COMP538: Introduction to Bayesian Networks

Lecture 2: Bayesian Networks

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Fall 2008

Objective

- Objective: Explain the concept of Bayesian networks
- Reading: Zhang & Guo, Chapter 2
- References: Russell & Norvig, Chapter 15

Outline

The probabilistic approach to reasoning under uncertainty

- A problem domain is modeled by a list of variables X_1, X_2, ..., X_n,
- Knowledge about the problem domain is represented by a joint probability $P(X_1, X_2, \ldots, X_n)$.

Example: Alarm (Pearl 1988)

- Story: In LA, burglary and earthquake are not uncommon. They both can cause alarm. In case of alarm, two neighbors John and Mary may call.
- Problem: Estimate the probability of a burglary based who has or has not called.
- Variables: Burglary (B), Earthquake (E), Alarm (A), JohnCalls (J), MaryCalls (M).
- Knowledge required by the probabilistic approach in order to solve this problem:

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

Joint probability distribution

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

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

Inference with joint probability distribution

- What is the probability of burglary given that Mary called, $P(B{=}y|M{=}y)$?
- Compute *marginal probability*:

$$P(B, M) = \sum_{E,A,J} P(B, E, A, J, M)$$

B	M	Prob
y	y	.000115
y	n	.000075
n	y	.00015
n	n	.99966

- Compute answer (reasoning by conditioning):

$$\begin{aligned} P(B{=}y|M{=}y) &= \frac{P(B{=}y, M{=}y)}{P(M{=}y)} \\ &= \frac{.000115}{.000115 + 000075} = 0.61 \end{aligned}$$

Advantages

- Probability theory well-established and well-understood.
- In theory, can perform arbitrary inference among the variables given a joint probability. This is because the joint probability contains information of all aspects of the relationships among the variables.
  - Diagnostic inference:
    - From effects to causes.
    - Example: $P(B{=}y|M{=}y)$
  - Predictive inference:
    - From causes to effects.
    - Example: $P(M{=}y|B{=}y)$
  - Combining evidence:

$$P(B{=}y|J{=}y, M{=}y, E{=}n)$$

- All inference sanctioned by laws of probability and hence has clear semantics.

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. [Exercise: Verify this.]
- 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.

Outline

Chain Rule and Factorization

Overcome the problem of exponential size by exploiting conditional independence

- The chain rule of probabilities:

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

- 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 continued

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

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

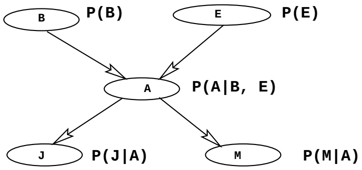
Outline

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:

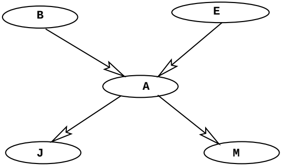
- An directed acyclic graph (DAG), where
- Each node represents a random variable
- And is associated with the conditional probability of the node given its parents.

Recall: In introduction, we said that

- Bayesian networks are networks of random variables.

Understanding Bayesian networks

- Qualitative level:
  - A directed acyclic graph (DAG) where arcs represent direct probabilistic dependence.



  - Absence of arc indicates conditional independence:
    - A variable is conditionally independent of all its **nondescendants** given its parents. (Will prove this later.)
  - The above DAG implies the following conditional independence relationships:
    - $B \perp E$; $J \perp B|A$; $J \perp E|A$; $M \perp B|A$; $M \perp E|A$; $M \perp J|A$
  - The following are not implied:
    - $J \perp B$; $J \perp E$; $J \perp M$; $B \perp E|A$

Understanding Bayesian networks

- Quantitative (numerical) level:
  - Conditional probability tables:

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

  - Describe how parents of a variable influence the variable.

Understanding Bayesian Networks

- As a whole:
  - A Bayesian network represents a **factorization** of a joint distribution.

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

  - Multiplying all the CPTs results in a joint distribution over all variables.

Example networks

Network repository

- Bayesian Network Repository:
  http://www.cs.huji.ac.il/labs/compbio/Repository/
- Genie & Smile Network Repository:
  http://genie.sis.pitt.edu/networks.html
- Netica Net Library: http://www.norsys.com/netlibrary/index.htm
- Hugin Case Studies: http://www.hugin.com/cases/

Software

- Genie & Smile: http://genie.sis.pitt.edu/. Free.
- Netica: http://www.norsys.com/. Free version for small nets.

Outline

Procedure for constructing Bayesian network structures

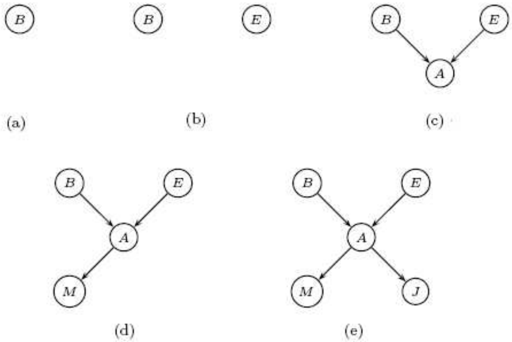
1. Choose a set of variables that describes the application domain.
2. Choose an ordering for the variables.
3. Start with the empty network and add variables to the network one by one according to the ordering.
4. To add the i-th variable X_i,
    1. Determine a subset $pa(X_i)$ of variables already in the network ($X_1, \ldots, X_{i-1}$) such that

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

       (Domain knowledge is needed here.)
    2. Draw an arc from each variable in $pa(X_i)$ to X_i.

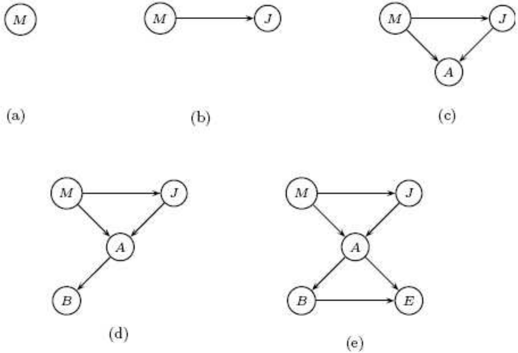
Examples

- Order 1: B, E, A, J, M
  - $pa(B) = \{\}$, $pa(E) = \{\}$,
    $pa(A) = \{B, E\}$, $pa(J) = \{A\}$, $pa(M) = \{A\}$.



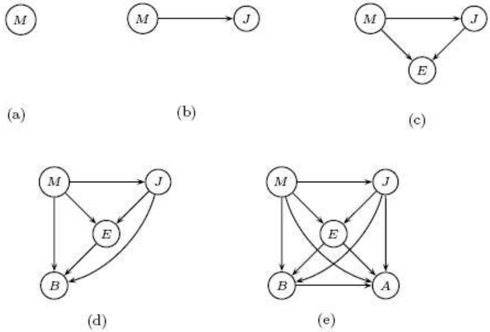
Examples

- Order 2: M, J, A, B, E
  - $pa(M) = \{\}$, $pa(J) = \{M\}$, $pa(A) = \{M, J\}$, $pa(B) = \{A\}$, $pa(E) = \{A, B\}$.



Examples

- Order 3: M, J, E, B, A
  - $pa(M) = \{\}$, $pa(J) = \{M\}$, $pa(E) = \{M, J\}$, $pa(B) = \{M, J, E\}$, $pa(A) = \{M, J, B, E\}$.



Building Bayesian network structures

Which order?

- Naturalness of probability assessment (Howard and Matheson).
    - (B, E, A, J, M) is a good ordering because the following distributions natural to assess
        - $P(B)$, $P(E)$: frequency of burglary and earthquake
        - $P(A|B, E)$: property of Alarm system.
        - $P(M|A)$: knowledge about Mary
        - $P(J|A)$: knowledge about John.
    - The order M, J, E, B, A is not good because, for instance, $P(B|J, M, E)$ is unnatural and hence difficult to assess directly.

Building Bayesian network structures

Which order?

- Minimize number of arcs (J. Q. Smith).
  - The order (M, J, E, B, A) is bad because too many arcs.
  - In contrast, the order (B, E, A, J, M) is good is because it results in a simple structure.
- Use causal relationships (Pearl): cause come before their effects.
  - The order (M, J, E, B, A) is not good because, for instance, M and J are effects of A but come before A.
  - In contrast, the order (B, E, A, J, M) is good is because it respects the causal relationships among variables.

Exercise in Structure Building

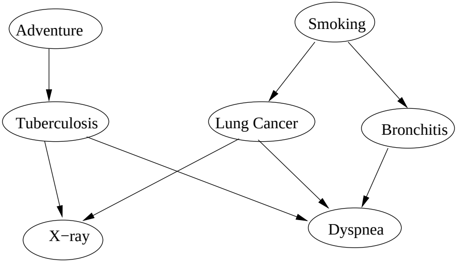
- Five variable about what happens to an office building
  - Fire: There is a fire in the building.
  - Smoke: There is smoke in the building.
  - Alarm: Fire alarm goes off.
  - Leave: People leaves the building.
  - Tampering: Someone tamper with the fire system (e.g., open fire exit)
- Build network structures using the following ordering. Clearly state your assumption.
  1. Order 1: tampering, fire, smoke, alarm, leave
  2. Order 2: leave, alarm, smoke, fire, tampering

Causal Bayesian networks

- Build a Bayesian network using casual relationships:
  - Choose a set of variables that describes the domain.
  - Draw an arc to a variable from each of its DIRECT causes. (Domain knowledge needed here.)
- What results in is a **causal Bayesian network**, or simply **causal networks**,
  - Arcs are interpreted as indicating cause-effect relationships.

Example:

- Travel (Lauritzen and Spiegelhalter)



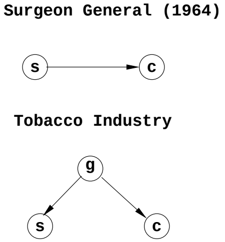
Use of Causality: Issue 1

Causality is not a well understood concept.

- No widely accepted definition.
- No consensus on
  - Whether it is a property of the world,
  - Or a concept in our minds helping us to organize our perception of the world.

Causality

- Sometimes causal relations are obvious:
  - Alarm causes people to leave building.
  - Lung Cancer causes mass on chest X-ray.
- At other times, they are not that clear.
  - Whether gender influences ability in technical sciences.
  - Most of us believe Smoking cause lung cancer,but the tobacco industry has a different story:



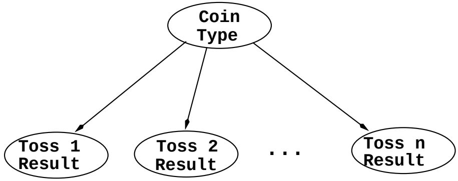
Working Definition of Causality

- Imagine an all powerful agent, GOD, who can change the states of variables .
  - X **causes** Y if *knowing that GOD has changed the state of* X changes your believe about Y.
- Example:
  - "Smoking" and "yellow finger" are correlated.
  - If we force someone to smoke for sometime, his finger will probably become yellow. So "Smoking" is a cause of "yellow finger".
  - If we paint someone's finger yellow, that will not affect our belief on whether s/he smokes. So "yellow finger" does not cause "smoking".
- Similar example with Earthquake and Alarm

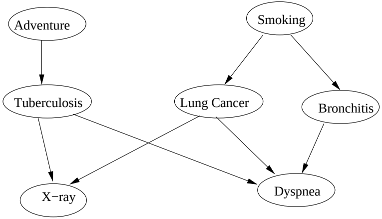
Causality

Coin tossing example revisited:

- Knowing that GOD somehow made sure the coin drawn from the bag is a fair coin would affect our belief on the results of tossing.
- Knowing that GOD somehow made sure that the first tossing resulted in a head does not affect our belief on the type of the coin.
- So arrows go from coin type to results of tossing.



Use of Causality: Issue 2



- Causality ⇒ network structure (building process)
- Network structure ⇒ conditional independence (Semantics of BN)

The **causal Markov assumption** bridges causality and conditional independence:

- A variable is independent of all its non-effects (non-descendants) given its direct causes (i.e. parents).

We make this assumption if we determine Bayesian network structure using causality.

Determining probability parameters

- Later in this course, we will discuss in detail how to learn parameters from data.
- We will not be so much concerned with eliciting probability values from experts.
- However, people do that some times. In such a case, one would want the number of parameters be as small as possible.
- The rest of the lecture describe two concepts for reducing the number of parameters:
  - Causal Independence.
  - Context-specific independence.
- Left to students as reading materials.

Determining probability parameters

- Sometimes conditional probabilities are given by domain theory¶
  - Genetic inheritance in Stud (horse) farm
    (Jensen, F. V. (2001). Bayesian networks and decision graphs. Springer.):

P(Child|Father, Mother)

	aa	aA	AA
aa	(1, 0, 0)	(.5, .5, 0)	(0, 1, 0)
aA	(.5, .5, 0)	(.25, .5, 25)	(0, .5, .5)
AA	(0, 1, 0)	(0, .5, .5)	(0, 0, 1)

Genotypes: aa - sick, aA - carrier, AA - pure.

Determining probability parameters

- Sometimes, we need to get the numbers from the experts.
  - This is a time-consuming and difficult process.
  - Nonetheless, many networks have been built. See Bayesian Network Repository at
    `http://www.cs.huji.ac.il/labs/compbio/Repository/`

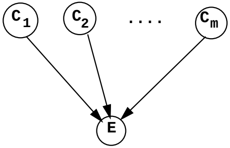
- Combine experts' knowledge and data
  - Use assessments by experts as a start point.
  - When data become available, combine data and experts' assessments.
  - As more and more data become available, influence of experts is automatically reduced.

  We will show how this can be done when discussing parameter learning.

Note: Much of the course will be about how to learning Bayesian networks (structures and parameters) from data.

Reducing the number of parameters

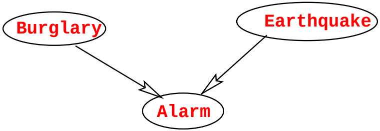
- Let E be a variable in a BN and let C_1, C_2, ..., C_m be its parents.

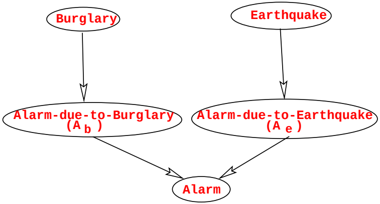


- The size of the conditional probability $P(E|C_1, C_2, \ldots, C_m)$ is exponential in m.
- This poses a problem for knowledge acquisition, learning, and inference.
- In application, there usually exist **local structures** that one can exploit to reduce the size of conditional probabilities

Causal independence

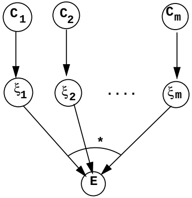
- **Causal independence** refers to the situation where
  - the causes C_1, C_2 ..., and C_m influence E independently.
  - In other words, the ways by which the C_i's influence e are independent.





- Example:
  - Burglary and earthquake trigger alarm independently.
  - Precise statement: A_b and A_e are independent.
  - $A = A_b \vee A_e$, hence **Noisy-OR gate** (Good 1960).

Causal Independence

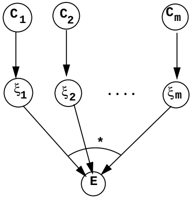


- Formally, C_1, C_2 ..., and C_m are said to be **causally independent** w.r.t effect E if
  - there exist random variables ξ_1, ξ_2 ..., and ξ_m such that
    1. For each i, ξ_i probabilistically depends on C_i and is conditionally independent of all other C_j's and all other ξ_j's given C_i, and
    2. There exists a commutative and associative binary operator $*$ over the domain of e such that

$$E = \xi_1 * \xi_2 * \ldots * \xi_m$$

.

Causal Independence



- In words, individual contributions from different causes are independent and the total influence on effect is a combination of the individual contributions.
- ξ_i – **contribution** of C_i to E.
- * – **base combination operator**.
- E – **independent cause (IC)** variable. Known as convergent variable in Zhang & Poole (1996).

Causal Independence

- Example: Lottery
  - C_i: money spent on buying lottery of type i.
  - E: change of wealth.
  - ξ_i: change in wealth due to buying the ith type lottery.
  - Base combination operator: "$+$". (**Noisy-adder**)

- Other causal independence models:
  1. Noisy MAX-gate — *max*
  2. Noisy AND-gate — $\wedge$

Causal Independence

Theorem (2.1)

If C_1, C_2, ..., C_m are causally independent w.r.t E, then the conditional probability $P(E|C_1, \ldots, C_m)$ can be obtained from the conditional probabilities $P(\xi_i|C_i)$ through

$$P(E{=}e|C_1, \ldots, C_m) = \sum_{\alpha_1 * \ldots * \alpha_k = e} P(\xi_1{=}\alpha_1|C_1)\ldots P(\xi_m{=}\alpha_m|C_m), \qquad (1)$$

for each value e of E. Here $*$ is the base combination operator of E.

See Zhang and Poole (1996) for the proof.

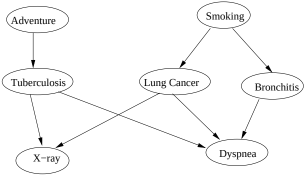
Causal Independence

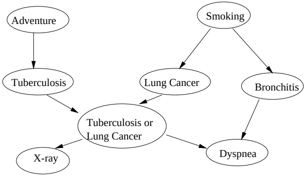
Notes:

- Causal independence reduces model size:
  - In the case of binary variable, it reduces model sizes from 2^{m+1} to $4m$.
  - Examples: CPSC, Carpo
- It can also be used to speed up inference (Zhang and Poole 1996).
- Relationship with logistic regression? (Potential term project)

Parent divorcing

Another technique to reduce the number of parameters





- Top figure: A more natural model for the Travel example. But it requires 1+1+2+2+2+4+8=20 parameters.
- Low figure: requires only 1+1+2+2+2+4+2+4=18 parameters.
- The difference would be bigger if, for example, D have other parents.
- The trick is to introduce a new node (TB-or-LC).
- It **divorces** T and L from the other parent B of D.
- Note that the trick would not help if the new node TB-or-LC has 4 or more states.

Context specific independence (CSI)

- Let **C** be a set of variables. A **context** on **C** is an assignment of one value to each variable in **C**.
- We denote a context by **C**=**c**, where **c** is a set of values of variables in **C**.
- Two contexts are **incompatible** if there exists a variable that is assigned different values in the contexts.
- They are **compatible** otherwise.

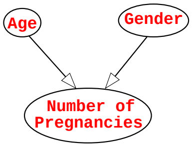
Context-specific independence

- Let **X**, **Y**, **Z**, and **C** be four disjoint sets of variables.
- **X** and **Y** are **independent given Z in context C=c** if

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

  whenever $P(\mathbf{Y}, \mathbf{Z}, \mathbf{C}{=}\mathbf{c}){>}0$.
- When **Z** is empty, one simply says that **X** and **Y** are **independent in context C=c**.

Context-specific independence

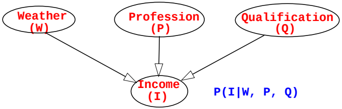


- Shafer's Example:
  - *Number of pregnancies (N)* is independent of *Age (A)* in the context *Gender=Male (G=m)*.

$$P(N|A, G{=}m) = P(N|G{=}m)$$

- Number of parameters reduced by $(|A|-1)(|N|-1)$.

Context-specific independence



- *Income* independent of *Weather* in context *Profession*=*Programmer*.

$$P(I|W, P{=}Prog, Q) = P(I|P{=}Prog, Q)$$

- *Income* independent of *Qualification* in context *Profession*=*Farmer*.

$$P(I|W, P{=}Farmer, Q) = P(I|W, P{=}Farmer)$$

- Number of parameters reduced by: $(|W|-1)|Q|(|I|-1) + (|Q|-1)|W|(|I|-1)$

- CSI can also be exploited to speed up inference (Zhang and Poole 1999).

Outline

Reasons for the popularity of Bayesian networks

- It's graphical language is intuitive and easy to understand because it captures what might be called "intuitive causality".
- Pearl (1986) claims that it is a model for human's inferential reasoning:
  - Notations of dependence and conditional dependence are basic to human reasoning.
  - The fundamental structure of human knowledge can be represented by dependence graphs.

Reasons for the popularity of Bayesian networks

- In practice, the graphical language
  - Functions as a convenient language to organizes one's knowledge about a domain.
  - Facilitates interpersonal communication.
- On the other hand, the language is well-defined enough to allow computer processing.
  - Correctness of results guaranteed by probability theory.
- For probability theory, Bayesian networks provide a whole new perspective:
- *"Probability is not really about numbers; It is about the structure of reasoning." (Glenn Shafer)*