Modelli Probabilistici per le Decisioni

UniShare

Davide Cozzi
@dlcgold

Indice

Capitolo 1

Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: `https://github.com/dlcgold/Appunti`.

Capitolo 2

Ripasso di Probabilità

Riprendiamo qualche definizione.

Definizione 1. *Definiamo **variabile casuale** come un'osservazione, un esito o un evento il cui valore è incerto.*

Definizione 2. *Definiamo **dominio o spazio degli eventi** come l'insieme dei possibili valore che può assumere una variabile casuale.*

Definizione 3. *Definiamo **spazio di probabilità o modello di probabilità** come uno spazio degli eventi corredato da un assegnamento:*

$$P(\omega), \omega \in \Omega$$

tale che:

- $0 \leq P(\omega) \leq 1$
- $\sum_{\omega} p(\omega) = 1$

con omega evento e Ω spazio degli eventi.

Definizione 4. *Definiamo **evento atomico o campione** una specificazione completa del calore delle variabili casuali di interesse.*
L'insieme di tutti i possibili eventi atomici è:

- *mutualmente esaustivo (non potendo accadere altro)*
- *mutualmente esclusivo (può accadere solo un evento atomico di quelli possibili)*

Definizione 5. *Definiamo un **evento** (non atomico) A può essere un qualunque sottoinsieme di Ω tale che:*

$$P(A) = \sum_{\omega \in A} P(\omega)$$

Definizione 6. *Definiamo una* ***variabile aleatoria*** *è una variabile che può assumere diversi valori in corrispondenza di altrettanti eventi che costituiscono una partizione dello spazio delle probabilità.*

Si ricorda che, per una variabile a e una b:

- $0 \leq P(a) \leq 1$
- $P(\top) = 1$ e $P(\bot) = 0$
- $P(a \vee b) = P(a) + p(b) - p(a \wedge b)$

Definizione 7. *Definiamo una* ***probabilità condizionata*** *rappresenta la verosimiglianza che un evento a si verifichi se b si verifica e si denota con:*

$$P(a|b)$$

Si ha quindi la specifica che alcuni eventi rendono altri eventi più o meno verosimili.
Si parla quindi di eventi ***dipendenti****.*

Definizione 8. *Due eventi sono* ***indipendenti*** *se un evento non influisce sulla realizzazione dell'altro:*

$$P(a|b) = P(a)$$

Si ha quindi la seguente regola.

Teorema 1 (Regola del prodotto)**.** *Possiamo calcolare che due eventi si verifichino contemporaneamente tramite la probabilità condizionata e quella dei singoli eventi:*

$$P(a, b) = P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

Con $P(a, b) = P(a \wedge b)$ è detta ***probabilità congiunta*** *("=" perché sono due modi per scriverla).*

Posso fare la tabella dei vari eventi condizionati.

Teorema 2 (Regola della somma)**.** *Si ha che, avendo la tabella degli eventi:*

$$P(x) = \sum_{y} P(x, y)$$

con $P(x)$ detta ***probabilità marginale****.*

La somma di tutte le possibili combinazioni di eventi, quindi dei valori della tabella, deve dare 1.
Su slide esempio di uso di quanto detto, dove si arriva al teorema di Bayes.
Si vuole infatti passare dal conoscere $P(a|b)$ al consocere $P(b|a)$.

Teorema 3 (Teorema di Bayes)**.** *Il teorema enuncia che:*

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Avendo:

- $P(h)$ *che è la probabilità conosciuta a priori di* h*. Tale probabilità riflette qualsiasi conoscenza di base sulla possibilità che* h *sia corretta*
- $P(D)$ *che è la probabilità conosciuta a priori di* D*, ovvero la probabilità che* D *sia osservato*
- $P(D|h)$ *che è la probabilità di osservare* D *in presenza dell'ipotesi* h
- $P(D|h)$ *che è la probabilità a posteriori di* h*. Tale probabilità riflette la "confidenza" di avere* h *dopo che* D *è stato osservato*

In altri termini, avendo:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

ho che:

$$P(a|b)P(b) = P(b|a)P(a)$$

arrivando a dire che:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

notando la correlazione tra probabilità congiunta e Bayes.
Che è il punto fondamentale della moderna teoria dell'intelligenza artificiale in quanto permette di raccogliere l'evidenza senza poi usare le tabelle delle probabilità congiunte, che sarebbero difficilissimi da osservare. Se pensiamo ad alcuni eventi come cause "nascoste" non necessariamente osservabili se modelliamo la verosimiglianza degli eventi osservabili date le cause nascoste si ha:

$$P(causa|effetto) = \frac{P(effetto|causa)P(causa)}{P(effetto)}$$

Si ha quindi un modello per inferire/derivare la verosimiglianza della causa nascosta e quindi rispondendo a:

$$P(causa|effetto)$$

Avendo quindi la probabilità di una causa dato un effetto. *Dato l'effetto modello la causa.*
Se non si ha una delle due probabilità a priori posso stimare per poi normalizzare. In altri termini il denominatore $P(D)$ è spesso solo una quantità di normalizzazione, essendo spesso difficile da stimare.
Le probabilità possono essere definite su due approcci:

- **approccio frequentista o oggettivista** che considera la probabilità come un'entità misurabile legata alla frequenza di accadimento, come si fa ne machine learning
- **approccio soggettivista** che considera la probabilità come una misura del grado di attesa soggettivo del verificarsi di un evento, come si fa nel corso di statistica

Capitolo 3

Modelli Probabilistici

Nel passato si sono usati **sistemi a regole**, dove codificando tutto quello che può succedere si cercava di giungere ad una decisione. Questo però era molto dispendioso, si arrivava o a vero a falso, senza via di mezzo, e si dovevano avere dati ipoteticamente completi e sicuri in partenza. Si parla in questo caso di **modelli logici**.
Viviamo in un'era dove si hanno molti dati, sia in ambito sociale, che di business che scientifico. Questi dati devono essere analizzati al fine di poter prendere **decisioni** e per farlo si deve per capire la situazione in cui ci si trova e spesso posso capirlo solo osservando i dati, non osservando la variabile specifica. Dalle osservazioni dobbiamo inferire il valore di variabili "nascoste". Spesso si ha a che fare con dati non completi, non consistenti, spesso errati, con rumore di trasmissione etc...
Tali dati sono comunque evidenze per percepire la situazione in cui ci si trova. L'obiettivo del corso è fornire strumenti modellistici per rappresentare l'incertezza nel modello, incertezza per struttura e parametri, e per rappresentare in termini probabilistici gli errori nei dati. Si vuole quindi implementare algoritmi di "ragionamento", automatizzati e adattivi, oltre che robusti e scalabili.
I **modelli probabilistici** sono anche detti **modelli generativi**. Si usa la teoria delle probabilità per esprimere incertezza e rumore associati al modello e ai dati, soprattutto usando la teoria Bayesana, per fare previsione e adattare i modelli. Questi modelli permettono di partire da una "credenza" iniziale, anche soggettiva, per poi raccogliere evidenze aggiustando tale modello.
I modelli probabilistici sono anche modelli di machine learning, in quanto apprendono.
Bisogna quindi partire dalle osservazioni generate rispetto ad un valore di variabile per poi inferire tale variabile (ad esempio parto dai risultati di un gioco per capire quando è bravo il giocatore, che non è una variabile che posso

sapere a priori). Si parte dai dati e si arriva al valore della variabile che ha generato questi dati (per questo *modello generativo*). Man mano che raccolgo informazioni raffino il modello, più o meno come fa un essere umano ("più rispondi alle domande all'orale e più il docente capisce il tuo voto, anche se alla fine non si ha la certezza che il voto rispecchi la preparazione"). I dati possono non portare alla certezza, ma più dati si hanno e più ci si avvicina, riducendo l'incertezza.
Un esempio pratico è il modello **Elo** (nato per gli scacchi) da cui deriva quello usato da *Xbox* per capire come appaiare giocatori online in base alle skill. Il valore di bravura viene rappresentato come una distribuzione, in primis con una Gaussiana, con una certa media e una certa varianza/deviazione standard, quindi solo due numeri. Cambiare il modello significa solo cambiare quei due valori. Per confrontare due giocatori capisco la distribuzione a partire dai dati del giocatore che si hanno, diminuendo l'incertezza all'aumentare dei dati. Con il modello probabilistico poi, a partire dal risultato modificherò le distribuzioni di partenza, cambiando la percezione su essi. Nel tempo posso tenere aggiornato i modelli probabilistici che rappresentano una certa variabile e usarli per fare confronti (ad esempio confrontando due giocatori per poi fare l'appaiamento).
Con i modelli probabilistici si ha una capacità espressiva maggiore di quella di un modello logico, avendo le distribuzioni di probabilità e potendo anche usare varie soglie.
Un *modello generativ*o parte dalle probabilità a priori e può "generare" possibili eventi, generando campioni verosimili con una certa distribuzione statistica.
Nella vita reale si osservano degli accadimenti e studiandoli si può risalire alla probabilità degli eventi, tramite l'approccio frequentista.

3.1 Incertezza

Si introduce quindi l'**incertezza**. Non sempre si ha a che fare con dati "certi" e precisi, che possono portare con più facilità ad una certa decisione, potendo giungere ad una decisione **ottimale** senza alcun dubbio su quale essa sia.
Con l'**incertezza** sui dati bisogna modificare l'idea di **soluzione ottimale**. Si arriva a dover capire quale sia la **soluzione ottima** in un contesto dove "non si sa cosa succederà", partendo da dati incerti.
Si ha che:

- un evento osservato può avere molte cause

- la verosimiglianza di un'ipotesi sulla causa cambia man mano che si raccolgono pezzi di evidenza
- usando modelli probabilistici di ragionamento possiamo calcolare quanto probabile è una certa ipotesi. Si ipotizza che le fonti di incertezza siano quantificabili

Vari esempi di vita in merito sulle slide.
Spesso si ha un approccio "frequentista", valutando la frequenza di un evento per capire la probabilità che tale evento accada, inferendo così una distribuzione di probabilità dalla frequenza con la quale si osservano i dati. Questo è piuù o meno come funziona il cervello umano ma bisogna fare la stessa cosa con un calcolatore e per questo ci verrà incontro il **teorema di Bayes**.
Si ha inoltre che un sistema che considera anche l'incertezza, che è presente in moltissime situazioni, dovrebbe funzionare meglio di uno che non lo fa ma ci serve in primis un modo per rappresentare l'incertezza stessa.
Più parametri ha il modello e più è difficile rappresentarlo.
Si ha il **Degree of Belief** che è una probabilità a priori sono ricavate da:

- osservazioni statistiche
- regole generali e conosciute
- combinazioni di sorgenti di evidenza

In ogni caso si hanno quindi evidenze empiriche.
Vediamo ora il rapporto tra i **modelli causali** e la **regola di Bayes**.
Ricordiamo che per Bayes si ha:

$$P(causa|effetto) = \frac{P(effetto|causa)P(causa)}{P(effetto)}$$

Con le reti causali vorremmo risalire dall'effetto alla causa ma normalmente si hanno più informazioni su $P(causa|effetto)$ che su $P(effetto|causa)$.
Conoscendo $P(effetto|causa)$ per ogni causa posso evitare di calcolare $P(effetto)$, infatti, dato $c = causa$ ed $e = effeto$:

$$P(c|e) = \frac{P(e|c)P(c)}{P(e)} = \frac{P(e|c)P(c)}{\sum_{\forall h \in causa} P(e|h)P(h)}$$

Vediamo un po' di notazione:

- con $< \top, \bot >$ indichiamo una distribuzione di probabilità

- α costante di normalizzazione per trascurare il denominatore di Bayes (lo sostituisce). È detto **fattore di normalizzazione**

Esempio 1. *Vediamo un esempio:*

$$P(meningite =< \top, \bot > | s = \top) = \alpha < P(s|m)P(m), P(s|\neg m)P(\neg m) >$$

SI assume che l'effetto deve essere scaturito a causa di una delle cause ipotizzate e non altre. A volte è più difficile calcolare $P(effeto|causa)$ per tutte le cause indipendentemente che calcolare direttamente $P(effetto)$.
Dato:

$$P(A|B) = \alpha P(B|A)P(A)$$

si ha che:

- $P(A)$ è la probabilità a priori
- $P(B|A)$ probabilità a posteriori
- $P(B|A)$ verosimiglianza

Se la probabilità a priori è nulla si assegna una probabilità ε (anche solo per un'osservazione) a tutti gli eventi che riteniamo possibili, anche se ancora non sono accaduti. Se un evento può realizzarsi deve avere una probabilità a priori, anche se molto piccola. Bisogna poi riscalare la probabilità di tutti per poter includere anche questi eventi rari. **Esempi su slide**.
Vediamo quindi come si **combinano le evidenze**. Qualora si abbiano più effetti il modello diventa più complesso. Per n effetti arei 2^n possibili combinazioni di evidenze da modellate. Si utilizza quindi la **catena di probabilità condizionali**, che, per esempio, per 4 eventi è:

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

ottenuta sfruttando la regola del prodotto:

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

La catena di probabilità condizionali è utile per rappresentare la probabilità congiunta in quanto permette una rappresentazione più compatta (potendo mettere anche insieme diverse fonti).

Definizione 9. *Considerato un insieme di eventi $E_1, \dots E_n$ e tutte le possibili combinazioni dei loro valori $\top$ e $\bot$. Supponiamo di conoscere tutti i valori $P(E_1, \dots, E_n)$. Supponiamo che un sottoinsieme di questi presenti un*

valore definito, ovvero $E_j = e = \top$ allora chiamo ***inferenza probabilistica*** *il processo di calcolo del valore:*

$$P(E_i = \top | E_i = e)$$

In generale l'inferenza probabilistica non è trattabile con questo metodo avendo una lista 2^n probabilità congiunte $P(E_1, \dots E_n)$ (lista che per di più spesso non abbiamo).
Si ragiona quindi spesso tramite ***metodi approssimati/qualitativi****, avendo magari centinaia di evidenze.*

Esempio su slide.

Per risolvere il problema viene anche incontro l'**indipendenza condizionata**.
Due eventi possono diventare indipendenti data la presenza di un altro evento, che è causa comune di entrambi. Si passa da una dipendenza causale diretta alla dipendenza dovuta ad un effetto causale indiretto (???). Se si conosce la causa i due eventi sono indipendenti se non la si conosce potrebbero essere dipendenti.

Definizione 10. *Definiamo la* ***regola di marginalizzazione*** *per due insiemi di variabili Y e Z come:*

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

In alternativa uso le probabilità condizionate usando la ***regola del condizionamento****:*

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

*Potrei anche usare l'****inferenza per enumerazione*** *dove semplicemente sommo i valori della tabella rispetto a ciò che mi interessa (se voglio $P(c = \top, m = \top)$ sommo tutti i valori con almeno uno dei due nella tabella).*

Su slide esempio di conto per tutti con anche conto per α.
Un principio generale di computazione è:

- specificare la variabile oggetto della "query"
- fissare lo stato delle variabili per le quali è disponibile l'evidenza
- calcolare la probabilità a posteriori sommando rispetto alle variabili sulle quali non è disponibile evidenza

Quindi indiciamo con x tale variabile oggetto di query. Data la realizzazione congiunta $e(evidenza)$ per un sottoinsieme E di variabili dette **variabili con evidenza** si indica con Y l'insieme restanti variabili. Y è detto insieme delle variabili senza evidenza. L'intero insieme delle variabili del problema è quindi:

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

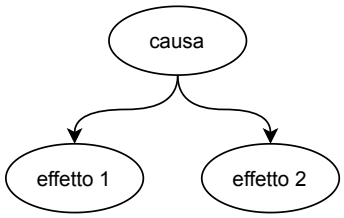
La distribuzione marginale a posteriori di X è ottenuto per marginalizzazione rispetto a Y:

$$P(X|E = e) = \alpha P(X, E = e) = \alpha \sum P(X, E = e, Y = y)$$

Posso quindi fare query per qualsiasi variabili avendo la tabella delle probabilità congiunte ma tale metodo non è efficiente.

3.1.1 Reti Bayesiane

Le relazioni di indipendenza condizionata può essere illustrata da un grafo, dove un nodo è collegato ad un altro con arco diretto sse il primo è causa dell'altro:



Esempio 2. *Vediamo un esempio:*

$$P(carie|mal\_di\_denti \wedge incastro) = \alpha P(|mal\_di\_denti \wedge incastro|carie)P(carie)$$

$$= \alpha P(|mal\_di\_denti|carie)P(incastro|carie)P(carie)$$

Ulteriori esempi su slide.
Le probabilità congiunte le posso rappresentare in una tabella.
Le **reti Bayesiane** sfruttano grafi diretti aciclici per rappresentare le assunzioni di indipendenza condizionale tra variabili in modo chiaro ed efficiente. Un arco diretto tra A e B rappresenta una relazione di causalità: A influenza B. Si ha quindi che "pattern" di ragionamenti sono un cammino tra un nodo e un altro.
Si passa da $O(2^n)$ a $O(n)$, per n numero di effetti.

Definizione 11. *Si ha che l'evento A è condizionalmente indipendente dall'evento B se, dato l'evento C:*

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

ovvero la conoscenza di B non porta a nessuna ulteriore variazione della probabilità di A rispetto a quella dell'avverarsi di C.
Dall'indipendenza di A e B dato C si ha che:

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

Se C è un insieme vuoto ho, non avendo correlazione:

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

Le reti Bayesiane quindi analizzano le cause dirette e indirette permettendo di rappresentare in modo efficiente la distribuzione congiunta di probabilità, tramite dipendenza e indipendenza condizionale.
L'inferenza basata su enumerazione è in $O(d^n)$ sia in spazio che tempo con:

- d massima cardinalità del supporto (se binario $d = 2$)
- n numero di variabili

è questo non va bene.
Una distribuzione congiunta può essere rappresentata come produttoria di n valori di probabilità di **eventi indipendenti**, passando da un arrivando a $O(n)$, avendo:

$$P(C_1, \dots C_n) = P(C_1) \cdots \dots P(C_n)$$

Rappresentando quindi, tramite l'indipendenza delle variabili, in modo compatto una distribuzione congiunta.
Nel mondo reale però non si ha indipendenza assoluta tra le variabili e spesso anche il gran numero di variabili rende difficile la specifica di una distribuzione congiunta.
Si usa quindi l'indipendenza condizionale, sfruttando le variabili condizionalmente indipendenti.
Per scrivere la definizione congiunta usiamo la **chain rule**. Dato $P(X,Y,Z)$ si ha che, avendo X e Y indipendenti:

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

riducendo quindi il numero di valori di probabilità necessari al conto.
Le asserzioni di indipendenza condizionate si basano sul dominio in analisi e

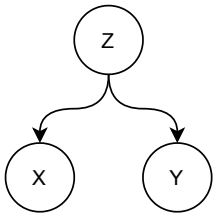
consente di limitare le complessità del modello. Il caso in cui tutte le variabili sono indipendenti ho la fattorizzazione delle probabilità delle singole variabili ed è un caso specifico di questo discorso.
Ricordando che per Bayes:

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

calcolando la probabilità della causa data la conoscenza dello stato degli effetti. Quindi, avendo X e Y indipendenti:

$$P(X|Y,Z) = \alpha P(X,Y|Z)P(Z) = \alpha P(X|Z)P(Y|Z)P(Z)$$

Avendo graficamente $P(X,Y,Z) = \alpha P(X|Z)P(Y|Z)P(Z)$:



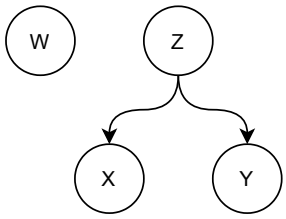
Possiamo quindi parlare meglio delle **reti Bayesiane**, spesso indicata con:

- **Bayesian Belief Network, BBN**
- **Probabilistic Network, PN**
- **Causal Network, CN**

Tali reti appartengono alla classe dei **modelli grafico-probabilistici**.

Definizione 12. *Una **rete Bayesiana** è un grafo cui nodi sono annotati da una informazione quantitativa, tramite tabelle di probabilità condizionata, e i cui archi definiscono dipendenza e indipendenza tra le variabili dei nodi. Si hanno solo archi orientati. Se X è causa diretta di Y ho un arco tra X e Y. X è detto genitore e Y figlio. Le variabili possono essere sia continue che discrete.*
La topologia della rete e le probabilità condizionate dei nodi dati genitori sono sufficienti a specificare (implicitamente) la distribuzione congiunta di tutte le variabili.
Uno o più nodi isolati segnalano indipendenza assoluta. Due figli di uno stesso

genitore segnalano che sono condizionalmente indipendenti. Nell'immagine X e Y sono condizionalmente indipendenti, W è indipendente dalle altre 3 variabili, Z è causa diretta di X e Y mentre tra queste ultime non esiste una relazione diretta di causalità:



Tale grafo non contiene cicli e quindi si parla di ***Directed Acyclic Graph (DAG)****, non essendo possibile che una variabile causi se stessa.*

Come detto la componente quantitativa è costruita da un insieme di tabella di probabilità condizionale. Ogni nodo ha associata quindi una *Conditioned Probability Table (CPT)* che traduce l'impatto dei genitori sulla variabile stessa.
Su slide esempio esteso.
Ricapitolando:

- ogni nodo ha CPT
- ogni riga della CPT somma ad uno (e se ho un solo valore non scrivo sia il vero che il falso visto che posso fare $1-$)
- la CPT di una variabile booleana con K variabili genitori contiene 2^K valori che possono essere specificati indipendentemente
- una variabile senza genitori ha una sola riga con i valori di probabilità a priori per ogni possibile valore che la variabile può assumere

Possiamo dire che si hanno due chiavi di lettura dal punto di vista della semantica, semanticamente equivalenti:

- la rete rappresenta una **distribuzione congiunta di probabilità**. Questa lettura è utile per progettare e implementare procedure di inferenza. Per questa lettura dice che ogni rete 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:

$$P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n) = P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

che è detta **formula di fattorizzazione**.
Avendo che con le maiuscole abbiamo le variabili (anche per $Parents$) e con le minuscole le realizzazioni (anche per $parents$). Si ha quindi che ogni elemento della distribuzione congiunta è rappresentato come prodotto di opportune componenti delle CPT che costituiscono quindi una rappresentazione decomposta della distribuzione di probabilità congiunta. Per questa rappresentazione posso usare le reti per rispondere a qualsiasi query relative al dominio che descrive tramite marginalizzazioni.
Esempio su slide.

- la rete codifica un **insieme di relazioni di indipendenza condizionale**. Questa lettura è utile per costruire un modello di rete Bayesiana.

Sfrutto la formula di fattorizzazione per determinate la componente topologica della rete. Si ricorda che per la **cchin rule**:

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) \cdot P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdot \ldots \cdot P(x_2 \mid x_1) \cdot P(x_1)$$

$$= \prod_{i=1}^{n} P(x_i \mid x_{n-1}, \ldots, x_1)$$

Noto che tale formula è confrontabile con la formula di fattorizzazione e ho che, a patto che $Parents(X_1) \subseteq \{x_{i-1}, \ldots, x_1\}$:

$$\mathbf{P}(x_i \mid x_{i-1}, \ldots, x_1) = \mathbf{P}(x_i \mid \text{ parents } (X_i))$$

Quindi 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. Quindi i genitori di X_i devono essere scelti da $\{X_1, \ldots, X_{i-1}\}$.
Si ha quindi la seguente procedura di costruzione incrementale della componente topologica;

1. Si seleziona un insieme di variabili $\{X_1, \ldots, X_n\}$ per descrivere il modello

2. scelgo un ordinamento per le variabili $\{X_{(1)}, \ldots, X_{(n)}\}$. Con un ordinamento sbagliato arrivo a definire reti sbagliate o reti più complesse del necessario, magari con informazioni ridondanti

3. inizializzo i nodi aggiunti alla rate partendo da $i = 1$

4. si seleziona la variabile $X_{(i)}$ alla rete, si pone $Parents(X_{(i)})$ uguale all'insieme minimale di nodi, attualmente appartenenti alla rete $X_{(1)}, \ldots, X_{(i-1)}$, che soddisfa la proprietà di indipendenza condizionale:
$$\mathbf{P}\left(X_{(i)} \mid X_{(i-1)}, \ldots, X\right) = \mathbf{P}\left(X_{(i)} \mid \text{ Parents } \left(X_{(i)}\right)\right.$$
per poi calcolare la CPT di $X_{(i)}$

5. tengo conto del nodo aggiunto nel conto totale dei nodi e faccio i++. Se ho finito le variabili termino altrimenti torno a 4)

Una rete è più compatta, solitamente, dell'intera distribuzione di probabilità congiunta e tale compattezza è un esempio della proprietà dei **sistemi strutturati localmente o sparsi**, dove ogni sotto-componente interagisce solo con un numero limitato di altre componenti, indipendentemente dal numero totale di componenti del sistema. La strutturazione locale normalmente ha crescita lineare e non esponenziale (mediamente con una rete Bayesiana è ragionevole pensare che ogni variabile sia influenzata da al massimo k, con k costante, variabili).
Nel caso in cui si consideri una Rete Bayesiana costituita da n variabili (nodi) booleane. Si ha che la quantità di informazione per specificare una CPT è limitata superiormente da 2^k per cui la rete completa richiede $n \cdot 2^k$ cifre contro i 2^n dell'intera distribuzione di probabilità congiunta.
Si hanno quindi vari tipi di rete, relativi a situazioni diametralmente opposte:

- **rete completamente connessa (*fully connected network*)**, dove ogni variabile può essere potenzialmente influenzata da tutte le altre e quindi specificare una CPT richiede le stesse informazioni della distribuzione congiunta

- **rete con relazioni di causalità tenui** dove se aggiungo delle relazioni, degli archi aumento la complessità della rete (???). Si valuta il guagno in accuratezza contro quelle in complessità

Anche in un dominio strutturato localmente la costruzione di una Rete Bayesiana non è semplice, volendo un numero limitato di influenze per ogni variabile e che la topologia della rete rifletta le relazioni di influenza diretta. La procedura di costruzione di una Rete Bayesiana funziona in modo tale che quando si giunge ad aggiungere un nodo, i nodi candidati ad essere suoi genitori, ovvero i nodi che lo influenzano direttamente o indirettamente, siano già stati aggiunti alla corrente struttura della rete. Il corretto ordine comporta che si devono prima stabilire le cause radice per poi aggiungere quelle variabili che da loro vengono influenzate,

procedendo fino alle foglie che non sono causa di nulla.
lato **semantica topologica** abbiamo due specificazioni equivalenti:

- un nodo è condizionalmente indipendente dai suoi non-discendenti dati suoi genitori
- 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. Tale insieme di nodi è detto **Markov Blanket**

Vediamo come capire se in una rete Bayesiana due variabili sono condizionalmente indipendenti.

Definizione 13. *Definiamo che X e X sono **d-separati** da un insieme E di variabili con evidenza (ovvero osservazioni) sse ogni cammino non orientato da X a Z è **bloccato** ovvero sse:*

- *lungo il cammino si ha una variabile $V \in E$ (quindi di cui conosco il valore) e gli archi che collegano V al cammino sono solo **tail to tail** (ovvero da V escono i due archi: $\rightarrow V \leftarrow$) e quindi il cammino è bloccato da V*
- *lungo il cammino si ha una variabile $V \in E$ (quindi di cui conosco il valore) e gli archi che collegano V al cammino sono solo **tail to head** (ovvero un arco entra in V ed uno esce: $\rightarrow V \rightarrow$) e quindi il cammino è bloccato da V*
- *lungo il cammino si ha una variabile $V \notin E$ (quindi di cui NON conosco il valore) tale che nessuno dei suoi discenti appartiene all'insieme E e gli archi che collegano V al cammino sono solo **head to head***

Esempi su slide.

Teorema 4 (teorema di Verma e Pearl)**.** *Se in una rete Bayesiana un insieme E di variabili con evidenza d-separa X e Y allora X e Y sono indipendenti.*

La d-separazione si calcola in tempo lineare avendo quindi un algoritmo efficiente per inferire automaticamente se apprendere il valore di una variabile può fornire informazioni aggiuntivi su altre variabili date le informazioni a disposizione quindi può essere utile per capire se apprendere il valore di una delle due variabili può aiutarci ad avere informazione aggiuntiva rispetto all'altra.
Si hanno algoritmi di propagazione dell'informazione che portano a calcolare la distribuzione di probabilità, incondizionata o marginale, su ogni nodo. Tale distribuzione è detta anche **belief function** del nodo.
L'algoritmo di propagazione aggiorna le belief di ogni nodo della rete ogni volta che si ha una nuova evidenza (???).

Esempi su slide. Su elearning file sulla *Moralizzazione*, argomento extra.
Anche avendo massimo k genitori per ogni variabile si ha che servono $O(2^k)$ parametri per ogni CPT.
Si ha una distribuzione canonica per rappresentare pattern standard per avere un numero limitato di parametri per compilare la CPT. Uno di questi pattern è detto **modi deterministici**. Un nodo deterministico ha che il suo valore è determinato unicamente da quello dei genitori, senza incertezza, avendo quindi una relazione ad esempio logica (*and, or,* etc...) o numerica (*max, min,* etc...). Ovviamente non tutta la rete è fatta da questi nodi altrimenti non userei una rete Bayesiana ma in una rete complessa si possono avere tali pattern. Tramite tali relazioni calcolo direttamente i valori delle CPT.
Un altro pattern, in presenza di incertezza, è quello **noisy logical relationship**, con generalizzazioni degli operatori logici. Un esempio è il **noisy-OR**, che introduce incertezza circa la capacità di causare $\top$ nella variabile figlio da parte del nodo genitore. Si ha che la relazione di causalità tra genitore e figlio potrebbe essere inibita. Con noisy-OR si hanno due ipotesi:

- tutte le possibili cause sono note (eventualmente si possono aggiungere *leak node*, un nodo fittizio per rendere esaustive la cause)
- l'inibizione di un genitore è indipendente da quella di altri genitori per il nodo considerato

L'inibizione è legata ad una certa probabilità, dette *probabilità di inibizione*, ovvero **noise parameters**.
La probabilità di un evento viene ottenuta tramite il prodotto delle probabilità di inibizione di ogni nodo genitore. Quindi per certe combinazioni di or non ho il valore preciso ma una certa probabilità. Se esattamente un genitore è $\top$ allora il figlio è $\top$ con probabilità pari al noise parameter di tale genitore. Se nessun genitore è nello stato $\top$ allora certamente il figlio è nello stato $\bot$. Negli altri casi appunto moltiplico i noisy parameter dei genitori con valore $\top$. Ho quindi limitato il numero di parametri. Quindi se una variabile dipende da k genitori di norma ho la noisy-OR con $O(k)$ parametri e non $O(2^k)$, per descrivere l'intera distribuzione di probabilità congiunta.
Esempi su slide.
Potrei avere **reti con nodi continui** e potendo assumere un numero infinito di valori si rende impossibile specificare esplicitamente i valori della distribuzione condizionale e di quella a priori. Si procede quindi **discetizzando** i valori in un numero finito di intervalli, perdendo comunque di informazione, soprattutto con grandi CPT. Una alternativa sono le **funzioni di densità di probabilità**, trascritte un numero finito e, solitamente, contenuto di parametri.
Un esempio di tale funzione è la **gaussiana univariata**:

$$N(\mu, \sigma^2)$$

che ha due parametri: *media* e *varianza*.
Una rete con nodi discreti e continui insieme è detta **rete Bayesiana ibrida** e per la sua definizione servono due 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

Usando quindi questa *gaussiana lineare* (comoda perché facendo operazioni tra gaussiane si ottiene comunque una gaussiana) perché il nodo figlio è distribuito secondo una distribuzione gaussiana la cui media μ varia linearmente con il valore del genitore mentre la deviazione standard σ è fissata. In pratica si fa una sorta di regressione lineare con rumore gaussiano. Si ha che la gaussiana lineare ha 2 proprietà:

1. una rete che contiene solo nodi continui con distribuzione Gaussiana Lineare ha una distribuzione di probabilità congiunta **Gaussiana multivariata**
2. 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**

Avendo una cosa del tipo $P(c|r, s = \top)$ si avrebbe:

$$N(a_\top \cdot r + b_\top, \sigma_\top^2)(c) = \frac{1}{\sqrt{2\pi} \cdot \sigma_\top} e^{-\frac{1}{2}\left(\frac{c-(a_\top \cdot r + b_\top)}{\sigma_\top}\right)^2}$$

dovendo poi fare un discorso analogo per $\perp$. Si ha che a e b sono il legame con la regressione lineare.
Esempio su slide.
Anche usando modelli semplici si possono ottenere distribuzioni flessibili e interessanti.
Passiamo ora alla **distribuzione associata a variabili discrete con genitori continui**, tramite una **funzione a soglia soft**.
Un modo di ottenere soglie soft è usare l'integrale della normale standard:

$$\phi(x) = \int_{-\infty}^{x} N(0,1)(x)\,\mathrm{d}x$$

usando la distribuzione **probit**:

$$P(X|Y = c)\phi(\frac{-c+\mu}{\sigma})$$

il che significa che la soglia di costo si verifica intorno al valore , lo spessore della regione soglia è proporzionale al valore σ, e la probabilità di acquistare diminuisce all'aumentare del costo.
In alternativa si ha la funzione **probit**, che utilizza la **funzione sigmoid** per ottenere una soglia soft:

$$P(X|Y = c)\phi(\frac{1}{1 + e^{-2 \cdot \frac{-c+\mu}{\sigma}}})$$

3.1.2 Inferenza nelle Reti Bayesiana

Lo scopo di un modello probabilistico è quello di ottenere la distribuzione a posteriori, per un certo insieme di variabili query, in presenza di un evento, ovvero un assegnamento congiunto di valori ad un insieme di variabili con evidenza. Sono dati:

- X variabile query
- E insieme delle evidenze E_i
- e evento specifico
- Y insieme variabili non evidenziate Y_i

L'insieme completo delle variabili è:

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

e si hanno query del tipo:

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

Dato un modello di rete Bayesiana si hanno 4 tipologie di inerenza;

- **diagnostica**, avendo l'effetto si vuole capire la causa
- **causale**, avendo la causa voglio stimare la probabilità degli effetti
- **intercausale**, avendo una causa, capire la probabilità un'altra causa di un effetto comune
- **mista**, avendo sia diagnostica che causale o sia diagnostica che intercausale

Avere osservato che uno degli eventi fa divenire meno probabile che un altro si sia verificato è il meccanismo di **explaining away**.
Con una rete Bayesiana quindi posso cercare:

- probabilità condizionata $P(X|e)$

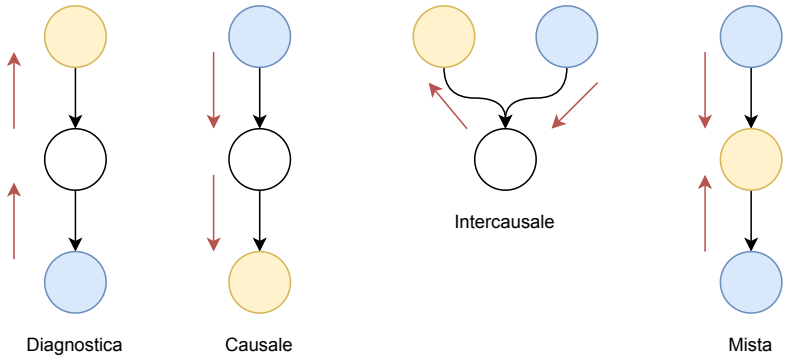


Figura 3.1: Tipi di inferenza, in blu l'evidenza e in giallo la variabile X

- quale valore massimizza $P(X|e)$, cercando la massima probabilità a posteriori
- quale è la distribuzione di probabilità di X dato e (che è il caso generale)

Ogni distribuzione condizionale può essere ottenuta tramite un procedimento di somma di opportuni elementi dell'intera distribuzione di probabilità congiunta (marginalizzazione). Si ha quindi che:

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

e con la regola di fattorizzazione delle reti Bayesiane:

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

e quindi $P(X, E = e, Y = y)$ nella distribuzione congiunta, possono essere scritti sotto forma di prodotti di probabilità condizionali della rete. In definitiva concludiamo che:

Ad una query è possibile rispondere utilizzando una Rete Bayesiana tramite la computazione di somme di prodotti di probabilità condizionali della rete