Probabilistic Reasoning; Network-based reasoning

COMPSCI 276, Spring 2017

Set 1: Introduction and Background

Rina Dechter

(Reading: Pearl chapter 1-2, Darwiche chapters 1,3)

Example of Common Sense Reasoning

- **Zebra on Pajama**: (7:30 pm): I told Susannah: you have a nice pajama, but it was just a dress. Why jump to that conclusion?: 1. because time is night time. 2. certain designs look like pajama.
- **Cars going out of a parking lot:** You enter a parking lot which is quite full (UCI), you see a car coming : you think ah… now there is a space (vacated), OR… there is no space and this guy is looking and leaving to another parking lot. What other clues can we have?
- **Robot gets out at a wrong level:** A robot goes down the elevator. stops at 2nd floor instead of ground floor. It steps out and should immediately recognize not being in the right level, and go back inside.
- **Turing quotes**
  - If machines will not be allowed to be fallible they cannot be intelligent
  - (Mathematicians are wrong from time to time so a machine should also be allowed)

Why/What/How Uncertainty?

- Why Uncertainty?
  - Answer: It is abundant
- What formalism to use?
  - Answer: Probability theory
- How to overcome exponential representation?
  - Answer: Graphs, graphs, graphs… to capture irrelevance, independence

Class Description

- Instructor: Rina Dechter

- Days: Monday & Wednesday
- Time: 11:00 - 12:20 pm
- Class page:
- http://www.ics.uci.edu/~dechter/courses/ics-275b/spring-17/

Outline

- Why/What/How… uncertainty?
- Basics of probability theory and modeling

Outline

- Why/What/How uncertainty?
- Basics of probability theory and modeling

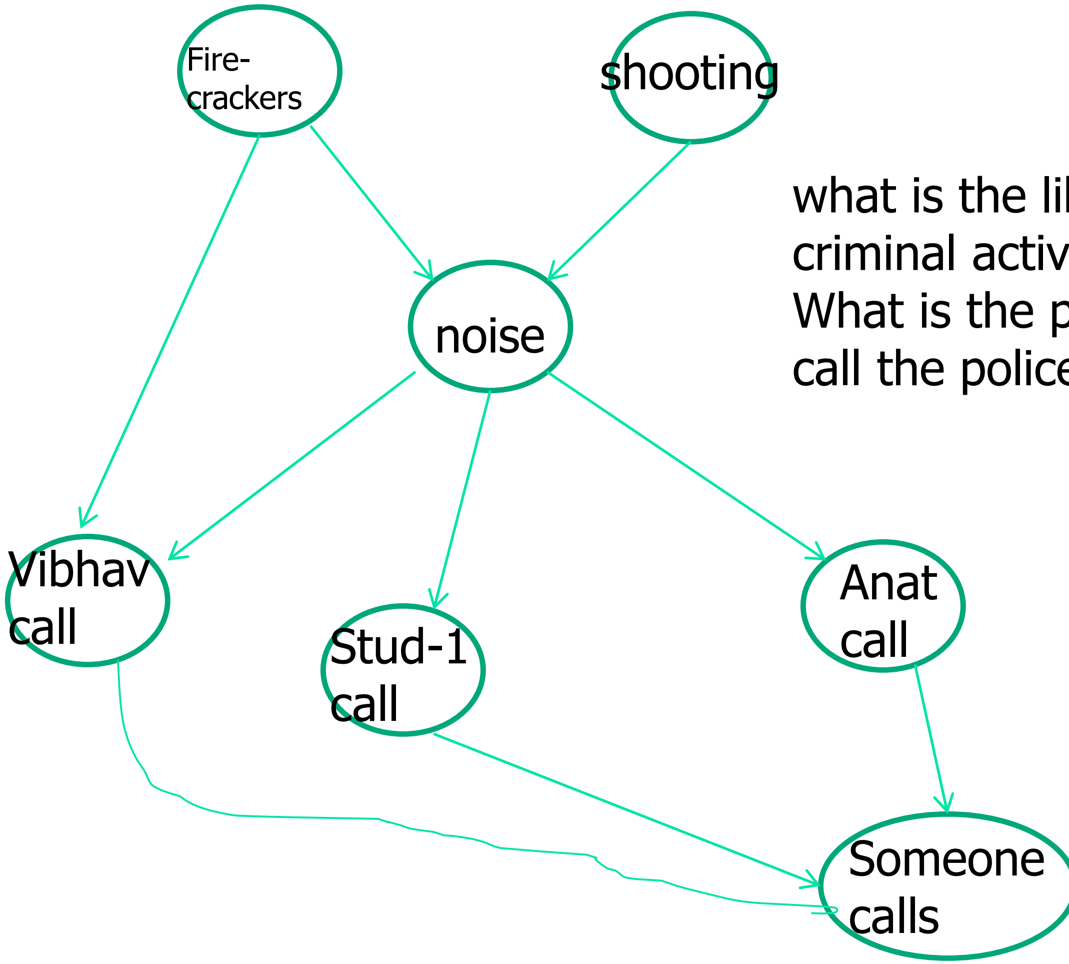
Why Uncertainty?

- AI goal: to have a declarative, model-based, framework that allows computer system to reason.
- People reason with partial information
- Sources of uncertainty:
  - Limitation in observing the world: e.g., a physician see symptoms and not exactly what goes in the body when he performs diagnosis. Observations are noisy (test results are inaccurate)
  - Limitation in modeling the world,
  - maybe the world is not deterministic.

Example of Common Sense Reasoning

- Explosive noise at UCI
- Parking in Cambridge
- The missing garage door
- Years to finish an undergrad degree in college
- The Ebola case
- Lots of abductive reasoning

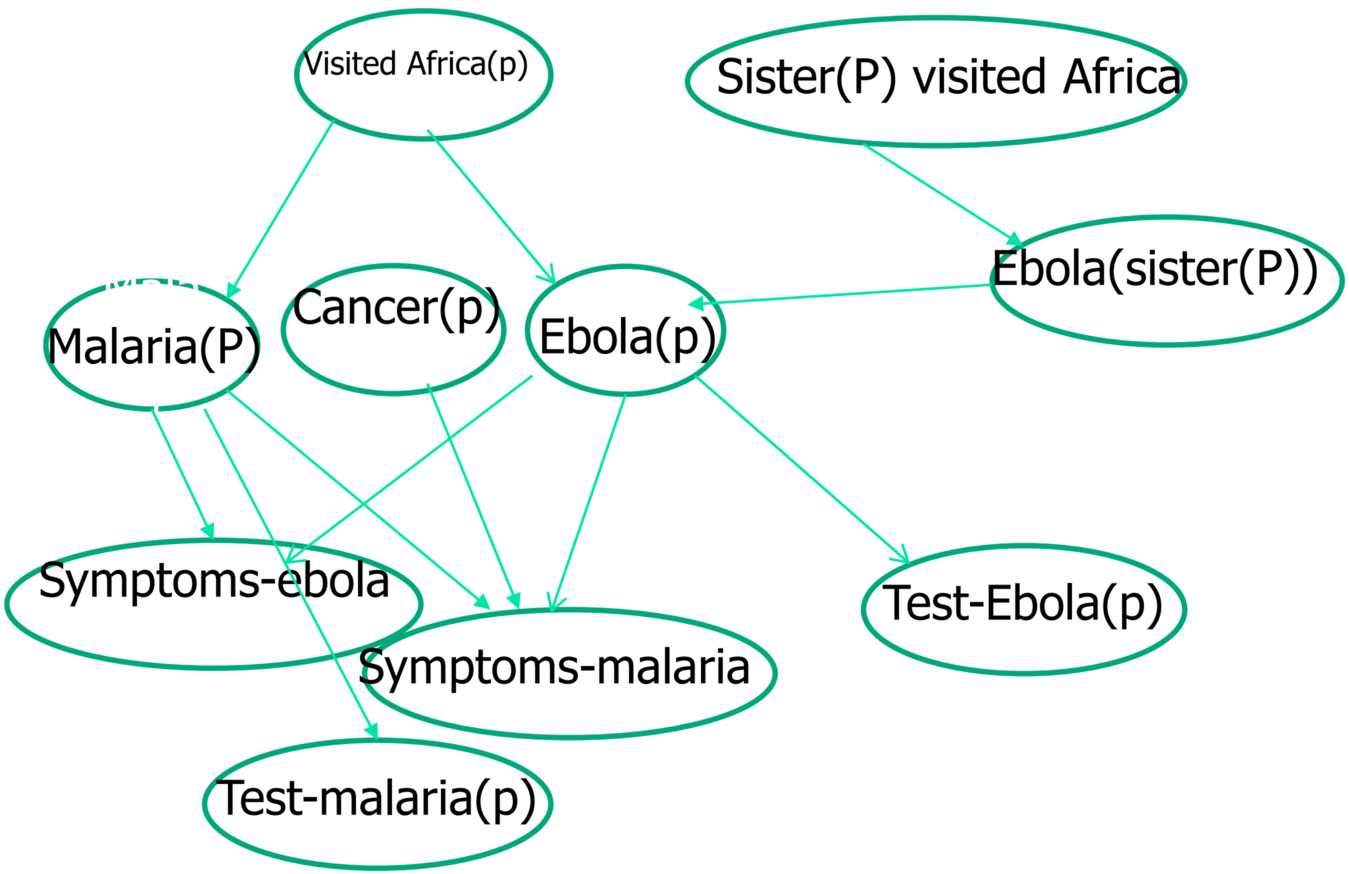
Shooting at UCI



what is the likelihood that there was a criminal activity if S1 called?
What is the probability that someone will call the police?

Ebola in the US

What is the likelihood that P has Ebola if he came from Africa?
If his sister came from Africa?
What is the probability P was in Africa given that he tested positive for Ebola?



Why Uncertainty

- **Summary of exceptions**
  - Birds fly, smoke means fire (cannot enumerate all exceptions.
- **Why is it difficult?**
  - Exception combines in intricate ways
  - e.g., we cannot tell from formulas how exceptions to rules interact:

A→C
B→C

A and B -→ C

Commonsense Reasoning(*)

Example: My car is still parked where I left it this morning. If I turn the key of my car, the engine will turn on. If I start driving now, I will get home in thirty minutes.

- None of these statements is factual as each is qualied by a set of assumptions. We tend to make these assumptions, use them to derive certain conclusions (e.g., I will arrive home in thirty minutes if I head out of the ofice now), and then use these conclusions to justify some of our decisions (I will head home now).
- We stand ready to retract any of these assumptions if we observe something to the contrary (e.g., a major accident on the road home).

The Problem

All men are mortal	T
All penguins are birds	T
…	
Socrates is a man	
Men are kind	p1
Birds fly	p2
T looks like a penguin	
Turn key –> car starts	P_n

True propositions (rows 1–4)

Uncertain propositions (rows 5–8)

Q: Does T fly? Logic?....but how we handle exceptions

P(Q)? Probability: astronomical

Managing Uncertainty

- Knowledge obtained from people is almost always loaded with uncertainty
- Most rules have exceptions which one cannot afford to enumerate
- Antecedent conditions are ambiguously defined or hard to satisfy precisely
- First-generation expert systems combined uncertainties according to simple and uniform principle
- Lead to unpredictable and counterintuitive results
- Early days: logicist, new-calculist, neo-probabilist

The Limits of Modularity

Deductive reasoning: modularity and detachment

$P \rightarrow Q$
P

Q

$P \rightarrow Q$
K and P

Q

$P \rightarrow Q$
$K \rightarrow P$
K

Q

Plausible Reasoning: violation of locality

Wet → rain
Wet
----
rain

wet → rain
Sprinkler and wet
----
rain?

Violation of Detachment

Deductive reasoning

P → Q
K→ P
K

Q

Plausible reasoning

Wet → rain
Sprinkler →wet
Sprinkler
----
rain?

Probabilistic Modeling with Joint Distributions

- All frameworks for reasoning with uncertainty today are “intentional” model-based. All are based on the probability theory implying calculus and semantics.

Outline

- Why uncertainty?
- Basics of probability theory and modeling

Degrees of Belief

- Assign a degree of belief or probability in $[0, 1]$ to each world ω and denote it by $\Pr(\omega)$.
- The belief in, or probability of, a sentence α:

$$\Pr(\alpha) \stackrel{def}{=} \sum_{\omega \models \alpha} \Pr(\omega).$$

world	Earthquake	Burglary	Alarm	$\Pr(.)$
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

Properties of Beliefs

- A bound on the belief in any sentence:

$$0 \leq \Pr(\alpha) \leq 1 \qquad \text{for any sentence } \alpha.$$

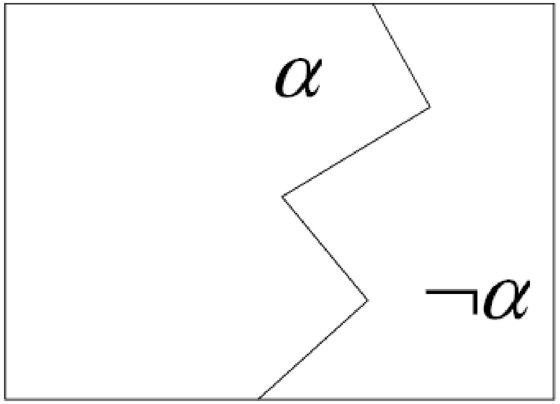
- A baseline for inconsistent sentences:

$$\Pr(\alpha) = 0 \qquad \text{when } \alpha \text{ is inconsistent.}$$

- A baseline for valid sentences:

$$\Pr(\alpha) = 1 \qquad \text{when } \alpha \text{ is valid.}$$

Properties of Beliefs



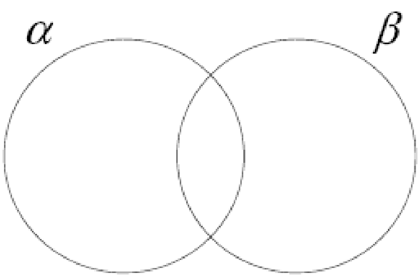
- The belief in a sentence given the belief in its negation:

$$\Pr(\alpha) + \Pr(\neg\alpha) = 1.$$

Example

$$\begin{aligned}
\Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\
\Pr(\neg\text{Burglary}) &= \Pr(\omega_3) + \Pr(\omega_4) + \Pr(\omega_7) + \Pr(\omega_8) = .8
\end{aligned}$$

Properties of Beliefs



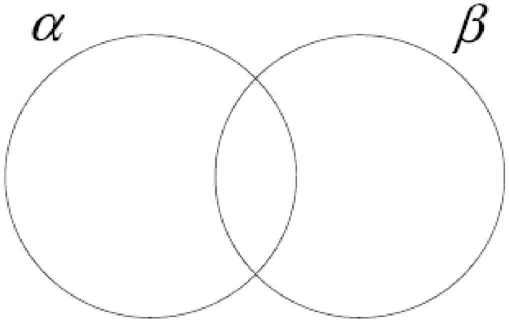
- The belief in a disjunction:

$$\Pr(\alpha \vee \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha \wedge \beta).$$

- Example:

$$\begin{aligned}
\Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_5) + \Pr(\omega_6) = .2 \\
\Pr(\text{Earthquake} \wedge \text{Burglary}) &= \Pr(\omega_1) + \Pr(\omega_2) = .02 \\
\Pr(\text{Earthquake} \vee \text{Burglary}) &= .1 + .2 - .02 = .28
\end{aligned}$$

Properties of Beliefs



- The belief in a disjunction:

  $\Pr(\alpha \vee \beta) = \Pr(\alpha) + \Pr(\beta)$ when α and β are mutually exclusive.

Entropy

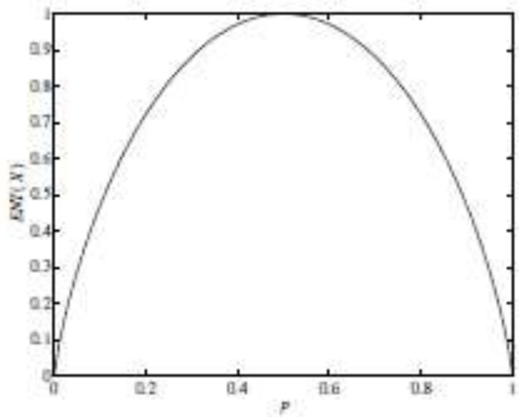
Quantify uncertainty about a variable X using the notion of entropy:

$$\mathrm{ENT}(X) \stackrel{def}{=} -\sum_{x} \Pr(x) \log_2 \Pr(x),$$

where $0 \log 0 = 0$ by convention.

	Earthquake	Burglary	Alarm
true	.1	.2	.2442
false	.9	.8	.7558
$\mathrm{ENT}(.)$	.469	.722	.802

Entropy



- The entropy for a binary variable X and varying $p = \Pr(X)$.
- Entropy is non-negative.
- When $p = 0$ or $p = 1$, the entropy of X is zero and at a minimum, indicating no uncertainty about the value of X.
- When $p = \frac{1}{2}$, we have $\Pr(X) = \Pr(\neg X)$ and the entropy is at a maximum (indicating complete uncertainty).

Bayes Conditioning

Alpha and beta are events

Closed form for Bayes conditioning:

$$\Pr(\alpha|\beta) = \frac{\Pr(\alpha \wedge \beta)}{\Pr(\beta)}.$$

Defined only when $\Pr(\beta) \neq 0$.

Degrees of Belief

world	Earthquake	Burglary	Alarm	$\Pr(.)$
ω_1	true	true	true	.0190
ω_2	true	true	false	.0010
ω_3	true	false	true	.0560
ω_4	true	false	false	.0240
ω_5	false	true	true	.1620
ω_6	false	true	false	.0180
ω_7	false	false	true	.0072
ω_8	false	false	false	.7128

$$\begin{aligned}
\Pr(\text{Earthquake}) &= \Pr(\omega_1) + \Pr(\omega_2) + \Pr(\omega_3) + \Pr(\omega_4) = .1 \\
\Pr(\text{Burglary}) &= .2 \\
\Pr(\neg\text{Burglary}) &= .8 \\
\Pr(\text{Alarm}) &= .2442
\end{aligned}$$

Belief Change

Burglary is independent of Earthquake

Conditioning on evidence Earthquake:

$$
\begin{aligned}
\Pr(\text{Burglary}) &= .2 \\
\Pr(\text{Burglary}|\text{Earthquake}) &= .2 \\
\\
\Pr(\text{Alarm}) &= .2442 \\
\Pr(\text{Alarm}|\text{Earthquake}) &\approx .75 \uparrow
\end{aligned}
$$

The belief in Burglary is not changed, but the belief in Alarm increases.

Belief Change

Earthquake is independent of burglary

Conditioning on evidence Burglary:

$$
\begin{aligned}
\Pr(\text{Alarm}) &= .2442 \\
\Pr(\text{Alarm}|\text{Burglary}) &\approx .905 \uparrow \\
\\
\Pr(\text{Earthquake}) &= .1 \\
\Pr(\text{Earthquake}|\text{Burglary}) &= .1
\end{aligned}
$$

The belief in Alarm increases in this case, but the belief in Earthquake stays the same.

Belief Change

The belief in Burglary increases when accepting the evidence Alarm. How would such a belief change further upon obtaining more evidence?

- Confirming that an Earthquake took place:

$$\begin{aligned} \Pr(\text{Burglary}|\text{Alarm}) &\approx .741 \\ \Pr(\text{Burglary}|\text{Alarm} \wedge \text{Earthquake}) &\approx .253 \downarrow \end{aligned}$$

  We now have an explanation of Alarm.

- Confirming that there was no Earthquake:

$$\begin{aligned} \Pr(\text{Burglary}|\text{Alarm}) &\approx .741 \\ \Pr(\text{Burglary}|\text{Alarm} \wedge \neg\text{Earthquake}) &\approx .957 \uparrow \end{aligned}$$

  New evidence will further establish burglary as an explanation.

Conditional Independence

$\Pr$ finds α conditionally independent of β given γ iff

$$\Pr(\alpha|\beta \wedge \gamma) = \Pr(\alpha|\gamma) \quad \text{or } \Pr(\beta \wedge \gamma) = 0.$$

Another definition

$$\Pr(\alpha \wedge \beta|\gamma) = \Pr(\alpha|\gamma)\Pr(\beta|\gamma) \quad \text{or } \Pr(\gamma) = 0.$$

Variable Independence

$\Pr$ finds $\mathbf{X}$ independent of $\mathbf{Y}$ given $\mathbf{Z}$, denoted $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, means that $\Pr$ finds $\mathbf{x}$ independent of $\mathbf{y}$ given $\mathbf{z}$ for all instantiations $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

Example

$\mathbf{X} = \{A, B\}$, $\mathbf{Y} = \{C\}$ and $\mathbf{Z} = \{D, E\}$, where A, B, C, D and E are all propositional variables. The statement $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is then a compact notation for a number of statements about independence:

$A \wedge B$ is independent of C given $D \wedge E$;

$A \wedge \neg B$ is independent of C given $D \wedge E$;

$\vdots$

$\neg A \wedge \neg B$ is independent of $\neg C$ given $\neg D \wedge \neg E$;

That is, $I_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ is a compact notation for $4 \times 2 \times 4 = 32$ independence statements of the above form.

Conditional Entropy

To quantify the average uncertainty about the value of X after observing the value of Y.

Conditional entropy of a variable X given another variable Y

$$\mathrm{ENT}(X|Y) \stackrel{def}{=} \sum_{y} \Pr(y)\mathrm{ENT}(X|y),$$

where

$$\mathrm{ENT}(X|y) \stackrel{def}{=} -\sum_{x} \Pr(x|y) \log_2 \Pr(x|y).$$

- Entropy never increases after conditioning:

$$\mathrm{ENT}(X|Y) \leq \mathrm{ENT}(X).$$

- Observing the value of Y reduces our uncertainty about X.
- For a particular value y, we may have $\mathrm{ENT}(X|y) > \mathrm{ENT}(X)$.

Conditional Entropy

	Burglary	Burglary$|$Alarm$=$true	Burglary$|$Alarm$=$false
true	.2	.741	.025
false	.8	.259	.975
ENT(.)	.722	.825	.169

The conditional entropy of Burglary given Alarm is then:

$$\begin{aligned} &\text{ENT}(\text{Burglary}|\text{Alarm}) \\ &= \text{ENT}(\text{Burglary}|\text{Alarm}=\text{true})\Pr(\text{Alarm}=\text{true}) + \\ &\quad \text{ENT}(\text{Burglary}|\text{Alarm}=\text{false})\Pr(\text{Alarm}=\text{false}) \\ &= .329, \end{aligned}$$

indicating a decrease in the uncertainty about variable Burglary.

Further Properties of Beliefs

Chain rule

$$\begin{aligned} & \Pr(\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n) \\ = \; & \Pr(\alpha_1 | \alpha_2 \wedge \ldots \wedge \alpha_n) \Pr(\alpha_2 | \alpha_3 \wedge \ldots \wedge \alpha_n) \ldots \Pr(\alpha_n). \end{aligned}$$

Case analysis (law of total probability)

$$\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha \wedge \beta_i),$$

where the events $\beta_1, \ldots, \beta_n$ are mutually exclusive and exhaustive.

Further Properties of Beliefs

Another version of case analysis

$$\Pr(\alpha) = \sum_{i=1}^{n} \Pr(\alpha|\beta_i)\Pr(\beta_i),$$

where the events $\beta_1, \ldots, \beta_n$ are mutually exclusive and exhaustive.

Two simple and useful forms of case analysis are these:

$$\begin{aligned}\Pr(\alpha) &= \Pr(\alpha \wedge \beta) + \Pr(\alpha \wedge \neg\beta) \\ \Pr(\alpha) &= \Pr(\alpha|\beta)\Pr(\beta) + \Pr(\alpha|\neg\beta)\Pr(\neg\beta).\end{aligned}$$

The main value of case analysis is that, in many situations, computing our beliefs in the cases is easier than computing our beliefs in α. We shall see many examples of this phenomena in later chapters.

Further Properties of Beliefs

Bayes rule

$$\Pr(\alpha|\beta) = \frac{\Pr(\beta|\alpha)\Pr(\alpha)}{\Pr(\beta)}.$$

- Classical usage: α is perceived to be a cause of β.
- Example: α is a disease and β is a symptom–
- Assess our belief in the cause given the effect.
- Belief in an effect given its cause, $\Pr(\beta|\alpha)$, is usually more readily available than the belief in a cause given one of its effects, $\Pr(\alpha|\beta)$.

Difficulty: Complexity in model construction and inference

- In Alarm example:
  - 31 numbers needed,
  - Quite unnatural to assess: e.g.

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

  - Computing $P(B{=}y|M{=}y)$ takes 29 additions.
- In general,
  - $P(X_1, X_2, \ldots, X_n)$ needs at least $2^n - 1$ numbers to specify the joint probability. Exponential model size.
  - Knowledge acquisition difficult (complex, unnatural),
  - Exponential storage and inference.

Chain Rule and Factorization

Overcome the problem of exponential size by exploiting conditional independence

- The chain rule of probabilities:

$$
\begin{aligned}
P(X_1, X_2) &= P(X_1)P(X_2|X_1) \\
P(X_1, X_2, X_3) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \\
&\dots \\
P(X_1, X_2, \dots, X_n) &= P(X_1)P(X_2|X_1)\dots P(X_n|X_1, \dots, X_{n-1}) \\
&= \prod_{i=1}^{n} P(X_i|X_1, \dots, X_{i-1}).
\end{aligned}
$$

- No gains yet. The number of parameters required by the factors is: $2^{n-1} + 2^{n-1} + \ldots + 1 = 2^n - 1$.

Conditional Independence

- About $P(X_i|X_1,\ldots,X_{i-1})$:
  - Domain knowledge usually allows one to identify a subset $pa(X_i) \subseteq \{X_1,\ldots,X_{i-1}\}$ such that
    - Given $pa(X_i)$, X_i is independent of all variables in $\{X_1,\ldots,X_{i-1}\} \setminus pa(X_i)$, i.e.

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

- Then

$$P(X_1,X_2,\ldots,X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i))$$

- Joint distribution factorized.
- The number of parameters might have been substantially reduced.

Example

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

Example continued

$$
\begin{aligned}
P(B,E,A,J,M) &= P(B)P(E|B)P(A|B,E)P(J|B,E,A)P(M|B,E,A,J)\\
&= P(B)P(E)P(A|B,E)P(J|A)P(M|A)(Factorization)
\end{aligned}
$$

- $pa(B)=\{\}$, $pa(E)=\{\}$,$pa(A)=\{B,E\}$, $pa(J)=\{A\}$,$pa(M)=\{A\}$.
- Conditional probabilities tables (CPT)

B	P(B)
Y	.01
N	.99

E	P(E)
Y	.02
N	.98

A	B	E	P(A|B, E)
Y	Y	Y	.95
N	Y	Y	.05
Y	Y	N	.94
N	Y	N	.06
Y	N	Y	.29
N	N	Y	.71
Y	N	N	.001
N	N	N	.999

M	A	P(M|A)
Y	Y	.9
N	Y	.1
Y	N	.05
N	N	.95

J	A	P(J|A)
Y	Y	.7
N	Y	.3
Y	N	.01
N	N	.99

Example continued

- Model size reduced from 31 to 1+1+4+2+2=10
- Model construction easier
  - Fewer parameters to assess.
  - Parameters more natural to assess:e.g.

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

$$P(J = Y|A = Y), P(M = Y|A = Y)$$

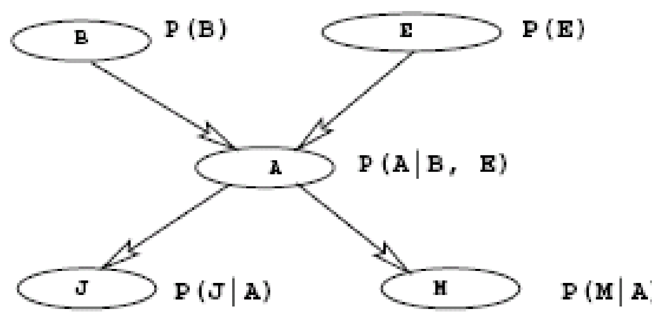
- Inference easier.Will see this later.

From Factorizations to Bayesian Networks

Graphically represent the conditional independency relationships:

- construct a directed graph by drawing an arc from X_j to X_i iff $X_j \in pa(X_i)$

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



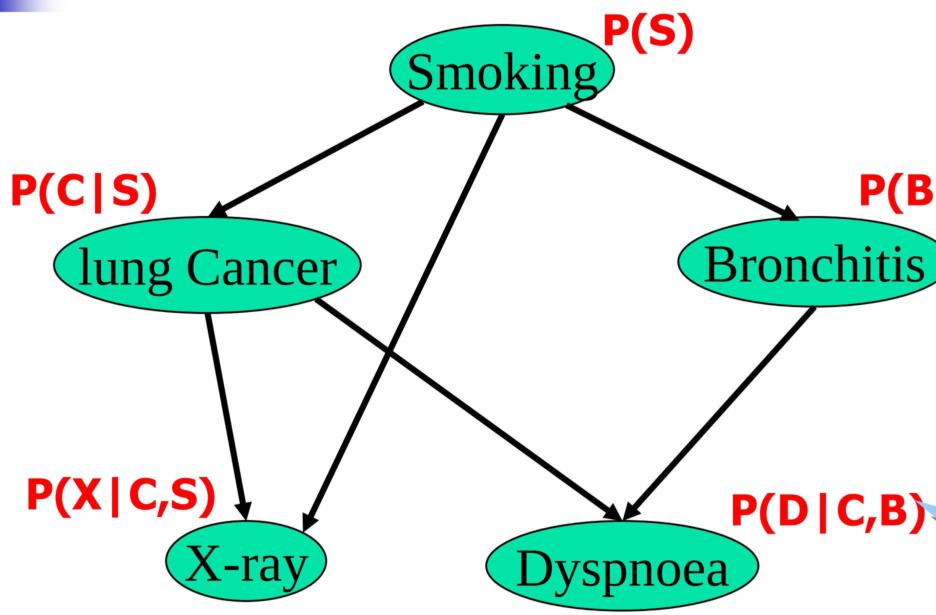
- Also attach the conditional probability (table) $P(X_i|pa(X_i))$ to node X_i.
- What results in is a Bayesian network. Also known as **belief network, probabilistic network.**

Formal Definition

A **Bayesian network** is:

- An directed acyclic graph (DAG), where
- Each node represents a random variable
- And is associated with the conditional probability of the node given its parents.

Bayesian Networks: Representation



BN = (G, Θ)

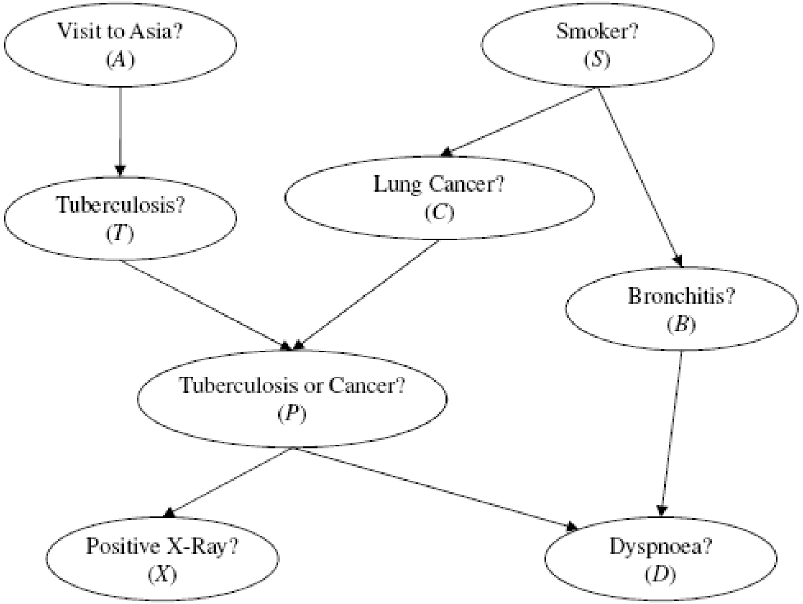
CPD:

C	B	D=0	D=1
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

$$P(S, C, B, X, D) = P(S)\, P(C|S)\, P(B|S)\, P(X|C,S)\, P(D|C,B)$$

Conditional Independencies ➡ Efficient Representation

Capturing Independence Graphically



We would clearly find a visit to Asia relevant to our belief in the X-Ray test coming out positive, but we would find the visit irrelevant if we know for sure that the patient does not have Tuberculosis. That is, X is dependent on A, but is independent of A given $\neg T$.