CS 1571 Introduction to AI
Lecture 17

Bayesian belief networks

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Uncertainty

- **Is an essential feature of many real-world problems**
- **Relations between components, states of the world are often uncertain**

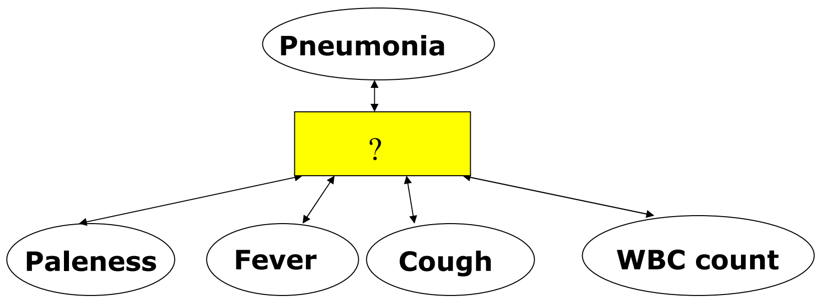
Examples:

- **Medical diagnosis**
  - A patient suffering from pneumonia may not have fever all the times, may or may not have a cough, white blood cell test can be in a normal range.
  - High fever is typical for many diseases (e.g. bacterial diseases) and does not point specifically to pneumonia
- **Therapy planning**
  - A response of a patient to a therapy is not deterministic, the patient's state can improve, stay the same, or worsen

Modeling the uncertainty.

Key issues:

- How to describe, represent the relations in the presence of uncertainty?
- How to manipulate such knowledge to make inferences?
  - **Humans can reason with uncertainty.**



Probability theory

a well-defined coherent theory for representing uncertainty and for reasoning with it

Representation:

Propositional statements – assignment of values to random variables

$$Pneumonia = True \qquad WBCcount = high$$

Probabilities over statements model the degree of belief in these statements

$$P(Pneumonia = True) = 0.001$$

$$P(WBCcount = high) = 0.005$$

$$P(Pneumonia = True, Fever = True) = 0.0009$$

$$P(Pneumonia = False, WBCcount = normal, Cough = False) = 0.97$$

Joint probability distribution

Joint probability distribution (for a set variables)

- Defines probabilities for all possible assignments of values to variables in the set

$\mathbf{P}(pneumonia, WBCcount)$ 2×3 table

		WBCcount			$\mathbf{P}(Pneumonia)$
		high	*normal*	*low*	
Pneumonia	*True*	0.0008	0.0001	0.0001	0.001
	False	0.0042	0.9929	0.0019	0.999
	$\mathbf{P}(WBCcount)$	0.005	0.993	0.002	

Marginalization - summing out variables

Variable independence

- **The joint distribution over a subset of variables can be always computed from the joint distribution through marginalization**
- **Not the other way around !!!**
  - **Only exception: when variables are independent**

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

$\mathbf{P}(pneumonia, WBCcount)$

		WBCcount			$\mathbf{P}(Pneumonia)$
		high	*normal*	*low*	
Pneumonia	*True*	0.0008	0.0001	0.0001	0.001
	False	0.0042	0.9929	0.0019	0.999
	$\mathbf{P}(WBCcount)$	0.005	0.993	0.002	

Conditional probability

Conditional probability :

- Probability of A given B

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

- Conditional probability is defined in terms of joint probabilities
- Joint probabilities can be expressed in terms of conditional probabilities

$$P(A,B) = P(A \mid B)P(B) \quad \textbf{(product rule)}$$

$$P(X_1, X_2, \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid X_{1,} \ldots X_{i-1}) \quad \textbf{(chain rule)}$$

- Conditional probability – is useful for **various probabilistic inferences**

$$P(Pneumonia = True \mid Fever = True, WBCcount = high, Cough = True)$$

Bayes rule

Conditional probability.

$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(A,B) = P(B \mid A)P(A)$$

Bayes rule:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

When is it useful?

- When we are interested in computing the diagnostic probability, from the causal probability

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

- **Reason:** It is often easier to assess causal probability
  - E.g. Probability of pneumonia causing fever
    vs. probability of pneumonia given fever

Probabilistic inference

Various inference tasks:

- **Diagnostic task. (from effect to cause)**

$$\mathbf{P}(Pneumonia \mid Fever = T)$$

- **Prediction task. (from cause to effect)**

$$\mathbf{P}(Fever \mid Pneumonia = T)$$

- **Other probabilistic queries (**queries on joint distributions).

$$\mathbf{P}(Fever)$$

$$\mathbf{P}(Fever, ChestPain)$$

Inference

Any query can be computed from the full joint distribution !!!

- **Joint over a subset of variables** is obtained through marginalization

$$P(A=a, C=c) = \sum_i \sum_j P(A=a, B=b_i, C=c, D=d_j)$$

- **Conditional probability over set of variables**, given other variables' values is obtained through marginalization and definition of conditionals

$$P(D=d \mid A=a, C=c) = \frac{P(A=a, C=c, D=d)}{P(A=a, C=c)}$$

$$= \frac{\sum_i P(A=a, B=b_i, C=c, D=d)}{\sum_i \sum_j P(A=a, B=b_i, C=c, D=d_j)}$$

Inference.

Any query can be computed from the full joint distribution !!!

- Any joint probability can be expressed as a product of conditionals via the **chain rule**.

$$P(X_1, X_2, \ldots X_n) = P(X_n \mid X_{1,} \ldots X_{n-1}) P(X_{1,} \ldots X_{n-1})$$

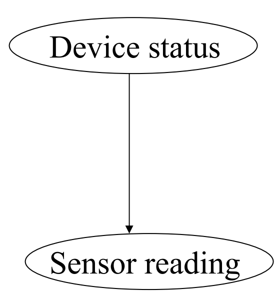
$$= P(X_n \mid X_{1,} \ldots X_{n-1}) P(X_{n-1} \mid X_{1,} \ldots X_{n-2}) P(X_{1,} \ldots X_{n-2})$$

$$= \prod_{i=1}^{n} P(X_i \mid X_{1,} \ldots X_{i-1})$$

- It is often easier to define the distribution in terms of conditional probabilities:
  - E.g. $\mathbf{P}(Fever \mid Pneumonia = T)$
    $\mathbf{P}(Fever \mid Pneumonia = F)$

Simple diagnostic inference. Example.

- **Device** (equipment) operating *normally* or *malfunctioning*.
  - Operation of the device sensed indirectly via a sensor
- **Sensor reading** is either *high* or *low*



P(Device status)

normal	malfunctioning
0.9	0.1

P(Sensor reading| Device status)

Device\Sensor	high	low
normal	0.1	0.9
malfunctioning	0.6	0.4

Diagnostic inference. Example.

- **Diagnostic inference:** compute the probability of device operating normally or malfunctioning given a sensor reading

$$\mathbf{P}(\text{Device status} \mid \text{Sensor reading} = high) = ?$$

$$= \begin{pmatrix} P(\text{Device status} = normal \mid \text{Sensor reading} = high) \\ P(\text{Device status} = malfunctioning \mid \text{Sensor reading} = high) \end{pmatrix}$$

- Note that the opposite conditional probabilities are available
- **Solution:** apply **Bayes rule** to reverse the conditioning variables

Modeling uncertainty with probabilities

- Defining the full joint distribution makes it possible to represent and reason with uncertainty in a uniform way
- We are able to handle an arbitrary inference problem

Problems:

- **Space complexity.** To store a full joint distribution we need to remember $O(d^n)$ numbers.

  n – number of random variables, d – number of values
- **Inference (time) complexity.** To compute some queries requires $O(d^n)$ steps.
- **Acquisition problem.** Who is going to define all of the probability entries?

Medical diagnosis example.

- **Space complexity**.
  - Pneumonia (2 values: T,F), Fever (2: T,F), Cough (2: T,F), WBCcount (3: high, normal, low), paleness (2: T,F)
  - Number of assignments: 2*2*2*3*2=48
  - We need to define at least 47 probabilities.
- **Time complexity**.
  - Assume we need to compute the marginal of Pneumonia=T from the full joint distribution

$$P(Pneumonia = T) =$$

$$= \sum_{i \in T,F} \sum_{j \in T,F} \sum_{k=h,n,l} \sum_{u \in T,F} P(Pneumonia = T, Fever = i, Cough = j, WBCcount = k, Pale = u)$$

  - Sum over: 2*2*3*2=24 combinations

Modeling uncertainty with probabilities

- **Knowledge based system era (70s – early 80's)**
  - **Extensional non-probabilistic models**
  - Solve the space, time and acquisition bottlenecks in probability-based models
  - froze the development and advancement of KB systems and contributed to the slow-down of AI in 80s in general
- Breakthrough (late 80s, beginning of 90s)
  - **Bayesian belief networks**
    - Give solutions to the space, acquisition bottlenecks
    - Partial solutions for time complexities
- Bayesian belief network

Bayesian belief networks (BBNs)

Bayesian belief networks.

- Represent the full joint distribution over the variables more compactly with a **smaller number of parameters**.
- Take advantage of **conditional and marginal independences** among random variables

- **A and B are independent**

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

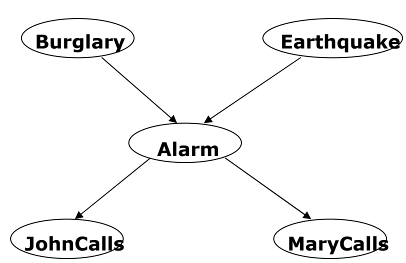
- **A and B are conditionally independent given C**

$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$
$$P(A \mid C,B) = P(A \mid C)$$

Alarm system example.

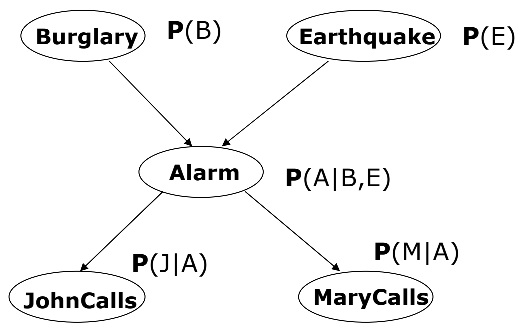
- Assume your house has an **alarm system** against **burglary**. You live in the seismically active area and the alarm system can get occasionally set off by an **earthquake**. You have two neighbors, **Mary** and **John**, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events:
  - Burglary, Earthquake, Alarm, Mary calls and John calls

Causal relations



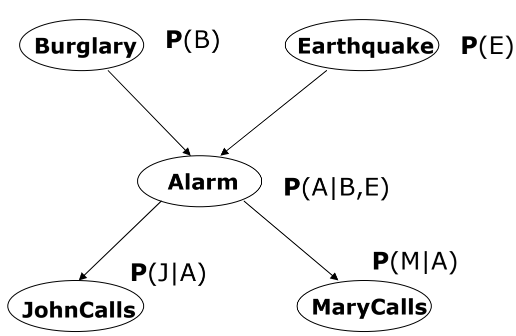
Bayesian belief network.

1. **Directed acyclic graph**
   - **Nodes** = random variables
     Burglary, Earthquake, Alarm, Mary calls and John calls
   - **Links** = direct (causal) dependencies between variables.
     The chance of Alarm being is influenced by Earthquake,
     The chance of John calling is affected by the Alarm

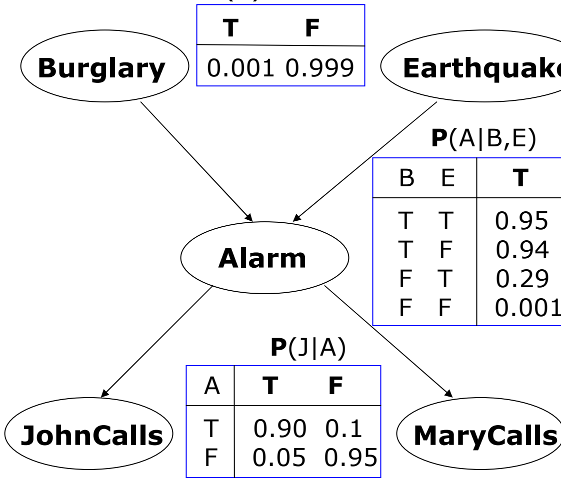


Bayesian belief network.

2. **Local conditional distributions**
   - relate variables and their parents



Bayesian belief network.



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A|B,E)

B	E	**T**	**F**
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J|A)

A	**T**	**F**
T	0.90	0.1
F	0.05	0.95

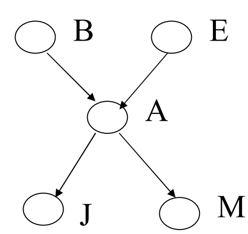
P(M|A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Bayesian belief networks (general)

Two components: $B = (S, \Theta_S)$

- **Directed acyclic graph**
  - Nodes correspond to random variables
  - (Missing) links encode independences



- **Parameters**
  - Local conditional probability distributions for every variable-parent configuration

  $\mathbf{P}(X_i \mid pa(X_i))$

  Where:

  $pa(X_i)$ - stand for parents of X_i

P(A|B,E)

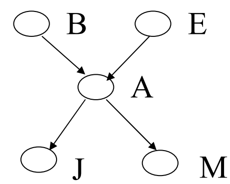
B	E	**T**	**F**
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

Full joint distribution in BBNs

Full joint distribution is defined in terms of local conditional distributions (obtained via the chain rule):

$$\mathbf{P}(X_1, X_2, .., X_n) = \prod_{i=1,..n} \mathbf{P}(X_i \mid pa(X_i))$$

Example:



Assume the following assignment of values to random variables

$B=T, E=T, A=T, J=T, M=F$

Then its probability is:

$$P(B=T, E=T, A=T, J=T, M=F) =$$
$$P(B=T)P(E=T)P(A=T \mid B=T, E=T)P(J=T \mid A=T)P(M=F \mid A=T)$$