Review of Probability and Bayesian Networks

Yuqing Tang

Doctoral Program in Computer Science
The Graduate Center
City University of New York
ytang@cs.gc.cuny.edu

October 6, 2010

Outline

Event Space

- Let U be the universe of all possible events
- For any possible event $X, Y \subseteq U$
- Set-theoretical operators
  - $X \cap Y \stackrel{\text{def}}{=} \{z | x \in X \text{ and } z \in Y\}$
  - $X \cup Y \stackrel{\text{def}}{=} \{z | x \in X \text{ or } z \in Y\}$
  - $X \setminus Y \stackrel{\text{def}}{=} \{z | x \in X \text{ but } z \notin Y\}$
  - $\bar{X} = U \setminus X$

Kolmogorov Axioms

1. $P(U) = 1$
2. For any $X \subseteq U$, $P(X) \geq 0$
3. For any two events $X, Y \subseteq U$
   if $X \cap Y = \emptyset$
   then $P(X \cup Y) = P(X) + P(Y)$

Conditional Probability

Definition (Conditional Probability)

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

Bayes' Rule

Definition

Bayes' Rule [Reverend Thomas Bayes (1764)]

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

- Read $P(E|H)$ as the likelihood of the event E given hypothesis H
- Read $P(H)$ as the prior of the hypothesis H
- Read $P(E)$ as the prior of the evidence E
- Read $P(H|E)$ as the posterior belief $Bel(H|E)$ of H given evidence E

Independence

Definition

Event X is said be independent of event Y, denoted by $X \perp\!\!\!\perp Y$,

- iff $P(X|Y) = P(X)$

- An equivalent definition is $P(X \cap Y) = P(X) \cdot P(Y)$
- Independence can be input knowledge
- Independence can arise from the probability

Conditional independence

Definition

Event X is said to be independent of event Y given Z, denoted $X \perp\!\!\!\perp Y|Z$

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

- An equivalent definition is $P(X \cap Y|Z) = P(X|Z) \cdot P(Y|Z)$

Outline

Representing the event space with random variables

Event space U can be represented by a set of random variables and the values assigned to these variables.

- A set of random variables $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$
- A domain $Domain(V_i)$ for each variable V_i
- A variable assignment $\omega : V_i \to Domain(V_i)$ corresponds to a possible world
- All the possible value assignments corresponds to the universe event $U = \{\omega_j\}$
  - Equivalently the universe event space is the cross product $U = Domain(V_1) \times Domain(V_2) \times \ldots \times Domain(V_n)$
  - U is the set of all possible combinations of the values that can be assigned to the variables
  - One more math notation: $U = \Pi_{V_i \in \mathcal{V}} Domain(V_i)$
- An expression $V_{i_1} = v_{i_1} \wedge V_{i_2} = v_{i_2} \wedge \ldots \wedge V_{i_k} = v_{i_k}$ corresponds to an event $X \subseteq U$ such that

$$X = \{\omega \mid \omega \in U \text{ and } \omega(V_{i_1}) = v_{i_1}\}$$

Examples I

- $\mathcal{V} = \{P, S, C\}$ where P for pollution, S for smoking, and C for having cancer
- The corresponding domains are $Domain(P) = \{low, high\}$, $Domain(S) = \{T, F\}$, $Doman(C) = \{T, F\}$
- All possible worlds are

$\langle P, S, C\rangle$
$\langle low, T, T\rangle$
$\langle low, F, T\rangle$
$\langle high, T, T\rangle$
$\langle high, F, T\rangle$
$\langle low, T, F\rangle$
$\langle low, F, F\rangle$
$\langle high, T, F\rangle$
$\langle high, F, F\rangle$

Examples II

- Event $P = low$ corresponds to

$$\{\langle P = low, S = T, C = T\rangle, \langle P = low, S = F, C = T\rangle,$$
$$\langle P = low, S = T, C = F\rangle, \langle P = low, S = F, C = F\rangle\}$$

- Event $S = T$ corresponds to

$$\{\langle P = low, S = T, C = T\rangle, \langle P = high, S = T, C = T\rangle,$$
$$\langle P = low, S = T, C = F\rangle, \langle P = high, S = T, C = F\rangle\}$$

- Event $P = low, S = T$ corresponds to

$$\{\langle P = low, S = T, C = T\rangle, \langle P = low, S = T, C = F\rangle\}$$

Multivariate

- A vector of variables $\mathbf{X} = \langle X_1, \ldots, X_k \rangle$ where $X_1, \ldots, X_k \in \mathcal{V}$
- A vector of values $\mathbf{x} = \langle x_1, \ldots, x_k \rangle$ where $x_i \in Domain(X_k)$ $(1 \leq i \leq k)$
- Multivariate notion of an event: $\mathbf{X} = \mathbf{x}$ means

$$X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_k = x_k$$

  which is usually abbreviated by $\mathbf{x}$ if the variables $\mathbf{X}$ is clear in the context
- $P(\mathbf{X})$ corresponds to a table of probabilities with each assignment $\mathbf{x}$ to $\mathbf{X}$ having an entry in the table
- $P(\mathbf{X}, \mathbf{Y})$ corresponds to a table of probabilities with each assignment $\langle \mathbf{x}, \mathbf{y} \rangle$ to $\langle \mathbf{X}, \mathbf{Y} \rangle$ having an entry in the table

Multivariate version of conditional probability

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

means for every assignment $\langle \mathbf{x}, \mathbf{y} \rangle$ to $\langle \mathbf{X}, \mathbf{Y} \rangle$, we have

$$P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{y})}$$

Multivariate version of conditional independence

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$$

means

$$P(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})$$

means for every assignment $\langle \mathbf{x}, \mathbf{y}, \mathbf{z} \rangle$ to $\langle \mathbf{X}, \mathbf{Y}, \mathbf{Z} \rangle$, we have

$$P(\mathbf{x}|\mathbf{y}, \mathbf{z}) = P(\mathbf{x}|\mathbf{z})$$

Multivariate version of Bayes' Rule

Bayes rule:

$$P(\mathbf{H}|\mathbf{E}) = \frac{P(\mathbf{E}|\mathbf{H}) \cdot P(\mathbf{H})}{P(\mathbf{E})}$$

For every assignment $\langle \mathbf{h}, \mathbf{e} \rangle$ to $\langle \mathbf{H}, \mathbf{E} \rangle$, we have

- If **e** is the only known evidence in the context
  - Read the likelihood of **e** given **h** simply as likelihood of **h**:

$$\lambda(\mathbf{h}) = P(\mathbf{e}|\mathbf{h})$$

  - Read the belief of h given e simply as belief:

$$Bel(\mathbf{h}) = Bel(\mathbf{h}|\mathbf{e}) = P(\mathbf{h}|\mathbf{e})$$

- Bayes' rule can then be read as

$$Posterior \;\; = \frac{Likelihood \times Prior}{Prob\ of\ \ evidence}$$

Marginalization

$$P(\mathbf{X} = \mathbf{x}) = \Sigma_{\mathbf{y} \in Domain(\mathbf{Y})} P(\mathbf{X} = \mathbf{x}, \mathbf{Y})$$

Example

P	S	$P(P,S)$
H	T	0.03
H	F	0.07
L	T	0.27
L	F	0.63

$$\begin{aligned} & P(P = low) \\ &= P(P = low, S = T) \\ &\quad + P(P = low, S = F) \\ &= 0.9 \end{aligned}$$

Multivariate version of chain rule

Each assignment $\langle \mathbf{x_1}, \ldots, \mathbf{x_n} \rangle$ to $\langle \mathbf{X}_1, \ldots, \mathbf{X_n} \rangle$ satisfies

$$
\begin{aligned}
P(\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}) = & \\
& P(\mathbf{X_1}) \\
& \times P(\mathbf{X_2} | \mathbf{X_1}) \\
& \times P(\mathbf{X_3} | \mathbf{X_1}, \mathbf{X_2}) \\
& \times \ldots \times P(\mathbf{X}_n | \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}) \\
= \quad & \Pi_{i=1,\ldots,n} P(\mathbf{X}_i | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1})
\end{aligned}
$$

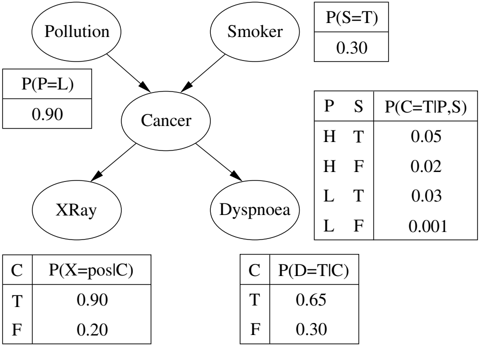
Example

Pollution-Smoking-Cancer

$$P(P, S, C) = P(P) \times P(S|P) \times P(C|P, S)$$

Outline

Bayesian networks



P(S=T)
0.30

P(P=L)
0.90

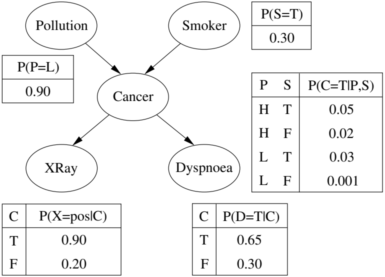
P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

- A Bayesian Network is a directed acyclic graph (DAG)
  - Random variables makes up the nodes
  - Directed links or arrows connects pairs of nodes representing the dependence between variables.
  - Each node has a conditional probability table that quantifies the effects the parents have on the node.
- Gives a concise specification of the joint probability distribution.

Structure terminology and layout



P(P=L)
0.90

P(S=T)
0.30

P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

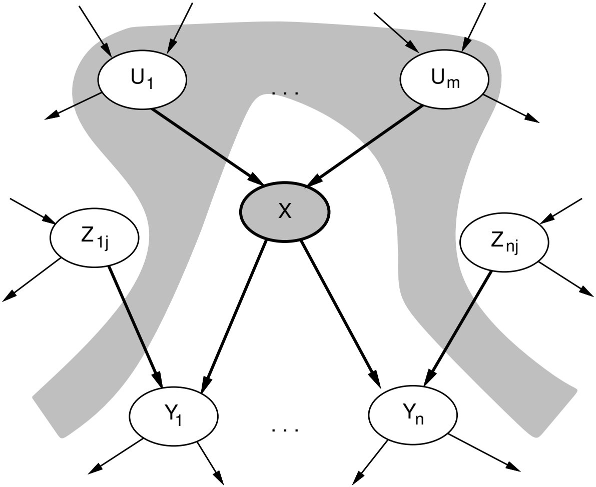
C	P(X=pos|C)
T	0.90
F	0.20

C	P(D=T|C)
T	0.65
F	0.30

- Family metaphor: *Parent* $\Rightarrow$ *Child*
  Ancestor $\Rightarrow \ldots \Rightarrow$ *Descendant*
- Tree analogy:
  - root node: no parents
  - leaf node: no children
  - intermediate node: non-leaf, non-root
- Layout convention: root notes at top, leaf nodes at bottom, arcs point down the page.

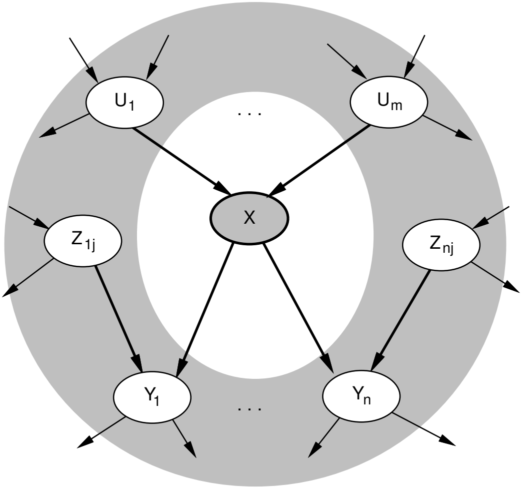
Local semantics

Local semantics: each node is conditionally independent of its nondescendants given its parents.

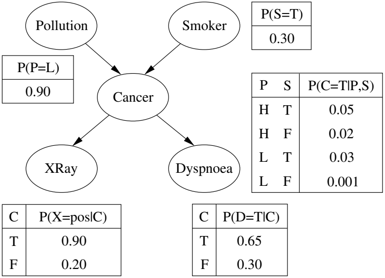


Markov blanket

Each node is conditionally independent of all others given its *Markov blanket*: parents + children + children's parents



Conditional probability tables



- One *conditional probability table (CPT)* for each node.
- Each row contains the conditional probability of every node value for a combination of its parents' values nodes.
- Each row sums to 1.
- A table for a Boolean var with n Boolean parents contain 2^{n+1} probabilities.
- A node with no parents has one row (the prior probabilities)

Bayesian network: A compact representation of joint probabilities

- Bayesian Network implies that the probability of a node is only conditional dependence of its parents

$$\begin{aligned} P(X_i|X_1,\ldots,X_{i-1}) &= P(X_i|Parents(X_i)) \\ P(X_1,\ldots,X_n) &= \Pi_{i=1,\ldots n} P(X_i|Parents(X_i)) \end{aligned}$$

- Bayesian network regulates an ordering of variables
  - $\langle X_1, X_2, \ldots, X_N \rangle$
  - $Parents(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$
- Factoraization of joint probability with Bayesian network

$$\begin{aligned} P(X_1,\ldots,X_n) &= P(X_1) \times \ldots \times P(X_n|X_1,\ldots,X_{n-1}) \\ &= \Pi_{i=1,\ldots n} P(X_i|Parents(X_i)) \end{aligned}$$

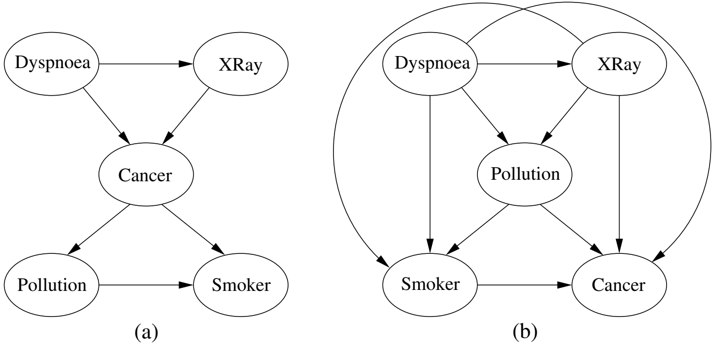
Example

Example

$$
\begin{aligned}
& P(S = F, P = low, C = T, D = T, X = pos) \\
= \;& P(S = F) \\
& \times P(P = low | S = F) \\
& \times P(C = T | P = low, S = F) \\
& \times P(D = T | C = T, P = low, S = F) \\
& \times P(X = pos | D = T, C = T, P = low, S = F) \\
= \;& P(S = F) \\
& \times P(P = low) \\
& \times P(C = T | P = low, S = F) \\
& \times P(D = T | C = T) \\
& \times P(X = pos | C = T)
\end{aligned}
$$

Different node ordering different compactness

- Variable order affect compactness
- Alternative structures using different orderings $(a)\langle D, X, CP, S\rangle$, $(b)\langle D, X, P, S, C\rangle$



- These BNs still represent the same joint distribution.
- Structure (b) requires many probabilities to compute the full joint distribution!

Pearl's network construction algorithm

1. Choose the set of relevant variables $\{X_i\}$ that describe the domain.
2. Choose an ordering for the variables, $\langle X_1, \ldots, X_n \rangle$.
3. While there are variables left:
    1. Add the next variable X_i to the network.
    2. Add arcs to the X_i nodes from some minimal set of nodes already in the net, $Parents(X_i)$, such that the following conditional independence property is satisfied:

       $$P(X_i | X'_1, \ldots, X'_m) = P(X_i | Parents(X_i))$$

       where $X'_1, \ldots, X'_m$ are all the variables preceding X_i, including $Parents(X_i)$.
    3. Define the CPT for X_i

References I