CS480 – Introduction to Artificial Intelligence

Topic: Bayesian Networks
Chapter: 14

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

JOINT DISTRIBUTION

- We have n random variables, $V_1, V_2, \ldots, V_n$
- We are interested in the probability of a possible world, where
  - $V_1=v_1, V_2=v_2, \ldots, V_n = v_n$
- $P(V_1, V_2, \ldots, V_n)$ associates a probability for each possible world ≡ the **joint distribution**
- How many independent parameters are needed, if V_i are all binary?

$2^n - 1$

Joint Distribution

- Extremely useful
  - Can answer any type of query
- Extremely inefficient
  - Requires exponential size memory
  - Inference using an exponential-size table requires exponential time
- Chapter 14 $\Rightarrow$ Efficient representation and inference

Chain Rule

- $P(V_1, V_2, \ldots, V_n) =$ $2^n - 1$
  $1 + 2 + 4 + 8 + 16 + \cdots + 2^{n-1} = 2^n - 1$
  - $P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_1, V_2) \ldots P(V_n \mid V_1, V_2, \ldots, V_{n-1})$
  - $P(V_2)P(V_1 \mid V_2)P(V_3 \mid V_2, V_1) \ldots P(V_n \mid V_2, V_1, \ldots, V_{n-1})$
  - …
- If all V_i are binary, $P(V_1, V_2, \ldots, V_n)$ requires 2^n-1 independent parameters
- $P(V_1)$: How many?
- $P(V_2 \mid V_1)$: How many?
- $P(V_3 \mid V_1, V_2)$: How many?
- …
- $P(V_n \mid V_1, V_2, \ldots, V_{n-1})$: How many?
- How many in total?

Marginal Independence

- Two random variables A and B are **marginally independent** if and only if
  - $P(A, B) = P(A)*P(B)$, equivalently
  - $P(A|B) = P(A)$, equivalently
  - $P(B|A) = P(B)$

THE JOINT REVISITED

- $P(V_1, V_2, \ldots, V_n) =$
  - $P(V_1)P(V_2 | V_1)P(V_3 | V_1,V_2) \ldots P(V_n | V_1, V_2, \ldots, V_{n-1})$
- If $V_i \perp V_j$ for all $i \neq j$
  - $P(V_1, V_2, \ldots, V_n) =$
    - $P(V_1)P(V_2 | V_1)P(V_3 | V_1,V_2) \ldots P(V_n | V_1, V_2, \ldots, V_{n-1})$
    - $P(V_1)P(V_2)P(V_3) \ldots P(V_n)$
    - $1 + 1 + 1 + \cdots + 1 = n$
    - How many independent parameters now?

CONDITIONAL INDEPENDENCE

- Marginal independence is not very common
- Two random variables A and B are conditionally independent given C if and only if
  - $P(A, B | C) = P(A | C)*P(B | C)$, equivalently
  - $P(A | B,C) = P(A | C)$, equivalently
  - $P(B | A,C) = P(B | C)$

Why Independence?

- The joint distribution for n binary random variables
  - $2^n - 1$ independent entries; exponential
- If the variables were all
  - Marginally independent, then
    - $1 + 1 + \ldots + 1 = n$ independent parameters; polynomial
  - Conditionally independent given one of them, then
    - $1 + 2 + 2 + \ldots + 2 = 1 + 2(n-1) = 2n - 1$ independent parameters; polynomial

Advantages of more compact representation

- Fewer parameters
  - Makes learning and reasoning easier
- Consider asking an expert the probability of specific entry in a huge probability table

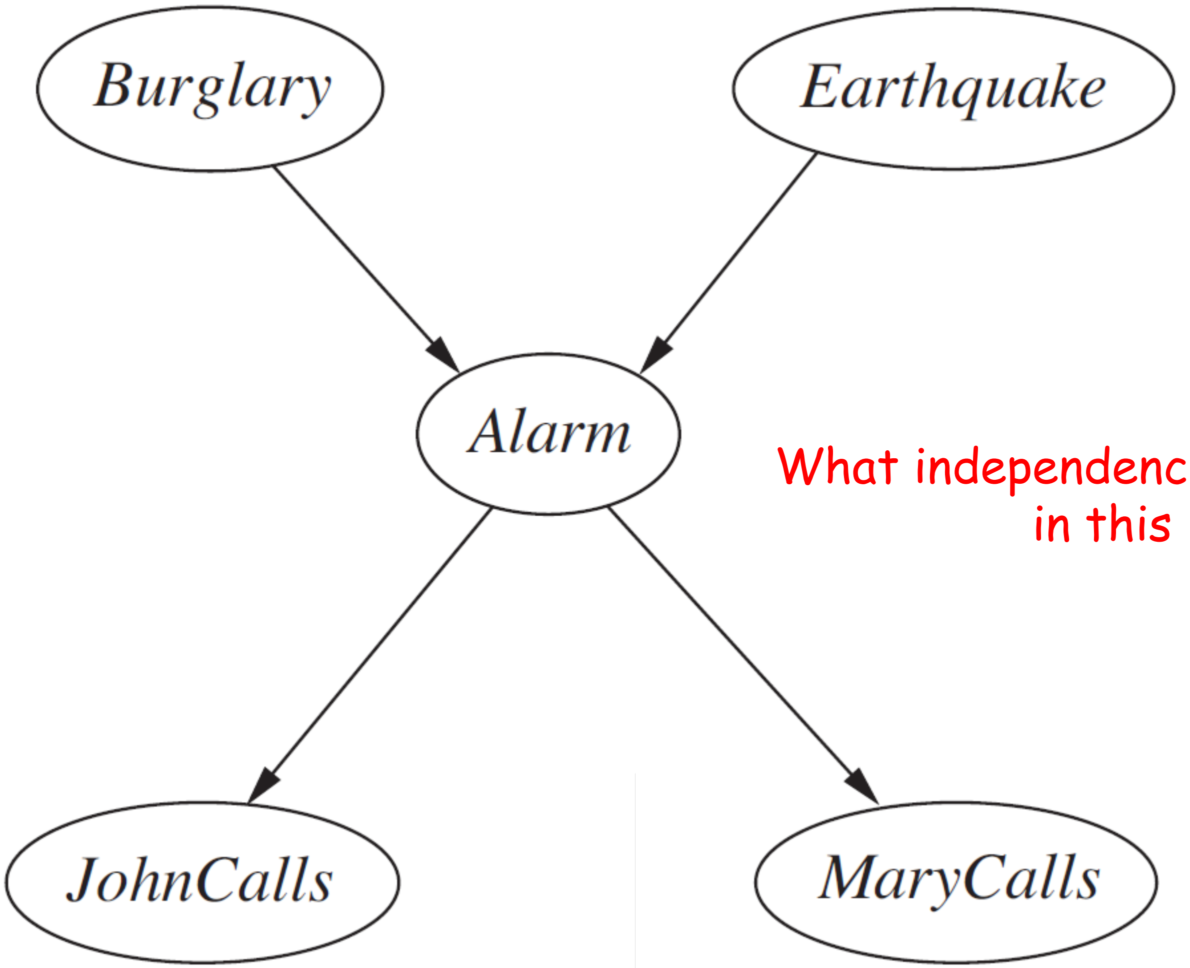
Bayesian Networks

- Random variables = nodes
- Direct relationships = directed edges
- BNs capture independencies
  - More compact than full joint representation
- Graphs provide
  - Graph theory / efficient reasoning
  - Intuition

Examples

- X causes Y and Y causes Z; no direct relationship between X and Z
  - $X \rightarrow Y \rightarrow Z$
  - Nothing is marginally independent of each other
  - $Z \perp X \mid Y$
- Y causes both X and Z; no direct relationship between X and Z
  - $X \leftarrow Y \rightarrow Z$
  - Nothing is marginally independent of each other
  - $Z \perp X \mid Y$
- Both X and Z cause Y; no direct relationship between X and Z
  - $X \rightarrow Y \leftarrow Z$
  - X and Z are marginally independent
  - X and Z become dependent when the value of Y is known

Burglary Example



What independencies are encoded in this BN?

Independencies – D-separation

- Definition: Observed ≡ It's value is known
- Causal trail
  - X → Y → Z; E.g., Burglary → Alarm → MaryCalls
  - X and Z are independent if Y is observed
- Evidential trail
  - X ← Y ← Z; E.g., MaryCalls ← Alarm ← Burglary
  - X and Z are independent if Y is observed
- Common cause
  - X ← Y → Z; E.g., JohnCalls ← Alarm → MaryCalls
  - X and Z are independent if Y is observed
- Common effect
  - X → Y ← Z; E.g., Burglary → Alarm ← Earthquake
  - X and Z are marginally independent but they become dependent if Y or any of Y's descendants are observed

D-separation

- Examples

Independencies - Parents

- X is independent of its non-descendants given its parents
  - $X \perp \text{Non-descendants}(X) \mid \text{Parents}(X)$
- What's a non-descendant?
- What are the independencies in the burglary example?

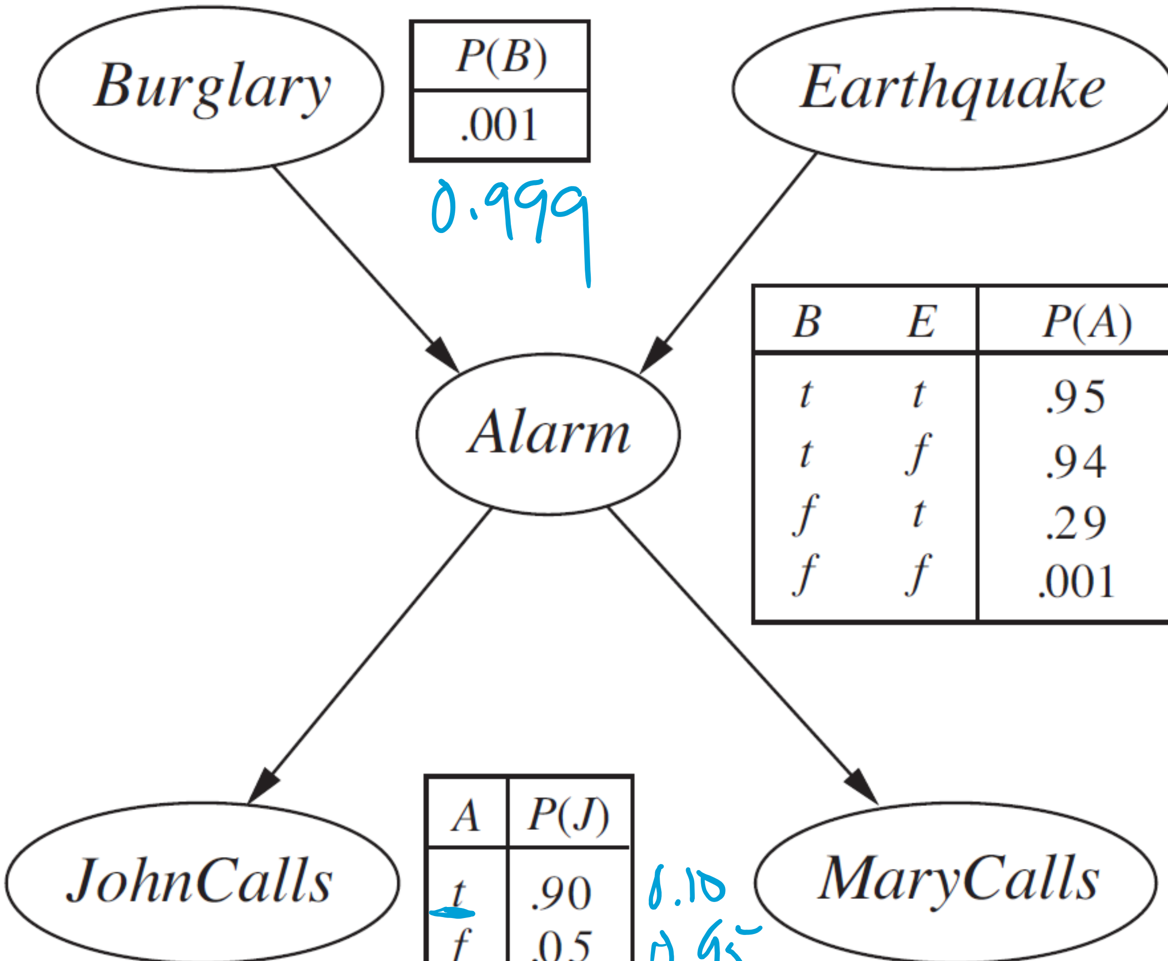
PARAMETERIZATION

Given the indecencies encoded in a BN, what are the parameters needed to capture the joint representation efficiently?

Bayesian network parameterization

$$P(\mathbf{V}) = \prod_i P(V_i \mid Pa(V_i))$$

BURGLARY EXAMPLE



P(B)
.001

0.999

P(E)
.002

0.998

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

0.05
0.06
0.71
0.999

A	P(J)
t	.90
f	.05

0.10
0.95

A	P(M)
t	.70
f	.01

0.30
0.99

THEOREMS

Ind Assumption $\Longleftrightarrow$ Factorizat $\Longrightarrow$

- **Theorem 1:** If a probability distribution P holds the independencies encoded in G, then P factorizes according to G
- **Theorem 2:** If P factorizes according to G, then it holds the independencies encoded in G
- Let's see a constructive proof for Theorem 1; we'll not prove Theorem 2

FROM INDEPENDENCE TO FACTORIZATION

- Linear chain example
  - $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$
- Burglary example

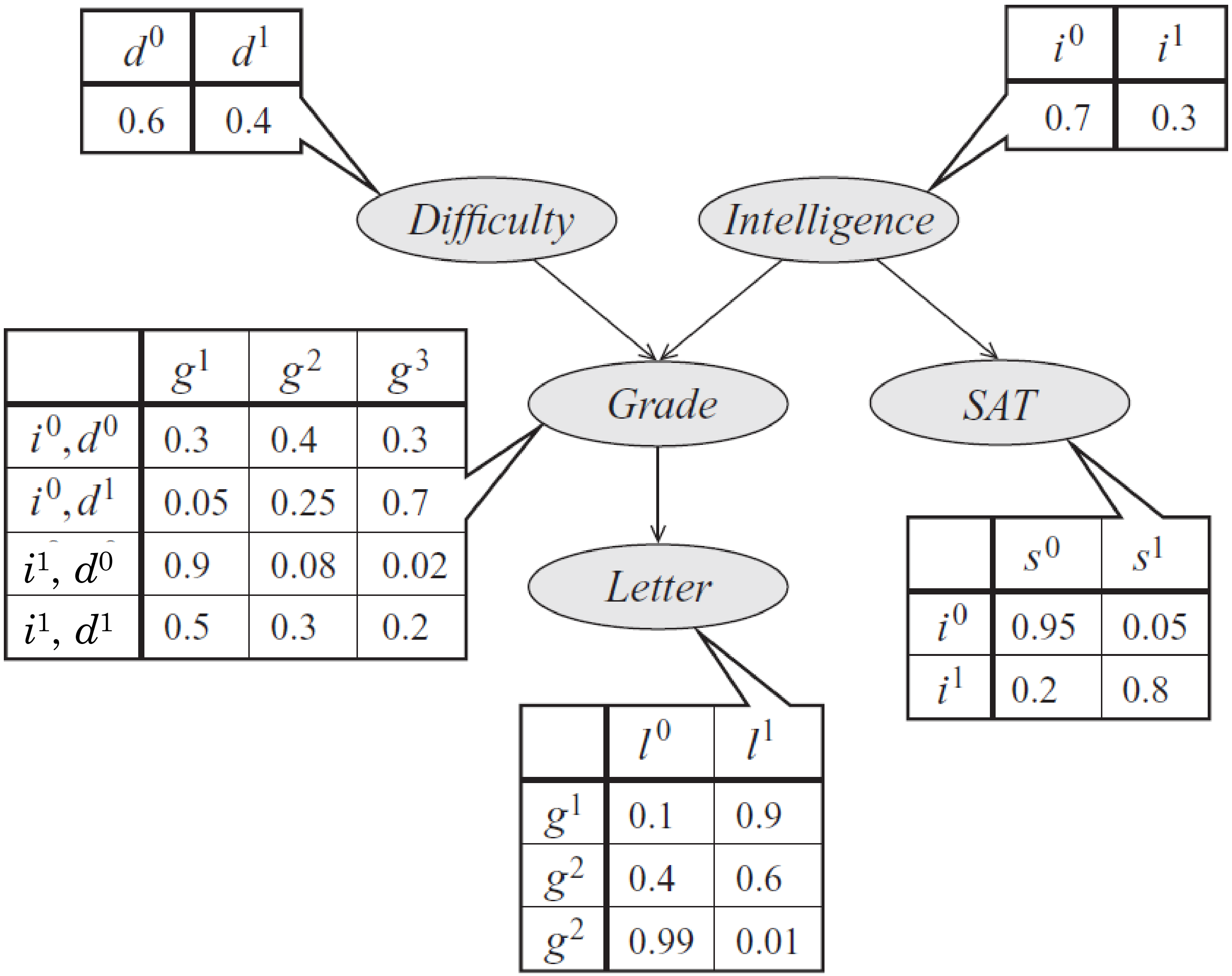
$$P(A,B,C,D,E) = P(A)\,P(B|A)\,P(C|B)\,P(D|C)\,P(E|D)$$

$$P(A)\,P(B|A)\,P(C|A,\cancel{B})\,P(D|\cancel{A},\cancel{B},C)\,P(E|\cancel{A},\cancel{B},\cancel{C},D)$$

Burglary Example

- The joint representation
  - Equation
- Contrast number of parameters for
  - Probability table for joint
  - Bayesian network

Student Example



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

Student Example

- The joint representation
  - Equation
- Contrast number of parameters for
  - Probability table for joint
  - Bayesian network

So far

- We've discussed the representation
- Now, it's time for inference

Reasoning Patterns

- **Causal reasoning**
  - From causes to effects
    - E.g., Burglary to Alarm to MaryCalls
    - E.g., Intelligence to Grade to Letter
- **Evidential reasoning**
  - From effects to the causes
    - E.g., JohnCalls to Alarm to Earthquake
    - E.g, Letter to Grade to Difficulty
- **Explaining away/inter-causal reasoning**
  - Causes of a common effect interact
    - E.g., Earthquake, Burglary, and Alarm (and Alarm's descendants)
    - E.g., Difficulty, Intelligence, and Grade (and Grade's descendants)

Inference in Bayesian Networks

- There are several methods, some are exact and some are approximate
- We will study only one in this class
- *Variable Elimination*

Variable Elimination

- Let
  - **V** be the set of all variables, **Q** be the set of query variables, **E** be the set of evidence variables
  - P(**Q** | **E**) be the query

1. Write down the joint dist. using the Bayesian network structure
2. Set the variables in **E** to their respective values
3. Sum over all variables in **V\(Q∪E)**
   a) Pick an order for variables in **V\(Q∪E)**
   b) For each variable V_i in **V\(Q∪E),** create a new factor by
      - Multiplying all the factors that contains V_i, and
      - Summing over possible values of V_i
4. Normalize the last remaining factor (this step is unnecessary if **E** is empty)

EXAMPLES

- Given the following BNs, compute the requested probabilities efficiently (without computing the full joint)
  - $A \rightarrow B \rightarrow C$;
    - $P(A) = \langle 0.6, 0.4 \rangle$,
    - $P(B \mid A{=}t) = \langle 0.8, 0.2 \rangle$, $P(B \mid A{=}f) = \langle 0.1, 0.9 \rangle$
    - $P(C \mid B{=}t) = \langle 0.7, 0.3 \rangle$, $P(C \mid B{=}f) = \langle 0.4, 0.6 \rangle$
  - Compute $P(A)$, $P(B)$, $P(C)$, $P(C \mid A{=}t)$, $P(A \mid C{=}t)$

Irrelevant

- Let
  - **V** be the set of all variables, **Q** be the set of query variables, **E** be the set of evidence variables
  - P(**Q** | **E**) be the query
- $Y \in \boldsymbol{V} \setminus \{\boldsymbol{Q} \cup \boldsymbol{E}\}$ is irrelevant iff
  - $Y \notin Ancestors\ of \{\boldsymbol{Q} \cup \boldsymbol{E}\}$
    - or
  - $Y \perp \boldsymbol{Q} \mid \boldsymbol{E}$
- Examples

Why Variable Elimination?

- We could compute P(D) by
  - Computing the full joint table, and then
  - Summing over the remaining variables
- Variable elimination, with a *good* ordering, can
  - Save memory, and
  - Save time

Applications of Bayesian Networks

- Too many to list
- Here is a book about it: http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470060301.html
- Chapters include:
  - Medical diagnosis
  - Complex genetic models
  - Crime risk factors analysis
  - Inference problems in forensic science
  - Classifiers for modeling of mineral potential
  - Reliability analysis of systems
  - Credit-rating of companies
  - Classification of Chilean wines
  - Complex industrial process operation
  - Probability of default for large corporates
  - Risk management in robotics