso comunque piccolo

7.2 Markov Chain Sampling

Differentemente dagli algoritmi di campionamento con rigetto e likelihood weighting che generano ogni evento partendo da zero, l'algoritmo Markov Chain

Monte Carlo (MCMC) genera ogni evento tramite modifiche casuali di un evento che lo precede nell'esecuzione. È utile pensare che il modello di rete Bayesiana si trovi in un determinato stato corrente; tale stato è identificato tramite l'assegnamento di un valore ad ogni nodo della rete. Lo stato prossimo viene ottenuto tramite campionamento di una delle variabili senza evidenza X_i condizionalmente ai valori correntemente assunti dalle variabili che costituiscono la Markov Blanket della variabile (nodo) X_i.

È importante mostrare come sia possibile estrarre un campione proveniente dalla distribuzione di probabilità di una variabile casuale X_i quanto sia noto lo stato delle variabili casuali che ne costituiscono la Markov Blanket di X_i.

Il campionamento in oggetto si basa sul fatto che

$$P(x_i|MB(X_i)) = \alpha P(x_i|parents(X_i)) \cdot \Pi_{y \in child(X_i)} P(y_i|parents(Y_i))$$

Ovvero la probabilità di una variabile casuale data la sua Markov Blanket è proporzionale alla probabilità della variabile dati i suoi genitori moltiplicata per la probabilità di ogni suo figlio dati i rispettivi genitori.

8. Processi stocastici

Il compito di prendere una decisione dipende da:
- Informazioni parziali
- Informazioni rumorose
- Incertezza sui cambiamenti dell'ambiente nel corso del tempo

Per descrivere un mondo mutevole si usano:
- Una serie di variabili casuali
- Descritte da uno stato
- In ogni istante di tempo

Le relazioni tra variabili casuali in istanti temporali diversi descrivono l'evoluzione dello stato.

Nei modelli statici il valore delle variabili non cambia nel tempo, mentre in modelli dinamici il valore delle variabili cambia nel tempo, lo stato corrente dipende dalla storia e il processo di cambiamento è descritto da una serie di "fotografie" ognuna delle quali contiene un insieme di variabili casuali.

8.1 Processo stocastico

un processo stocastico $\{X(t),\ t \in T\}$ è:
- Un insieme di variabili casuali (per ogni t, X(t) è una variabile casuale)
- Una variabile casuale che evolve nel tempo

L'insieme T degli indici e lo spazio X degli stati possono essere continui o discreti.

- Processi stocastici a tempo continuo: {X(t), t>0}
- Processi stocastici a tempo discreto: {X(t), t=0,1,…}
- Processi stocastici a stati continui
- Processi stocastici a stati discreti

X(t) è il valore dello stato del sistema al tempo t, ovvero il valore di una variabile casuale che descrive lo stato del sistema al tempo t.

Una proprietà dei processi stocastici è la proprietà markoviana. Tale proprietà assicura che la distribuzione di probabilità per tutti i possibili valori futuri del processo dipende solo dal loro valore corrente e non dai valori passati:

$$P(X_{t+1} = i_{t+1} | X_t = i_t, X_{t-1} = i_{t-1}, \dots, X_1 = i_1, X_0 = i_0) = P(X_{t+1} = i_{t+1} | X_t = i_t)$$

Dove t è lo stato corrente. I processi stocastici che soddisfano questa proprietà sono detti processi di Markov.

8.2 Catene di Markov

Un processo stocastico a tempi discreti è una catena di Markov se per t=1,2,3,… e per tutti gli stati si ha che

$$P(X_{t+1} = j | x_t = i, X_{t-1} = i_{t-1}, \dots, X_1 = i_1, X_0 = i_0) = P(x_{t+1} = j | X_t = i)$$

$P(X_0 = i) = q_i$, dove $q = [q_1, \dots, q_i, \dots, q_n]$ è la distribuzione di probabilità iniziale.

Se la probabilità di un certo evento è indipendente dal tempo t la catena di Markov si definisce stazionaria e si ha che

$$P(X_{t+1} = j | x_t = i) = p_{ij}$$

Dove p_{ij} è la probabilità che al tempo t+1 il sistema sarà nello stato j, essendo nello stato i al tempo t, ovvero è la probabilità di raggiungere uno stato j partendo da uno stato i della catena.

$$p_{ij} \geq 0, \qquad i, j \geq 0, \qquad \sum_{j=0}^{n} p_{ij} = 1$$

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \dots & \dots & \dots & \dots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{bmatrix} = \text{matrice di transizione}$$

Una matrice P delle probabilità di transizione è rappresentabile graficamente con un grafo. Ogni nodo rappresenta uno stato e l'arco (i,j) rappresenta la probabilità di transizione p_{ij}:

Ingredienti di una catena di Markov:

- Stati $\{S_1, \dots, S_N\}$
- Probabilità di transizione tra stati $p_{ij} = P\left(X_{t+1} = S_i \middle| X_t = S_j\right)$
- Distribuzione iniziale degli stati $\pi_i = P[X_1 = S_i]$

8.2.1 Probabilità di transizione a n-passi

se una catena di Markov è in uno stato i al tempo m, qual è la probabilità che dopo n passi sarà in uno stato j?

$$P(X_{m+n} = j | X_m = i) = P(X_n = j | X_0 = i) = P_{ij}(n)$$

Si avrà che

$$P_{ij}(2) = \sum_{k=1}^{n} p_{ik} \cdot p_{kj} \quad \text{prodotto scalare riga i colonna j}$$

$$P_{ij}(n) = \text{ij} - \text{esimo elemento di } P^n$$

8.2.2 Equazioni Chapman-Kolmogorov

la probabilità di transizione a n-passi

$$P_{ij}^n = P\{X_{n+k} = j | X_k = i\}, \qquad n \geq 0, \qquad i, j \geq 0$$

Può essere calcolata tramite le equazioni di Chapman-Kolmogorov:

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n \cdot P_{kj}^m, \qquad \forall n, m \geq 0, \qquad i, j \geq 0$$

$$P(n+m) = P(n)P(m)$$

$$P_{ij}^{n+m} = P\{X_{n+m} = j | X_0 = i\}$$
$$= \sum_{k=0}^{\infty} P\{X_{n+m} = j, X_n = k | X_0 = i\}$$
$$= \sum_{k=0}^{\infty} P\{X_{n+m} = j | X_n = k, X_0 = i\} P\{X_n = k | X_0 = i\} = \sum_{k=0}^{\infty} P_{kj}^{m} \cdot P_{ik}^{n}$$

8.2.3 Probabilità di transizione

la probabilità di essere in un certo stato j al tempo n non conoscendo lo stato di una catena di Markov al tempo 0 è

$$\sum_i q_i \cdot P_{ij}(n) = q \cdot (\text{colonna } j \text{ di } P^n)$$

Dove q_i è la probabilità che la catena sia nello stato i al tempo 0.

8.2.4 Classificazione degli stati

- Uno stato j è raggiungibile da uno stato i se esiste un cammino che da i arriva a j

$$P_{ij}^{n} > 0, \qquad \text{per qualche } n \geq 0$$

- Due stati i e j si dice che comunicano se j è raggiungibile da i, e viceversa. Ogni stato comunica con se stesso per definizione e vale anche la proprietà transitiva.
- Una catena di Markov è detta irriducibile se tutti i suoi stati sono comunicanti tra loro.
- Un insieme di stati S in una catena di Markov è un insieme chiuso se nessuno stato fuori da S è raggiungibile dagli stati in S.
- Uno stato i si definisce stato assorbente se

$$p_{ii} = 1$$

- Uno stato i si definisce stato transiente se esiste uno stato j raggiungibile da i, ma i non è raggiungibile da j

$$\sum_{n=1}^{\infty} P_{ii}^{n} < \infty$$

- Uno stato che non è transiente viene definito stato ricorrente

$$\sum_{n=1}^{\infty} P_{ii}^{n} = \infty$$

- La ricorrenza è una proprietà di classe: se lo stato i è ricorrente e lo stato j comunica con i, allora lo stato j è ricorrente.
- Anche essere transiente è una proprietà di classe.
- Tutti gli stati di una catena di Markov finita (numero di stati finito) e irriducibile sono ricorrenti.
- Uno stato i è periodico di periodo k>1 se k è il più piccolo numero tale che tutti i cammini che dallo stato i ritornano ad i hanno una lunghezza che è un multiplo di k.
- Se uno stato non è periodico si definisce aperiodico.
- Se tutti gli stati in una catena sono ricorrenti, aperiodici e comunicano l'uno con l'altro, la catena si definisce ergodica.
- Una catena è periodica se produce ciclicamente la stessa sequenza di stati.

8.2.5 Distribuzione d'equilibrio (steady state)

sia P una matrice delle probabilità per una catena ergodica di N stati, allora vale che

$$\lim_{t\to\infty} P_{ij}(t) = \pi_j$$

Con $\pi = [\pi_1, \pi_2, \dots, \pi_n]$ vettore distribuzione d'equilibrio dove

$$\pi = \pi \cdot P$$

8.2.6 Transitorio

Il comportamento di una catena di Markov prima di raggiungere la distribuzione d'equilibrio è chiamato transitorio.

8.2.7 Passaggio intermedio

Numero di transizioni attese prima di raggiungere lo stato j essendo nello stato i in una catena ergodica:

$$m_{ij} = p_{ij}(1) + \sum_{k\neq j} p_{ik} \cdot (1 + m_{kj}) = 1 + \sum_{k\neq j} p_{ik} \cdot m_{kj}$$

$$m_{ii} = \frac{1}{\pi_i}$$

Numero di passi da k a j
Probabilità da i a k
Probabilità da i a j in 1 passo
Probabilità di trovarsi nello stato i

Catene assorbenti:

le catene assorbenti sono catene di Markov nelle quali alcuni stati sono assorbenti, mentre tutti gli altri sono stati transienti.

Remember: uno stato i si definisce stato assorbente se $p_{ii} = 1$.

Possibili domande:
1. Qual è il numero di passi che intercorrono prima che, da uno stato transiente, venga raggiunto uno stato assorbente?
2. Se una catena parte da uno stato transiente, qual è la probabilità che termini in uno stato assorbente?

8.2.8 Matrice di transizione

La matrice di transizione per una catena assorbente può essere descritta come

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

- Q matrice che rappresenta le relazioni tra gli stati transienti
- R matrice che rappresenta le transizioni da stati transienti a stati assorbenti

8.2.9 Matrice fondamentale

Se siamo in uno stato transiente i, il numero di periodi che si trascorreranno in uno stato transiente j prima dell'assorbimento nello stato è

$$\text{ij} - \text{esimo elemento della matrice } (I - Q)^{-1}$$

8.2.10 Probabilità d'assorbimento

Se siamo in uno stato transiente i, la probabilità di arrivare in uno stato assorbente j è:

$$\text{ij} - \text{esimo elemento della matrice } (I - Q)^{-1} \cdot R$$

9. Google's Search Engine

Assunzione: un link da una pagina A ad una pagina B è una "raccomandazione" della pagina B da parte di un autore della pagina A.
La qualità di una pagina è misurata in base al numero dei suoi links entranti.

Teorema (Markov Chain):
- Esiste un'unica distribuzione stazionaria q con $q_i > 0 \quad \forall i$
- Sia N(i,t) il numero di volte in cui la Markov Chain visita lo stato i in t passi, allora

$$\lim_{t\to\infty} \frac{N(i,t)}{t} = \pi_i$$

Numero di volte che la catena di Markov visita lo stato i in t steps

$$pageRank(u) = \frac{p}{n} + (1-p) \cdot \sum_{(u,v)\in E} \frac{pageRank(v)}{outDegree(v)}$$

Dove n è il numero totale di nodi del grafo e p è la probabilità di saltare in modo random.

10. Ragionamento probabilistico nel tempo

Il compito di prendere una decisione dipende da:
- Informazioni parziali
- Informazioni rumorose
- Incertezza sui cambiamenti dell'ambiente nel corso del tempo

10.1 Tempo e incertezza

- Modello statici:
  - Il valore delle variabili di stato non cambia nel tempo
- Modelli dinamici:
  - Il valore delle variabili di stato cambia nel tempo
  - Lo stato corrente dipende dalla storia
  - Il processo di cambiamento è descritto da una serie di "fotografie" ognuna delle quali contiene un insieme di variabili casuali
  - Per descrivere un mondo mutevole si usano una serie di variabili casuali descritte da uno stato in ogni istante temporale
  - Le relazioni fra variabili casuali in istanti temporali diversi descrivono l'evoluzione dello stato

Principi di modellazione:
- Set X_t di variabili di stato non osservabili al tempo t
- Set E_t di variabili osservabili al tempo t
- Le dipendenze tra di esse
- L'ipotesi che i cambiamenti del mondo siano regolati da un processo stazionario, ovvero la probabilità di passare da uno stato i ad uno stato j non varia nel tempo
- Ipotesi di Markov (del primo ordine), ovvero tutta la storia è riassunta nello stato corrente

- Il teorema di esistenza e unicità afferma che data una catena di Markov omogenea a stati discreti con probabilità di transizione P_{ij} e spazio degli stati S, se la catena di Markov è irriducibile e ricorrente positiva allora esiste un'unica distribuzione stazionaria π per la catena di Markov.
- Il teorema della convergenza afferma che data una catena di Markov omogenea a stati discreti con probabilità di transizione P_{ij} e spazio degli stati S, se la catena di Markov è irriducibile, aperiodica e ricorrente positiva allora la distribuzione di probabilità $\tilde{\pi}_n$ al tempo t_n, converge alla distribuzione stazionaria π per ogni distribuzione iniziale di probabilità $\tilde{\pi}_0$ scelta. Si ha cioè $\forall \tilde{\pi}_0, \forall i, j \in S, \lim_{n\to\infty} \sum_{i\in S} (\tilde{\pi}_0)_i (P^{(n)})_{ij} = \pi_j$.

11. Catene di Markov a stati nascosti: Hidden Markov Model – HMM

Una catena di Markov a stati nascosti è un processo stocastico continuo/discreto caratterizzato da:

- Un insieme di stati X_t
- Un insieme di osservazioni E_t
- Una matrice delle probabilità di transizione

$$P(X|X_{0:t-1}) = P(X|X_{t-1})$$

- Una matrice delle probabilità di emissione delle osservazioni

$$P(E_t|X_{0:t-1}, E_{0:t-1}) = P(E_t|X_t)$$

- Matrice delle probabilità iniziali degli stati al tempo 0: $P(X_0)$

Catena di Markov a stati nascosti:

Modello di transizione: $P(X|X_{0:t-1}) = P(X|X_{t-1})$

Modello sensoriale: $P(E_t|X_{0:t-1}, E_{0:t-1}) = P(E_t|X_t)$

Per ogni t finito, la distribuzione congiunta risulta:

$$P(X_0, \dots, X_t, E_1, \dots, E_t) = P(X_0) \cdot \Pi_{i=1}^{t} P(X_i|X_{i-1}) P(E_i|X_i)$$

Esempio di HMM: un detenuto in carcere vuole sapere se fuori piove, nevica o c'è il sole. Lui non può guardare fuori dalla finestra e l'unica cosa che può guardare è il comportamento/vestiario del direttore del carcere. Lui può osservare che, nei vari giorni con una certa probabilità, possono vestirsi leggeri, pesanti, col cappotto, possono portare l'ombrello, etc… e da qui dedurre se fuori piove, c'è il sole oppure nevica.

L'ipotesi di Markov in alcuni casi può essere un'ipotesi forte, infatti suppone che le variabili di stato contengano tutte le informazioni necessarie per caratterizzare la distribuzione di probabilità dell'istante successivo, ma potrebbe non essere sempre così. Due possibili rimedi possono essere:

1. Aumentare l'ordine del modello di processo di Markov
2. Aumentare l'insieme delle variabili di stato

In questo modo, però, aumenta inevitabilmente la complessità del modello.

Task di inferenza nei modelli temporali (descriviamoli con degli esempi):

- Filtraggio: qual è la probabilità che oggi piova avendo osservato il direttore (ombrello) tutti i giorni fino ad oggi?
- Previsione: qual è la probabilità di pioggia tra n giorni avendo osservato il direttore (ombrello) tutti i giorni fino ad oggi?
- Smoothing: qual è la probabilità che ieri abbia piovuto avendo osservato il direttore (ombrello) fino ad oggi?
- Spiegazione più probabile: se il direttore è apparso con l'ombrello i primi tre giorni ma non il quarto, qual è la sequenza di stati (pioggia/non pioggia) più probabile che può aver indotto il direttore a tale comportamento?
- Apprendimento del modello di transizione e del modello sensoriale a partire dalle osservazioni. L'apprendimento si ottiene come sottoprodotto dell'inferenza: questa fornisce una stima che può essere utilizzata per aggiornare il modello. L'apprendimento richiede di effettuare uno smoothing completo

11.1 Filtraggio

Il filtraggio è una stima ricorsiva, ovvero prevede il calcolo della distribuzione a posteriori dello stato corrente date tutte le osservazioni

$$P(X_t|e_{1:t})$$

Date le distribuzioni fino al tempo t-1, si può calcolare la distribuzione al tempo t della nuova prova e_t:

$$P(X_t|e_{1:t}) = f\big(e_t, P(X_{t-1}|e_{1:t-1})\big)$$

Nuova osservazione
Filtraggio al tempo t-1

Il filtraggio è suddiviso in due step: proiezione e aggiorname

- Dato il risultato del filtraggio al tempo t, si proietta in avanti la distribuzione dello stato corrente da t a t+1
- Si aggiorna in base alla nuova prova e_{t+1}:

$$\begin{aligned} P(X_{t+1}|e_{1:t+1}) &= P(X_{t+1}|e_{1:t}, e_{t+1}) \\ &= \alpha P(e_{t+1}|X_{t+1}, e_{1:t}) P(X_{t+1}|e_{1:t}) \\ &= \alpha P(e_{t+1}|X_{t+1}) P(X_{t+1}|e_{1:t}) \end{aligned}$$

Divido l'evidenza
Th di Bayes
Propr Markov dell'evidenza

Predizione del prossimo step
Aggiornamento in base all'evidenza

Dove α è la costante di normalizzazione della somma di probabilità.

Il primo termine aggiorna lo stato rispetto alla nuova evidenza e il secondo termine rappresenta la predizione del prossimo stato.
Condizioniamo allo stato corrente X_t per ottenere la predizione del prossimo stato:

$$P(X_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|X_{t+1}) \cdot \sum_{x_t} P(X_{t+1}|x_t, e_{1:t}) P(x_t|e_{1:t})$$

Modello di transizione
Modello sensoriale

$$= \alpha P(e_{t+1}|X_{t+1}) \sum_{x_t} P(X_{t+1}|x_t) P(x_t|e_{1:t})$$

Propr Markov dell'evidenza

$$f_{1:t+1} = \alpha \cdot \text{forward}(f_{1:t}; e_{t+1})$$

11.2 Predizione

la predizione è un <u>filtraggio privo dell'aggiunta di nuove osservazioni</u>

$$P(X_{t-k}|e_{1:t})$$

Cosa succede quando cerchiamo di predire sempre più avanti nel tempo?
Maggiore è l'incertezza del modello di transizione, più breve sarà il tempo per raggiungere un punto fisso (stato stazionario) per una predizione e più ignoto sarà il futuro.

Forward
Ricorsione

$$P(X_t) = \sum_{x_{t-1}} P(X_t, x_{t-1}) = \sum_{x_{t-1}} P(X_t|x_{t-1}) P(x_{t-1})$$

11.3 Calcolo della verosimiglianza

possiamo usare una ricorsione per il calcolo della verosimiglianza di una sequenza di prove:

$$P(e_{1:t})$$

La verosimiglianza è utile per confrontare modelli diversi che potrebbero aver prodotto la stessa sequenza di prove:

$$l_{1:t} = P(X_t, e_{1:t})$$
$$l_{1:t+1} = \text{forward}(l_{1:t}; e_{t+1})$$
$$L_{1:t} = P(e_{1:t}) = \sum_{x_t} l_{1:t}(x_t)$$

11.4 Smoothing (regolarizzazione)

Lo smoothing è il processo della distribuzione di stati passati, date le osservazioni fino allo stato corrente

$$P(X_k|e_{1:t}), \qquad \text{con } 1 \leq k < t$$

Consideriamo separatamente le osservazioni fino a k e quelle da k+1 a t, con t>k:

Regola di Bayes
Indipendenza condizionata
Forward
Backward

$$\begin{aligned} P(X_k|e_{1:t}) &= P(X_k|e_{1:k}, e_{k+1:t}) \\ &= \alpha P(X_k|e_{1:t}) P(e_{k+1:t}|X_k, e_{1:t}) \\ &= \alpha P(X_k|e_{1:t}) P(e_{k+1:t}|X_k) \\ &= \alpha f_{(1:k)}\, b_{k+1:t} \end{aligned}$$

Dove con forward si intende il filtraggio in avanti da 1 a k.

$b_{k+1:t}$ viene calcolato mediante una procedura ricorsiva di backward che procede all'indietro da t:

$$\begin{aligned} P(e_{k+1:t}|X_k) &= \sum_{x_{k+1}} P(e_{k+1:t}|X_k, x_{k+1}) P(x_{k+1}|X_k) \\ &= \sum_{x_{k+1}} P(e_{k+1:t}|x_{t+1}) P(x_{k+1}|X_k) \\ &= \sum_{x_{k+1}} P(e_{k+1}, e_{k+2:t}|x_{k+1}) P(x_{k+1}|X_k) \end{aligned}$$

Modello
Chiamata ricorsiva

$$= \sum_{x_{k+1}} P(e_{k+1}|x_{k+1}) P(e_{k+2:t}|x_{k+1}) P(x_{k+1}|X_k)$$

11.5 Sequenza più probabile

Data una sequenza di osservazioni vogliamo trovare la sequenza di stati che più probabilmente ha generato il mio set di osservazioni

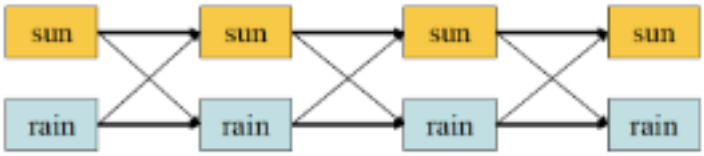
$$\arg\max_{x(1:t)} P(x_{1:t}|e_{1:t})$$

Supponiamo che ogni stato possa assumere solo valori binari e che la mia sequenza di osservazioni sia lunga n, quindi potremmo avere 2^n possibili sequenze di stati.

Consideriamo ogni sequenza come un cammino lungo un grafo: la probabilità di ogni cammino è il prodotto delle probabilità di transizione per le probabilità delle osservazioni rilevate ad ogni stato.

L'algoritmo di Viterbi si basa sulla seguente assunzione: esiste una relazione ricorsiva fra i cammini più probabili verso ogni stato x_{t+1} e i cammini più probabili verso ogni stato x_t.

Esempio:



Tale relazione ricorsiva può essere scritta come:

$$\text{Forma ricorsiva} \quad \max_{x_1,\dots,x_t} P(x_1,\dots,x_t,X_{t+1}|e_{1:t+1})$$
$$= \alpha P(e_{t+1}) \cdot \max_{x_t}\left(P(X_{t+1}|x_t) \cdot \max_{x_1,\dots,x_{t-1}} P(x_1,\dots,x_{t-1},x_t|e_{1:t})\right)$$

Al termine sarà disponibile la probabilità della sequenza più probabile che raggiunge ogni stato finale.

Siccome il procedimento diventa troppo costoso, si fanno degli aggiornamenti incrementali:

$$\begin{aligned} m_t[x] &= \max_{x_{1:t-1}} P(x_{1:t-1},x) \\ &= \max_{x_{1:t-1}} P(x_{1:t-1})P(x|x_{t-1}) \\ &= \max_{x_{t-1}} P(x_t|x_{t-1}) \cdot \max_{x_{1:t-2}} P(x_{1:t-1}) \\ &= \max_{x_{t-1}} P(x_t|x_{t-1}) \cdot m_{t-1}[x] \quad \text{Forma ricorsiva} \end{aligned}$$

$$m_1[x] = P(x_1)$$

Qual è la sequenza più probabile di stati date le osservazioni?

$$\begin{aligned} x^*_{1:T} &= \arg\max_{x_{1:T}} P(x_{1:T}|e_{1:T}) \\ m_t[x_t] &= \max_{x_{1:t-1}} P(x_{1:t-1},x_t,e_{1:t}) \\ &= \max_{x_{1:t-1}} P(x_{1:t-1},e_{1:t-1})P(x_t|x_{t-1})P(e_t|x_t) \\ &= P(e_t|x_t) \cdot \max_{x_{1:t-1}} P(x_t|x_{t-1}) \cdot \max_{x_{1:t-2}} P(x_{1:t-1},e_{1:t-1}) \\ &= P(e_t|x_t) \cdot \max_{x_{t-1}} P(x_t|x_{t-1}) \cdot m_{t-1}[x_{t-1}] \end{aligned}$$

$$m_1[x] = P(x_1)$$

Reti Bayesiane dinamiche:

- Modelli di Markov nascosti, ovvero sono modelli probabilistici temporali nei quali gli stati del processo sono descritti da una singola variabile discreta

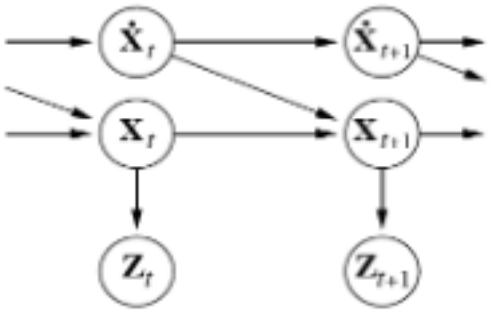
- Modello basato sul filtro di Kalman, ovvero stima lo stato di un sistema fisico a partire da una sequenza di dati rumorosi (conosciuti anche come sistemi dinamici lineari)

12. Filtri di Kalman

Obiettivo: stimare lo stato (es: posizione, velocità) di un sistema dinamico partendo da una sequenza di osservazioni rumorose.

Sono necessari:
- Modello di transizione, che descrive la fisica di moto
- Modello sensoriale, che descrive il processo di misurazione

Il sistema è descritto da un insieme di variabili continue (es: posizione (X,Y,Z) e velocità (X,Y,Z)).



Supponendo che l'intervallo di tempo tra due osservazioni sia Δ e he la velocità tra due istanti temporali sia costante, allora l'aggiornamento della posizione avviene tramite la formula

$$X_{t+\Delta} = X_t + \dot{X}\Delta$$

Aggiungendo del rumore gaussiano si ottiene un modello di transizione gaussiano lineare

$$P(X_{t+\Delta} = x_{t+\Delta} \,|\, X_t = x_t, \dot{X}_t = \dot{x}_t) = N(x_t + \dot{x}_t\Delta, \sigma^2)(x_{t+\Delta})$$

La gaussiana è unimodale e ha un solo massimo; la probabilità a posteriori si focalizza attorno al vero stato con "poca" incertezza.

Idea base: belief rappresentate da distribuzioni normali multivariate.

- Se la distribuzione corrente $P(X_t|e_{1:t})$ è gaussiana e il modello di transizione $P(X_{t+1}|x_t)$ è lineare, allora la predizione ad un passo sarà anch'essa gaussiana:

$$P(X_{t+1}|e_{1:t}) = \int_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})dx_t$$

- Se la distribuzione corrente $P(X_t|e_{1:t})$ è gaussiana e il modello sensoriale $P(e_{t+1}|X_{t+1})$ è gaussiano lineare, allora dopo aver aggiornato il modello rispetto alla nuova evidenza anche la distribuzione aggiornata è gaussiana lineare

$$P(X_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t})$$

L'operatore forward per il filtro di Kalman accetta un messaggio in avanti gaussiano $f_{1:t}$ specificato da una media μ e da una matrice di covarianza Σ_{t+1}. Quindi se iniziamo dalla distribuzione iniziale $P(X_0) = N(\mu_0, \Sigma_0)$ filtrando con un modello gaussiano lineare otterremo ancora una distribuzione statale gaussiana. Questo sembra essere un risultato piacevole ed elegante, ma perché è così importante? Il motivo è che ci permette di descrivere la distribuzione a posteriori con una distribuzione gaussiana utilizzando i due parametri media e varianza. Infatti, se la distribuzione non fosse gaussiana, il filtraggio nel continuo o ibrido (discreto e continuo) genera distribuzioni dello stato la cui rappresentazione cresce senza limiti col tempo (assume forme sempre più complesse).

12.1 Filtro di Bayes

L'algoritmo del filtro di Bayes calcola la probabilità a posteriori dello stato X_t condizionato alle misure e ai controlli fino al tempo t.
Sono richieste tre distribuzioni di probabilità:

1. La belief iniziale $P(X_0)$; KF assume la distribuzione dello stato iniziale gaussiana
2. La probabilità della misura $P(Z_t|X_t)$; KF assume la distribuzione di misura gaussiana lineare
3. La probabilità di transizione $P(X_t|X_{t-1})$; KF assume il sistema dinamico lineare