CIS 7414x Expert Systems

Lecture 3 & 4: Probability in AI and Bayesian Networks

Yuqing Tang

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

September 22nd, 2010

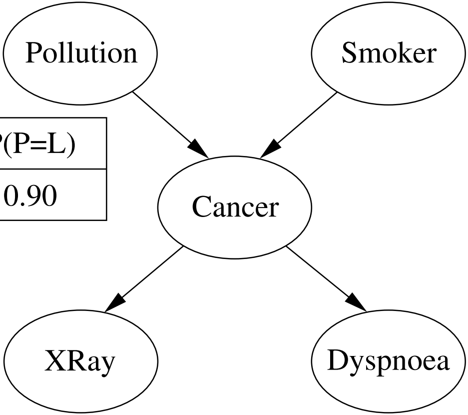
Outline

Topics

- Probability calculus
  - Kolmogorov's axioms
  - Joint probability
  - Conditionals
  - Independence
- Bayes' rules
- Bayesian networks at a glance

Outline

Probabilities and Bayesian network at a glance



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

Probability Calculus

- Classic approach to reasoning under uncertainty. (origin: Blaise Pascal and Fermat).
- Event space
  - Let U be the universe of all possible events
  - For any possible event X, $X \subseteq U$
- Kolmogorov's Axioms — constraints on valid assignments of uncertainty measures on events
  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 $P(X \cup Y) = P(X) + P(Y)$

Example

Example

Event space $U = \{(Pollution = Low), (pollution = high)\}$

- $(Pollution = Low)$ — all the possibilities that the level of pollution is low
- $(Pollution = high)$ — all the possibilities that the level of pollution is high

1. $P(pollution = Low) = 0.9 \geq 0$
2. $P(pollution = High) = 0.1 \geq 0$
3. $P(pollution = Low) + P(pollution = High) = 1$ as
   $(Pollution = Low) \cap (Pollution = High) = \emptyset$
   $(Pollution = Low) \cup (Pollution = High) = U$

Random variables and event space I

- A set of random variables $\mathcal{V}$: e.g. $P \in \mathcal{V}$ for pollution, $S \in \mathcal{V}$ for smoking, $C \in \mathcal{V}$ for having cancer
- A set of values, denoted by $Domain(V_i)$, of variable V_i — the domain of V_i: e.g. $Domain(P) = \{low, high\}$, $Domain(S) = \{T, F\}$, $Doman(C) = \{T, F\}$
- The joint universe event space $U = \Pi_{V_i \in \mathcal{V}} Domain(V_i)$ is the set of all possible combinations of the values can be assigned to the variables: e.g. the joint universe of $\mathcal{V} = \{P, S, C\}$ is

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

Random variables and event space II

- For a value $a \in Domain(V_i)$, the event $V_i = a$ corresponds to the cross product

$$\{V_i = a\} \times \Pi_{V_j \in \mathcal{V} \text{ and} j \neq i} Domain(V_j)$$

e.g. $P = low$ corresponds to

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

$S = T$ corresponds to

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

Random variables and event space III

- For two variables V_i and V_j, the joint event of $V_i = a$ and $V_j = b$, denoted by

$$V_i = a, V_j = b$$

or

$$V_i = a \wedge V_j = b$$

or

$$(V_i = a) \cap (V_j = b)$$

in the event space:
e.g.

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

Bayes' Theorem

Definition (Conditional Probability)

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

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 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

Conditionalization as posterior belief

Bayes rule:

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

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

    $$\lambda(h) = P(e|h)$$

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

    $$Bel(h) = Bel(h|e) = P(h|e)$$

- Bayes' rule can then be read as

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

A Bayes' rule example

Assume we know

$$
\begin{aligned}
P(C = T) &= 0.0116 \\
P(X = pos) &= 0.20812 \\
P(X = pos | C = T) &= 0.9
\end{aligned}
$$

With Bayes' rule, we can compute the following

$$
\begin{aligned}
P(C = T | X = pos) &= \frac{P(X = pos | C = T) \times P(C = T)}{P(X = pos)} \\
&= \frac{0.9 \times 0.0116}{0.2081} \\
&= 0.050
\end{aligned}
$$

A Bayes' rule example (cont.)

If we know $P(X = pos|C = F) = 0.2$, we don't need to know $P(X = pos) = 0.20812$. With the Bayes' rule, we can compute

$$
\begin{aligned}
P(C = T|X = pos) &= \frac{P(X = pos|C = T) \times P(C = T)}{P(X = pos)} \\
&= \frac{0.9 \times 0.0116}{P(X = pos)} \\
P(C = F|X = pos) &= \frac{P(X = pos|C = F) \cdot P(C = F)}{P(X = pos)} \\
&= \frac{0.2 \times 0.9884}{P(X = pos)}
\end{aligned}
$$

By $P(C = T|X = pos) + P(C = F|X = Pos) = 1$, we can solve the above three equations and obtain $P(X = pos) = 0.01044 + 0.19768 = 0.20812$. Put it back to the first equation, we will have

$$P(C = T|X = pos) = \frac{0.01044}{0.20812} = 0.050$$

Independence and conditional independence

- Independence $X \perp\!\!\!\perp Y$ iff $P(X|Y) = P(X)$ iff $P(X \cap Y) = P(X) \cdot P(Y)$
  - Independence can be input knowledge — $P(X \cap Y) = P(X) \cdot P(Y)$ is a constraint arising from the problem domain in hands
  - Independence can arise from the probability analysis of the joint probability
- Conditional independence $X \perp\!\!\!\perp Y|Z$ iff $P(X|Y,Z) = P(X|Z)$ iff $P(X \cap Y|Z) = P(X|Z) \cdot P(Y|Z)$

Marginalization

From Kolmogorov's second axiom: if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$, we have

$$P(X = a) = \Sigma_{y_i \in Domain(Y)} P(X = a, Y = y_i)$$

- $P(X, Y)$ is a joint distribution
- The summation is over all possible values of $Y = \{y_i\}$
- For any two values y_i and y_j $(i \neq j)$ of Y, the $(X = a, Y = y_i) \cap (X = a, Y = y_j) = \emptyset$
- X and Y can be generalized into vectors, i.e. multivariate variables.

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

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

Chain rule: From conditional probabilities to joint probability

$P(X|Y) = \frac{P(X\cap Y)}{P(Y)}$ implies $P(X \cap Y) = P(X|Y)P(Y)$
We can generalize this into a chain rule:

$$\begin{aligned} P(x_1, x_2, \ldots, x_n) &= P(x_1) \times P(x_2|x_1) \times P(x_3|x_1, x_2) \\ &\quad \times \ldots \times P(x_n|x_1, \ldots, x_{n-1}) \\ &= \Pi_{i=1,\ldots,n} P(x_i|x_1, x_2, \ldots, x_{i-1}) \end{aligned}$$

Example

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

Chain rule: From conditional probabilities to joint probability (cont.)

Example

If $C \perp\!\!\!\perp A|B$, then $P(C|A,B) = P(C|B)$, the chain can be simplified

$$P(A,B,C) \;=\; P(A) \times P(B|A) \times P(C|B)$$

Bayesian networks are about representing various kinds of independence between variables so that

- the joint probability can be compactly represented, and
- efficient algorithms can be devised to repeatedly apply the Bayes' rules on inferring about the posterior beliefs out of any new evidences.

Bayesian Decision Theory

- Frank Ramsey (1926)
  Decision making under uncertainty: what action to take (plan to adopt) when future state of the world is not known.

Bayesian answer: Find utility of each possible outcome (action-state pair) and take the action that maximizes expected utility.

Example

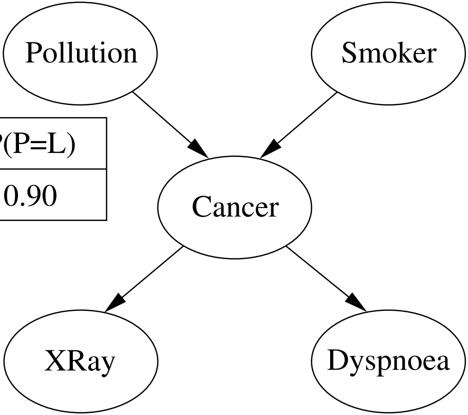
action	Rain ($p = 0.4$)	Shine ($1 - p = 0.6$)
Take umbrella	30	10
Leave umbrella	-100	50

Expected utilities:

$$\begin{aligned} E(\mathit{Take\ umbrella}) &= (30)(0.4) + (10)(0.6) = 18 \\ E(\mathit{Leave\ umbrella}) &= (-100)(0.4) + (50)(0.6) = -10 \end{aligned}$$

Outline

Probabilities and Bayesian network at a glance



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

Bayesian Networks

- Data Structure which represents the dependence between variables.
- Gives concise specification of the joint probability distribution.
- A Bayesian Network is a graph in which the following holds:
  - A set of random variables makes up the nodes in the network.
  - A set of directed links or arrows connects pairs of nodes.
  - Each node has a conditional probability table that quantifies the effects the parents have on the node.
  - Directed, acyclic graph (DAG), i.e. no directed cycles.

Nodes and values

- Nodes can be discrete or continuous; will focus on discrete for now.
- Boolean nodes: represent propositions, taking binary values true (T) and false (F).
  Example: Cancer node represents proposition "the patient has cancer".
- Ordered values.
  Example: Pollution node with values $\{low, medium, high\}$.
- Integral values. Example: Age node with possible values from 1 to 120.

Lung cancer example: nodes and values

Node name	Type	Values
Pollution	Binary	$\{low, high\}$
Smoker	Boolean	$\{T, F\}$
Cancer	Boolean	$\{T, F\}$
Dyspnoea	Boolean	$\{T, F\}$
X-ray	Binary	$\{pos, neg\}$

Structure terminology and layout

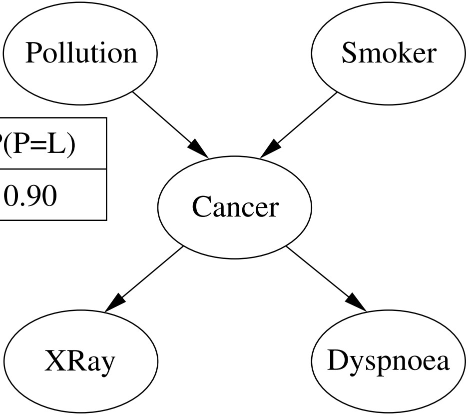
- Family metaphor: *Parent* $\Rightarrow$ *Child*
  Ancestor $\Rightarrow \ldots \Rightarrow$ *Descendant*
- Markov Blanket = parents + children + children's parents
- Tree analogy:
  - root node: no parents
  - leaf node: no children
  - intermediate node: non-leaf, non-root
- Layout convention: root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

Conditional Probability Tables

Once specified topology, need to specify *conditional probability table (CPT)* for each node.

- Each row contains the conditional probability of each node value for a each possible combination of values of its parent nodes.
- Each row must sum 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)

Lung cancer example: CPTs



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

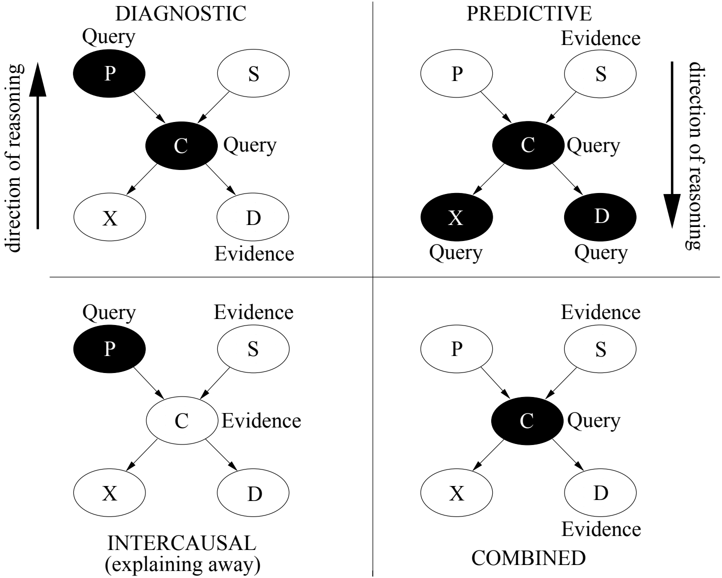
The Markov Property

- Modeling with BNs requires the assumption of the *Markov Property*: there are no direct dependencies in the system being modeled which are not already explicitly shown via arcs.
- Example: there is no way for smoking to influence dyspnoea except by way of causing cancer.
- BNs which have the Markov propertly are called Independence-Maps (I-Maps).
- Note: existence of arc does not have to correspond to real dependency in the system being modeled — can be nullified in the CPT.

Reasoning with Bayesian Networks

- Basic task for any probabilistic inference system:
  Compute the posterior probability distribution for a set of query variables, given new information about some evidence variables.
- Also called conditioning or belief updating or inference. Bayesian

Types of reasoning



Types of evidence

- Specific evidence: a definite finding that a node X has a particular value, x.
  Example: $Smoker = T$
- Negative evidence: a finding that node Y is not in state y_1 (but may take any other values).
- "Virtual" or "likelihood" evidence: source of information is not sure about it.
  Example:
  - $e =$ Radiologist is 80% sure that $Xray = pos$
  - Want e.g.:

$$
\begin{aligned}
P(Cancer|e) &= P(Cancer|Xray, e) \cdot P(Xray|e) + \\
&\quad P(Cancer|\neg Xray, e) \cdot P(\neg Xray|e)
\end{aligned}
$$

  - Jeffrey Conditionalization (will introduced later when it is encountered)

Reasoning with numbers

See a demo.

Understanding of Bayesian Networks (Semantics)

- A (more compact) representation of the joint probability distribution.
  - helpful in understanding how to construct network
- Encoding a collection of conditional independence statements.
  - helpful in understanding how to design inference procedures
  - via Markov property/I-map: Each conditional independence implied by the graph is present in the probability distribution

Bayesian Network (Conditional Independence), Chain Rule, and Joint Distribution I

- Write $P(X_1 = x_1, \ldots, X_n = x_n)$ as $P(x_1, \ldots, x_n)$.
- Factoraization (chain rule):

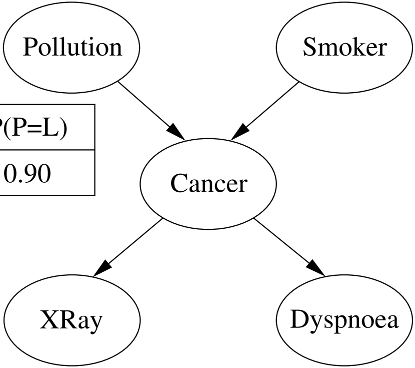
$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_1) \times \ldots \times P(x_n | x_1, \ldots, x_{n-1}) \\ &= \Pi_{i=1,\ldots n} P(x_i | x_1, \ldots, x_{i-1}) \end{aligned}$$

- Bayesian Network implies that the value of particular node is only conditional dependence of its parent nodes

$$\begin{aligned} P(x_i | x_1, \ldots, x_{i-1}) &= P(x_i | Parents(X_i)) \\ P(x_1, \ldots, x_n) &= \Pi_{i=1,\ldots n} P(x_i | Parents(X_i)) \end{aligned}$$

- In the above, we need an ordering of variables: $Parents(X_i) \subseteq \{X_1, \ldots, X_{n-1}\}$

Bayesian Network (Conditional Independence), Chain Rule, and Joint Distribution II



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

Bayesian Network (Conditional Independence), Chain Rule, and Joint Distribution III

Example

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

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:

    $$P(X_i | X_1', \ldots, X_m') = P(X_i | Parents(X_i))$$

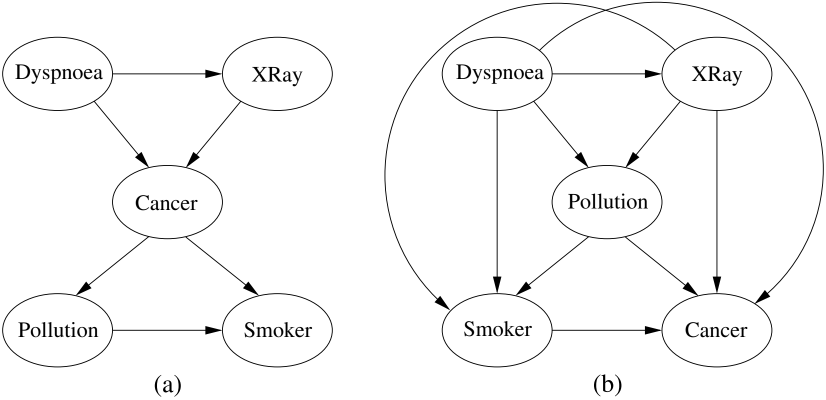
       where $X_1', \ldots, X_m'$ are all the variables preceding X_i, including $Parents(X_i)$.
    3. Define the CPT for X_i

Compactness and node ordering

- Compactness of BN depends upon sparseness of the systems
- The best order to add nodes is to add the "root causes" first, then the variable they influence, so on until "leaves" reached.
  - Causal structure

Different node ordering different compactness

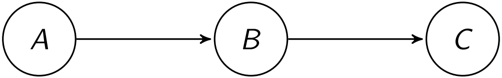
- Variable order affect compactness
- Alternative structures using different orderings $(a)\langle D, X, CP, S\rangle$, $(b)\langle D, X, P, S, C\rangle$



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

Condtional Independence in Causal Chains

Causal chians give rise to conditioanl independence: $A \perp\!\!\!\perp C | B$



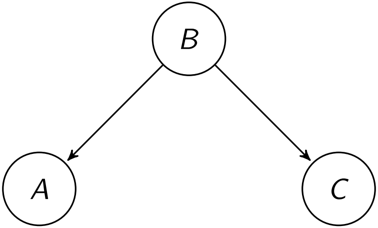
$$P(C|A, B) = P(C|B)$$

Example: "smoking causes cancer which causes dyspnoea".

Condtional Independence in Common Causes

Common causes (or ancestors) give rise to conditioanl independence:

$$A \perp\!\!\!\perp C | B$$

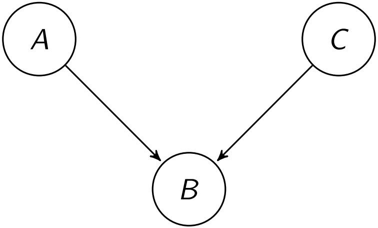


$$P(C|A, B) = P(C|B)$$

Example: "Cancer is common cause of the two symptons, a positive XRay resutl and dyspnoea."

Condtional Dependence in Common Effects

Causual effects (or their descendants) give rise to conditioanl independence: $\neg(A \perp\!\!\!\perp C|B)$



$$P(C|A,B) \neq P(C|B)$$

altough marginal dependence

$$P(A,C) = P(A) \cdot P(C)$$

Example: "Cancer is a common effect of pollution and smoking."

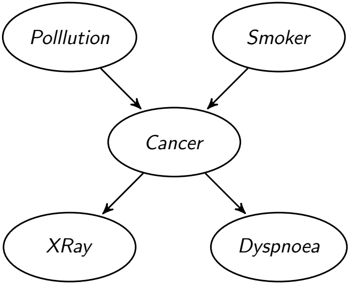
Given lung cancer, smoking "explains away" pollution.

Direction-dependent Speration: D-Speration

- Graphical criterion of conditional independence. X and Y are *d-seperated* by Z:

$$X \perp Y | Z$$

- We can determine whether a set of nodes X is independent of another set Y, given a set of evidence nodes E, via the Markov propoerty: $X \perp Y | E \rightarrow X \perp\!\!\!\perp Y | E$.
- Example



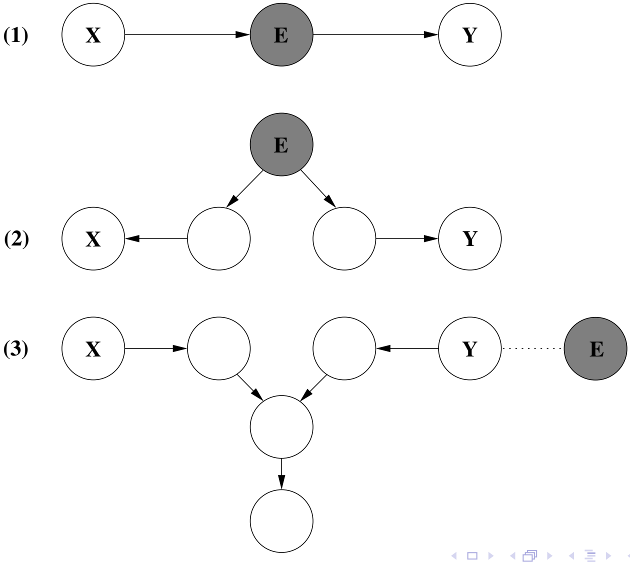
Determine D-Seperation

How to determine $X \perp Y | E$:

- If every undirected path from a node in X to a node in Y is d-separated by E, then X and Y are conditionally independent given E.
- A set of nodes E d-separates two sets of nodes X and Y if every undirected path from a node in X to a node in Y is blocked given E.
- A path is blocked given a set of nodes E if there is a node Z on the path for which one of three conditions holds:
  1. Z is in E and Z has one arrow on the path leading in and one arrow out (chain), or
  2. Z is in E and Z has both path arrows leading out (common cause), or
  3. Neither Z nor any descendant of Z is in E, and both path arrows lead in to Z (common effect).

D-Seperation

Evidence nodes **E** shown shaded.

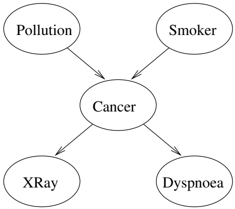


Causual Ordering

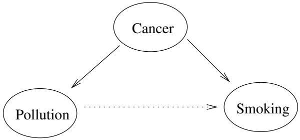
Why does variable order affect network density?

- Using the causal order allows direct representation of conditional independencies
- Violating causal order requires new arcs to re-establish conditional independencies

Causual Ordering (cont.)



Pollution and Smoking are marginally indepdendent.
Ordering: Cancer, Pollution, Smoking:



Marginal independence of Pollution and Smoking must be re-established by adding *Pllution* $\rightarrow$ *Smoking* or *Smoking* $\leftarrow$ *Pollution*.

Summary of Bayesian Networks

- Bayes' rule allows unknown probabilities to be computed from known ones.
- Conditional independence (due to causal relationships) allows efficient updating
- BNs are a natural way to represent conditional independence info.
  - links between nodes: qualitative aspects;
  - conditional probability tables: quantitative aspects.
- Probabilistic inference: compute the probability distribution for query variables, given evidence variables
- BN Inference is very flexible: can enter evidence about any node and update beliefs in any other nodes.

Outline

Justifications of probability I

- *The principle of indifference* — all elementary outcomes are equally likely
  - In the absence of any other information, there is no reason to consider one more likely than another
  - Application in handling statistical information: e.g. 40 percent of a doctor's are over 60, $P(PatientAge > 60) = 0.4$
  - Problem: Different choices of elementary outcomes lead to different probability assignments of the same situation
- Frequentism — the probability numbers represent relative frequencies
  - e.g. a coin lands heads with $1/2$ of the outcomes if it is tossed "sufficiently often"
  - Problem: How many times is "sufficiently often"? How about something can not be repeated? How about it is costly to repeat?
- Subjective view — the probability numbers reflect subjective assessments of likelihood as long as the numbers satisfied the Kolmogorov's axioms; a famous argument is of Ramsey's:

Justifications of probability II

- Probability is justified in terms of bet, denoted by $\langle X, \alpha \rangle$ $(0 \leq \alpha \leq 1)$: If event $X \subseteq U$ happens, the agent wins $100(1-\alpha)$ dollars otherwise it loses 100α; the complementary bet is $\langle \neg X, 1-\alpha \rangle$
- If an agent bets according to a set of rational criteria, then the probability measure of the event X is α_X such that
  - ★ $\langle X, \alpha \rangle$ is preferred to $\langle \neg X, 1-\alpha \rangle$ by the agent for all $\alpha < \alpha_X$, and
  - ★ $\langle \neg X, 1-\alpha \rangle$ is preferred to $\langle X, \alpha \rangle$ by the agent for all $\alpha > \alpha_X$

Taken from [Halpern, 2003]

A big picture

- Kolmogorov's axioms
- Conditional probability
- Belief as conditional probability
- Joint probability
- Marginal probability
- Belief update as posterior conditionalization via marginalization of joint probability and/or the application of Bayes' rules
- Joint probability computation and belief update can be simplified by employing the conditional independence
- Bayesian networks is the structural way to achieve this simplification
  - Graphical representation of independence and conditional independence
  - Factoring the computation of unknown conditional probabilities (the unknown post-evidence beliefs) into
    - Traversing the nodes and edges in the network, and
    - Carrying out simpler computation steps associated with the nodes and edges

Acknowledgments

Lecture 3 is composed the instructor's own understanding of the subject, and materials from [Korb and Nicholson, 2003, Chapter 1, Chapter 2] with the instructor's own interpretations. The instructor takes full responsibility of any mistakes in the slides.

References I

Joseph Y. Halpern.
Reasoning about Uncertainty.
MIT press, Cambridge, MA, 2003.

K. Korb and A. E. Nicholson.
Bayesian Artificial Intelligence.
Chapman & Hall /CRC, 2003.