Introduzione alle Reti Bayesiane

Giovedì, 18 Novembre 2004

Francesco Folino

Riferimenti:

Chapter 6, Mitchell

A Tutorial on Learning with Bayesian Networks, Heckerman

Perchè ci interessano?

- **Sono una struttura teorica molto utilizzata nell'ambito dell'apprendimeno, della classificazione, della rappresentazione della conoscenza**
- **I metodi Bayesiani sono importanti perchè capaci di gestire data sets incompleti e rumorosi**
- **I metodi Bayesiani vengono oggi comunemente applicati**

Approccio Bayesiano alla probabilità ed alla statistica

- **Probabilità classica : è una proprietà fisica del mondo. E' la vera probabilità.**
- **Probabilità Bayesiana : è il grado con il quale una persona crede in un evento X. E' una probabilità personale.**
- **Al contrario della probabilità classica, le probabilità bayesiane beneficiano del fatto che non sono richieste prove ripetute. Il focus è sul prossimo evento (es. qual'è la probabilità che la Reggina vinca il campionato di serie A?)**

Il teorema di Bayes

Regola del prodotto:

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

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

Regola di Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

i.e.

$$P(Class|evidence) = \frac{P(evidence|Class)P(Class)}{P(evidence)}$$

Tutti i metodi di classificazione possono essere visti come stime della regola di Bayes facendo uso di differenti tecniche per determinare P(evidence|Class).

Introduzione

- **Il dominio del problema A è modellato attraverso una lista di variabili $X_1, \ldots, X_n$**
- **La conoscenza inerente il dominio del problema è rappresentata dalla sua distribuzione di probabilità congiunta $P(X_1, \ldots, X_n)$**

Esempio

- **Consideriamo il seguente problema: l'allarme**
  - **La storia: A Los Angeles scassinatori e terremoto sono abbastanza comuni. Entrambe fanno scattare l'allarme. In caso di allarme, due vicini John and Mary possono telefonare**
  - **Problema: Stimare la probabilità che ci sia uno scassinatore sulla base di chi ha o no telefonato**
  - **Variabili: Scassinatore (B), Terremoto (E), Allarme (A), Chiamata di John (J), Chiamata di Mary (M)**
  - **Conoscenza richiesta per risolvere il problema:**

    P(B, E, A, J, M) (distribuzione congiunta di probabilità)

Distribuzione congiunta

$P(B, E, A, J, M)$

B	E	A	J	M	Prob	B	E	A	J	M	Prob
y	y	y	y	y	.00001	n	y	y	y	y	.0002
y	y	y	y	n	.000025	n	y	y	y	n	.0004
y	y	y	n	y	.000025	n	y	y	n	y	.0004
y	y	y	n	n	.00000	n	y	y	n	n	.0002
y	y	n	y	y	.00001	n	y	n	y	y	.0002
y	y	n	y	n	.000015	n	y	n	y	n	.0002
y	y	n	n	y	.000015	n	y	n	n	y	.0002
y	y	n	n	n	.0000	n	y	n	n	n	.0002
y	n	y	y	y	.00001	n	n	y	y	y	.0001
y	n	y	y	n	.000025	n	n	y	y	n	.0002
y	n	y	n	y	.000025	n	n	y	n	y	.0002
y	n	y	n	n	.0000	n	n	y	n	n	.0001
y	n	n	y	y	.00001	n	n	n	y	y	.0001
y	n	n	y	n	.00001	n	n	n	y	n	.0001
y	n	n	n	y	.00001	n	n	n	n	y	.0001
y	n	n	n	n	.00000	n	n	n	n	n	.996

Come si usa la distribuzione congiunta

- **Qual'è la probabilità che ci sia uno scassinatore dato che Mary ha telefonato P(B = y | M = y)?**
- **Si calcola la probabilità marginale: P(B , M) = $\sum_{E, A, J}$ P(B, E, A, J, M)**
- **Si usa infine la definizione di probabilità condizionata:**

$$P(B = y \mid M = y) = P(B = y, M = y) \,/\, P(M = y)$$

Introduzione

- **Difficoltà: complessità nella costruzione del modello e nell'inferenza**
- **Nell'esempio dell'allarme:**
  - **Sono richiesti 31 valori di probabilità (2^5-1)**
  - **Calcolare P(B = y | M = y) richiede un gran numero di addizioni(29)**
- **In generale**
  - **$P(X_1, \ldots X_n)$ richiede almemo $2^n - 1$ valori per specificare la probabilità congiunta**
  - **Inferenza e spazio di memorizzazione esponenziali**

Indipendenza Condizionata

- **Il superamento del problema della taglia esponenziale può essere ottenuto sfruttando il concetto di Indipendenza Condizionata.**
- **La “chain-rule” è:**

$$
\begin{aligned}
p(x) &= p(x_1, \ldots, x_n) \\
&= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots \\
&= \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1})
\end{aligned}
$$

Indipendenza Condizionata

- **Independenza condizionata nel dominio del problema:**
  - **Il dominio generalmente consente di identificare un sottoinsieme *pa(X_i)* (genitori di X_i) di *{X_1, …, X_{i-1}}* tale che, dato *pa(X_i)*, X_i è independente da tutte le variabili in *{X_1, …, X_{i-1}}* tranne *pa{X_i}*, cioè:**

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

Possiamo perciò riscrivere la chain-rule così:

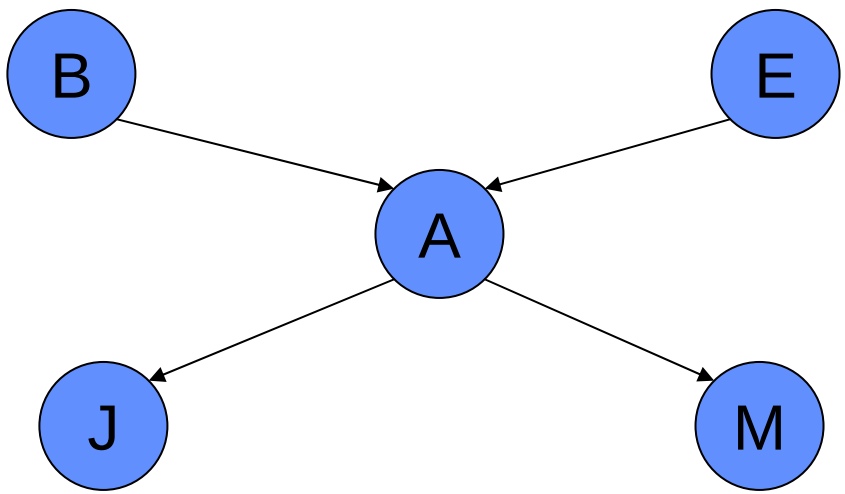
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$

Independenza Condizionata

- **Come risultati otteniamo:**
  - **la probabilità congiunta $P(X_1, \ldots, X_n)$ può essere rappresentata attraverso le probabilità condizionate $P(X_i \mid pa(X_i))$**
  - **riduzione della dimensione del modello**
  - **più semplice costruzione del modello**
  - **inferenza più semplice**

Rappresentazione grafica dell'indipendenza condizionata

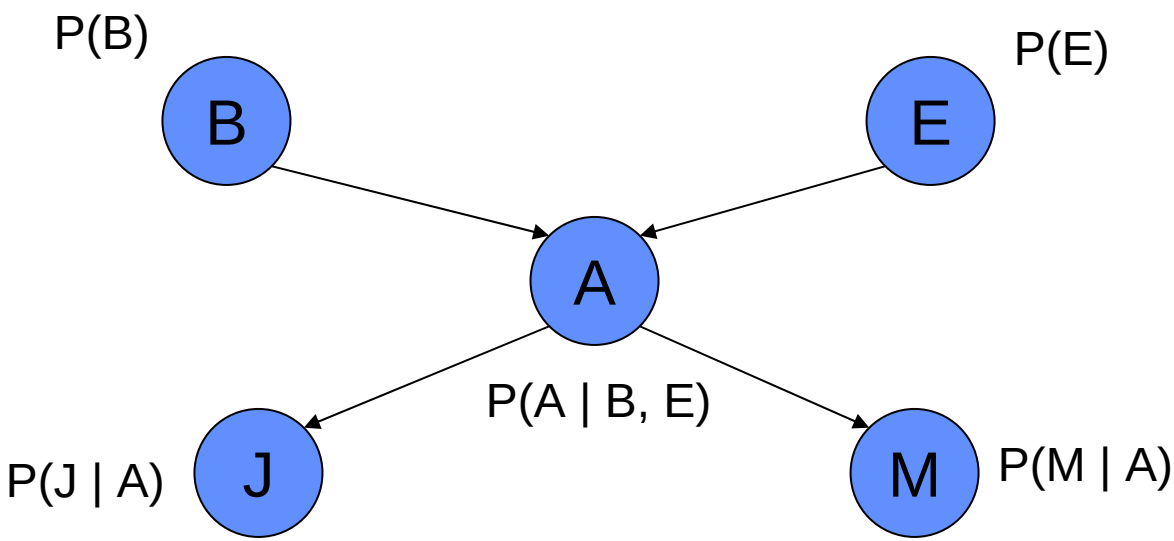
- **Per rappresentare graficamente le relazioni di dipendenza condizionata, basta costruire un grafo diretto tracciando un arco da X_j to X_i se e solo se X_j appartiene a pa(X_i).**
  - **Se questi sono i parents calcolati per l'esempio dell'allarme:**



pa(B) = {}
pa(E) = {}
pa(A) = {B, E}
pa{J} = {A}
pa{M} = {A}

Rappresentazione Grafica

- **Se aggiungiamo alla rete anche le tabelle delle probabilità condizionata $P(X_i \mid pa(X_i))$ per ogni nodo X_i, finalmente otteniamo la rete Bayesiana:**

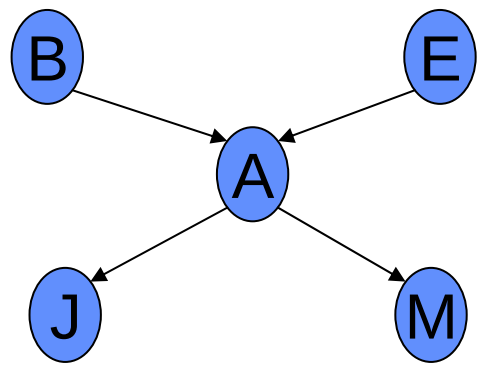


Definizione formale

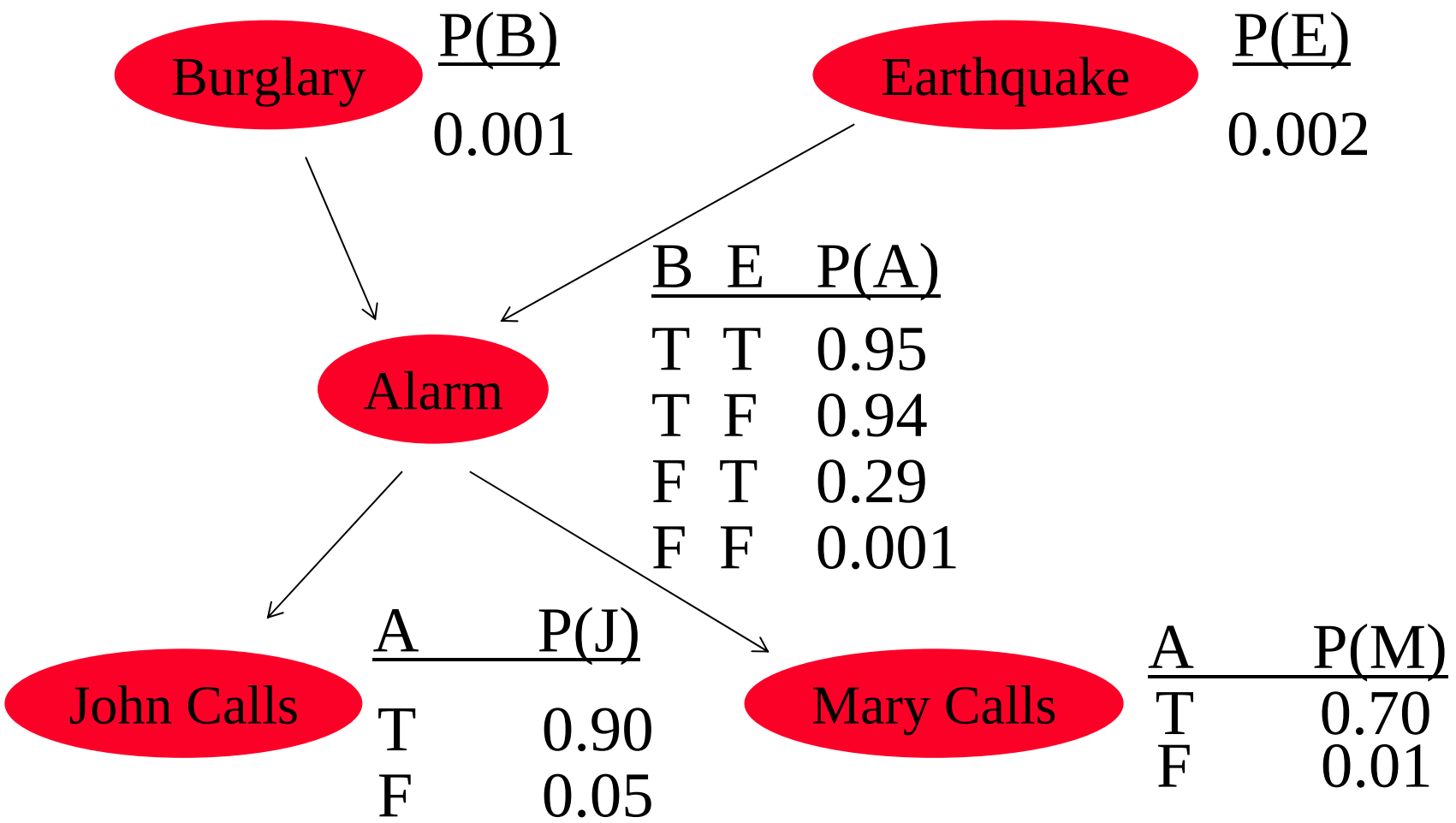
- **Una rete Bayesiana è:**
  - **Un D(ircect)A(cyclic)G(raph) che specifica le dependenze fra variabili e rappresenta la distribuzione di probabilità congiunta dove:**
    - **le variabili random costituiscono i nodi della rete**
    - **gli archi reppresentano influenze causali dirette**
    - **ad ogni nodo è associata una tabella contenente le probabilità condizionate dagli effetti dei propri parents**
    - **Non ci sono cicli diretti**

Intuitivamente…

- **Una BN può essere vista come un DAG dove gli archi reppresentano la dipendenza diretta**
- **L'assenza di un arco, invece, indica l'indipendenza: una variabile è *condizionalmente independente* da tutti i suoi nondiscendenti dati i suoi parents.**
  - **Def: diciamo che X è discendente da Y se esiste un percorso diretto da Y a X**



La rete Bayesiana completa (l'Allarme)



Costruzione formale delle BN

- **Passi per la costruzione di una BN:**
  - **Scegliere un insieme di variabili che descrivono il dominio dell'applicazione**
  - **Scegliere un ordine per le variabili**
  - **Partire dalla rete vuota ed aggiungere le variabili alla rete una per una in accordo all'ordine prescelto**

Construzione della BN (algoritmo)

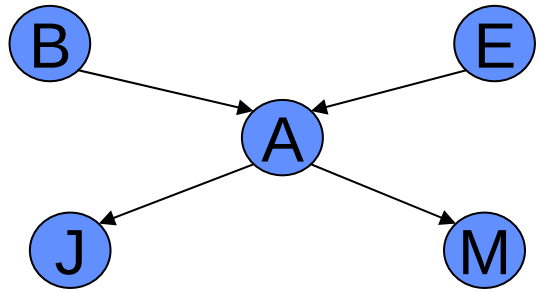
- **Aggiungere l' i-sima variabile X_i:**
  - **Determinare pa(X_i) delle variabili già nella rete (X_1, …, X_{i-1}) tale che:**

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

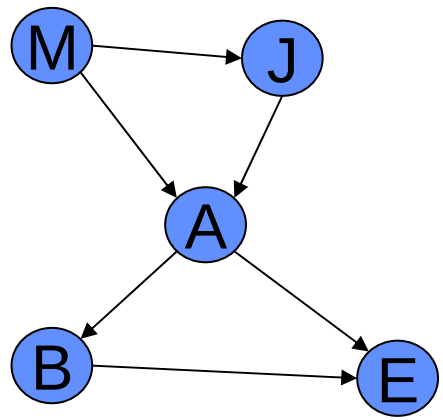
  - **Tracciare un arco da ognuna delle variabili in pa(X_i) a X_i**

Esempi

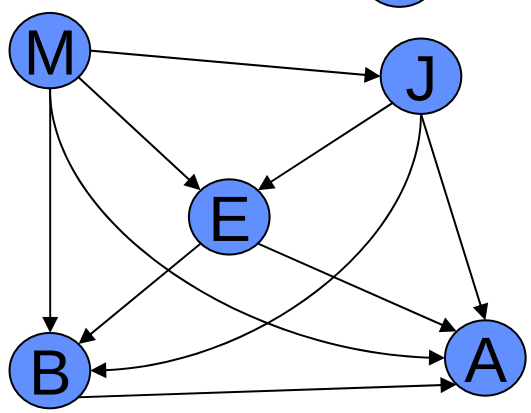
- **Ordine: B, E, A, J, M**
  - **pa(B)=pa(E)={}, pa(A)={B,E}, pa(J)={A}, pa{M}={A}**



- **Ordine: M, J, A, B, E**
  - **pa{M}={}, pa{J}={M}, pa{A}={M,J}, pa{B}={A}, pa{E}={A,B}**



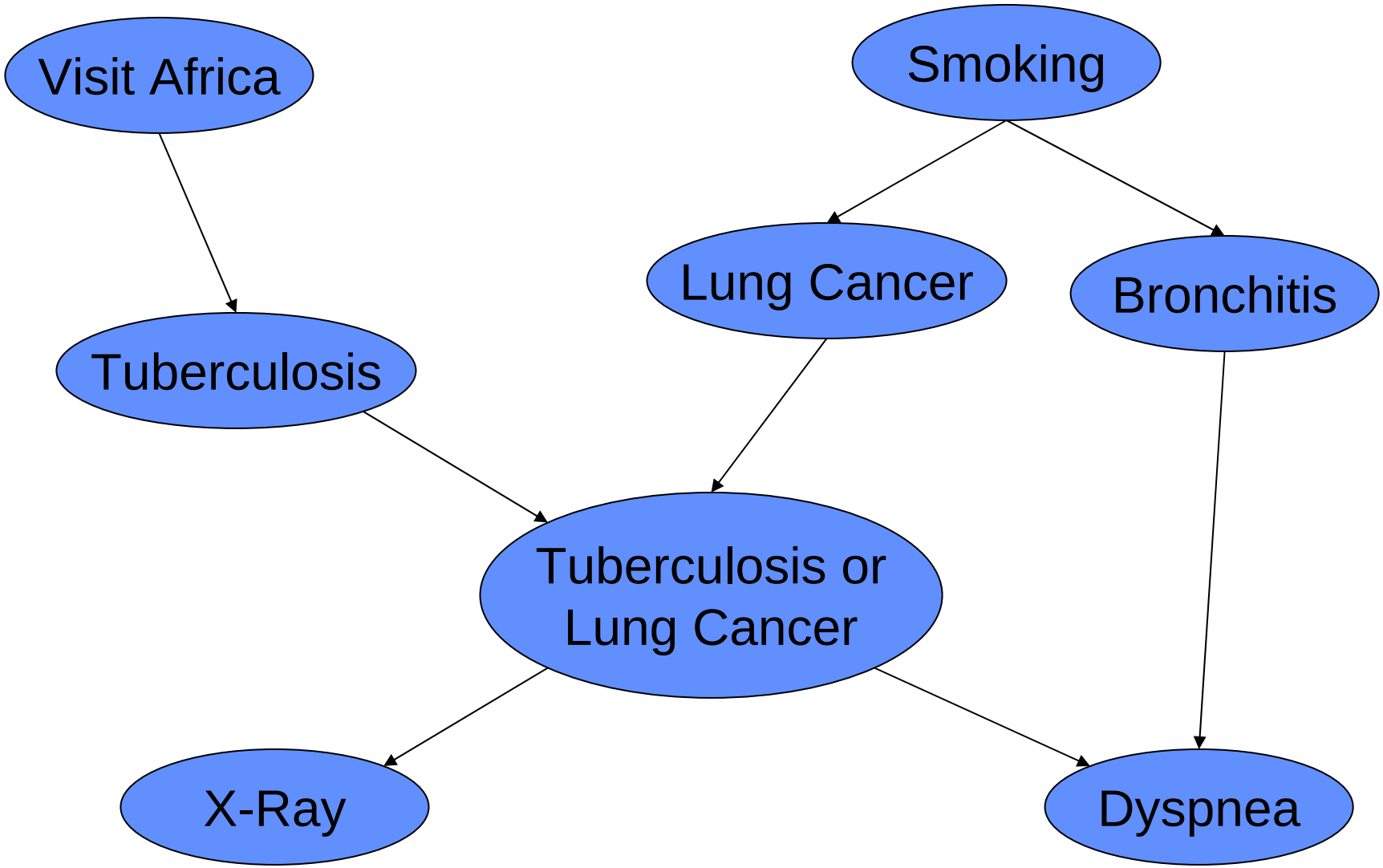
- **Ordine: M, J, E, B, A**
  - **Grafo completamente connesso**



Reti Bayesiane Causali

- Il precedente metodo per costruire una BN risente di alcune controindicazioni. In particolare, la scelta dell'ordine delle variabili è un task delicato
- Scegliere un ordine sbagliato può portare la BN a degenerare verso un grafo completamente connesso (il caso peggiore)
- Si usa allora un approccio causale
- Una rete Bayesiana causale, o simplicemente una rete causale, è una rete Bayesiana i cui archi sono interpretati come indicanti una relazione di causa-effetto.
- Per costruire una rete causale:
  - Scegliamo un insieme di variabili che rappresentano il dominio
  - Tracciare un arco verso una variabile a partire da ognuna delle sua cause dirette.

Esempio


Visit Africa
Smoking
Lung Cancer
Bronchitis
Tuberculosis
Tuberculosis or
Lung Cancer
X-Ray
Dyspnea

Ricapitoliamo…

- **Sia U={x1,……,xk} un insieme di variabili. Una *Bayesian Network* su U è un grafo aciclico diretto (DAG) su U più un insieme di tabelle di probabilità p(u|pa(u)) dove pa(u) è l'insieme dei parents(genitori) di u.**
- **Una BN rappresenta la distribuzione di probabilità:**

$$P(U) = \prod_{u \in U} p(u \mid pa(u))$$

Inferenza

- **Per usare una BN come un classificatore, occorre calcolare:**
  - $\arg\max_y P(y \mid \vec{x})$ **dove y rappresenta la *variabile classe* e x l'istanza da classificare.**

- **Usando la distribuzione di probabilità P(U) rappresentata dalla BN, possiamo scrivere che:**

$$P(y \mid \vec{x}) = P(U) \,/\, P(\vec{x}) \propto P(U) = \prod_{u \in U} p(u \mid pa(u))$$

- **Per ottenere la classificazione basta applicare la precedente per tutti i valori della classe**