Bayesian Networks and Decision Graphs
A 3-week course at Reykjavik University

Finn V. Jensen & Uffe Kjærulff ({fvj,uk}@cs.aau.dk)

Group of Machine Intelligence
Department of Computer Science, Aalborg University

26 April – 13 May, 2005
Outline

1 Introduction
2 Contents of Course
3 Causal Networks and Relevance Analysis
4 Bayesian Probability Theory
Introduction

1. Artificial Intelligence
2. Expert Systems
3. Normative Expert Systems
4. Sample Application Areas
What is artificial intelligence?

- Device or service that
  - reasons and makes decisions under uncertainty,
  - extracts knowledge from data/experience, and
  - solves problems efficiently and adapts to new situations.
What is artificial intelligence?
- Device or service that
  - reasons and makes decisions under uncertainty,
  - extracts knowledge from data/experience, and
  - solves problems efficiently and adapts to new situations.

Why use artificial intelligence?
- Automate tasks.
- Automate reasoning and decision making.
- Extract knowledge and information from data.
The first expert systems were constructed in the late 1960s.

*Expert System = Knowledge Base + Inference Engine*
The first expert systems were constructed in the late 1960s.

\[ \text{Expert System} = \text{Knowledge Base} + \text{Inference Engine} \]

The first expert systems were constructed as computer models of the expert, e.g. production rules like:

- if condition, then fact
- if condition, then action
The first expert systems were constructed in the late 1960s.

\[
\text{Expert System} = \text{Knowledge Base} + \text{Inference Engine}
\]

The first expert systems were constructed as computer models of the expert, e.g. production rules like:

- if \textit{condition}, then \textit{fact}
- if \textit{condition}, then \textit{action}

In most systems there is a need for handling uncertainty:

- if \textit{condition} with certainty \( x \), then \textit{fact} with certainty \( f(x) \)
- The algebras for combining \textit{certainty factors} are not mathematically coherent and can lead to \textit{incorrect conclusions}. 
Model the problem domain, not the expert.

Support the expert, don’t substitute the expert.

Use classical probability calculus and decision theory, not a non-coherent uncertainty calculus.

Closed-world representation of a given problem domain (i.e., the domain model assumes some given background conditions or context in which the model is valid).
Some important motivations for using model-based systems:

- Procedure-based (extensional) systems are semantically sloppy, model-based (intensional) systems are not.
- Speak the language of *causality*, use a single knowledge base to provide simulation, diagnosis, and prognosis.
- Both *knowledge and data* can be used to construct Bayesian networks.
- Adapt to individual settings.
- Probabilities make it easy to interface with decision and utility theory.
Bayes’ Rule

\[ P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \]

- The rule, as generalized by Laplace, is the basic starting point for inference problems using probability theory as logic.
Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or none of them, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. The results of a single chest X-ray do not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of dyspnoea.

This is a typical diagnostic situation.
Bayesian networks can be augmented with explicit representation of decisions and utilities. Such augmented models are denoted influence diagrams (or decision networks).

- Bayesian decision theory provides a solid foundation for assessing and thinking about actions under uncertainty.
- Intuitive, graphical specification of a decision problem.
- Automatic determination of a optimal strategy and computation of the maximal expected utility of this strategy.
Example: Oil Wildcatter

Oil Wildcatter

An oil wildcatter must decide either to drill or not to drill. He is uncertain whether the hole is dry, wet, or soaking. The wildcatter could perform a seismic soundings test that will help determine the geological structure of the site. The soundings will give a closed reflection pattern (indication of much oil), an open pattern (indication of some oil), or a diffuse pattern (almost no hope of oil). The cost of testing is $10,000 whereas the cost of drilling is $70,000. The utility of drilling is $270,000, $120,000, and $0 for a soaking, wet, and dry hole, respectively.

This is a typical decision scenario.
Systems based on Bayesian networks and influence diagrams are *normative*, and have the following characteristics:

- Graph representing causal relations.
- Strength of relations by probabilities.
- Preferences represented by utilities.
- Recommendations based on maximizing expected utility.
The class of tasks suitable for normative systems can be divided into three broad subclasses:

- **Forecasting:**
  - Computing probability distributions for future events.

- **Interpretation:**
  - Pattern identification (diagnosis, classification).

- **Planning:**
  - Generation of optimal sequences of decisions/actions.
Sample Application Areas of Normative Systems

- Medical
- Software
- Info. proc.
- Industry
- Economy
- Military
- Agriculture
- Mining
- Law enforcement
- Etc.
Outline

1. Introduction
2. Contents of Course
3. Causal Networks and Relevance Analysis
4. Bayesian Probability Theory
Contents of Course

Lecture plan:
1. Causal networks and Bayesian probability calculus  Tue 26/4  UK
2. Construction of Bayesian networks  Wed 27/4  UK
3. Workshop: Construction of Bayesian networks  Thu 28/4  UK
4. Inference and analyses in Bayesian networks  Fri 29/4  UK
5. Decisions, utilities and decision trees  Mon 2/5  FVJ
6. Troubleshooting and influence diagrams  Tue 3/5  FVJ
7. Solution of influence diagrams  Wed 4/5  FVJ
8. Methods for analysing an ID spec. dec. scenario  Fri 6/5  FVJ
9. Workshop: Construction of influence diagrams  Mon 9/5  FVJ
10. Learning parameters from data  Tue 10/5  FVJ
11. Bayesian networks as classifiers  Wed 11/5  FVJ
12. Learning the structure of Bayesian networks  Thu 12/5  FVJ
13. Continuous variables  Fri 13/5  FVJ

The two workshops have a duration of 4 hours.
All lectures start at 10:00.
The plan is subject to changes.
• Literature:
  • Finn V. Jensen (2001), *Bayesian Networks and Decision Graphs*, Springer-Verlag.

• Exercises suggested after each lecture.

• To pass the course you are required to
  • attend all lectures and
  • hand in written answers to home assignments.

• A number of the exercises require access to the HUGIN Tool. See the course home page for instructions on downloading and installing the HUGIN Tool.

• Home page: www.cs.aau.dk/~uk/teaching/Reykjavik-05/.
Outline

1. Introduction
2. Contents of Course
3. Causal Networks and Relevance Analysis
4. Bayesian Probability Theory
Causal Networks and Relevance Analysis

1. Causal networks, variables and DAGs
2. Relevance analysis (transmission of evidence)
   - Three types of connections
   - Explaining away
   - d-separation
Mr. Holmes is working in his office when he receives a phone call from his neighbor Dr. Watson, who tells him that Holmes’ burglar alarm has gone off. Convinced that a burglar has broken into his house, Holmes rushes to his car and heads for home. On his way, he listens to the radio, and in the news it is reported that there has been a small earthquake in the area. Knowing that earthquakes have a tendency to turn burglar alarms on, he returns to his work.
Each node in the graph represents a random variable.

In this example, each variable has state space \{no, yes\}.
Each node in the graph represents a random variable.
In this example, each variable has state space \{no, yes\}.

Three types of connections:
- Serial
- Diverging
- Converging
“Burglary” has a causal influence on “Alarm”, which in turn has a causal influence on “Watson calls”.

Burglary → Alarm → WatsonCalls
“Burglary” has a causal influence on “Alarm”, which in turn has a causal influence on “Watson calls”.

If we observe “Alarm”, any information about the state of “Burglary” is irrelevant to our belief about “Watson calls” and vice versa.
$X$ has a causal influence on $Y$ that has a causal influence on $Z$:

$\begin{align*}
\text{Serial Connections} \\
\text{Information may be transmitted through a } \textit{serial connection} \text{ unless the state of the variable (} Y \text{) in the connection is known.}
\end{align*}$
“Earthquake” has a causal influence on both “Alarm” and “Radio news”.

Diagram: [Diagram showing the causal connections between Earthquake, Alarm, and Radio News]
“Earthquake” has a causal influence on both “Alarm” and “Radio news”.

If we observe “Earthquake”, any information about the state of “Alarm” is irrelevant for our belief about an earthquake report in the “Radio news” and vice versa.
Y has a causal influence on both X and Z:

Diverging Connections

Information may be transmitted through a *diverging connection*, unless the state of the variable (Y) in the connection is known.
“Alarm” is causally influenced by both “Burglary” and “Earthquake”.

![Diagram showing causal connections between Burglary, Alarm, and Earthquake]
“Alarm” is causally influenced by both “Burglary” and “Earthquake”.

If we observe “Alarm” and “Burglary”, then this will effect our belief about “Earthquake”: Burglary explains the alarm, reducing our belief that earthquake is the triggering factor, and vice versa.
Both $X$ and $Z$ have a causal influence on $Y$:

\[ X \rightarrow Y \leftarrow Z \]

**Diverging Connections**

Information may only be transmitted through a *converging connection* if either information about the state of the variable in the connection ($Y$) or one of its descendants is available.
Transmission of Evidence

Serial
Evidence may be transmitted unless the state of $B$ is known.

Diverging
Evidence may be transmitted unless the state of $B$ is known.

Converging
Evidence may only be transmitted if $B$ or one of its descendants has received evidence.

It takes *hard evidence* to block a serial or diverging connection, whereas to open a converging connection *soft evidence* suffices.

Notice the *explaining away* effect in a converging connection: $B$ has been observed; then if $A_i$ is observed, it explains the observation of $B$ and the other causes are explained away (i.e., the beliefs in them are reduced).
The converging connection realizes the *explaining away mechanism*: The news about the earthquake strongly suggests that the earthquake is the cause of the alarm, and thereby explains away burglary as the cause.

The ability to perform this kind of *intercausal reasoning* is unique for graphical models and is one of the main differences between automatic reasoning systems based on graphical models and those based on e.g. production rules.
Assume that we have observed the symptom RunnyNose, and that there are two competing causes of it: Cold and Allergy. Observing Fever, however, provides strong evidence that cold is the cause of the problem, while our belief in Allergy being the cause decreases substantially (i.e., it is explained away by the observation of Fever).

The ability of probabilistic networks to automatically perform such intercausal inference is a key contribution to their reasoning power.
The rules for transmission of evidence over serial, diverging, and converging connections can be combined into one general rule known as *d-separation*:

**d-separation**

A path $\pi = \langle u, \ldots, v \rangle$ in a DAG, $G = (V, E)$, is blocked by $S \subseteq V$ if $\pi$ contains a node $w$ such that either

- $w \in S$ and the connections in $\pi$ do not meet head-to-head at $w$, or
- $w \notin S$, $w$ has no descendants in $S$, and the connections in $\pi$ meet head-to-head at $w$.

For three (not necessarily disjoint) subsets $A, B, S$ of $V$, $A$ and $B$ are said to be *d-separated* if all paths between $A$ and $B$ are blocked by $S$. 
If $X$ and $Y$ are not d-separated, they are \textit{d-connected}.

d-separation provides a criterion for reading statements of (conditional) dependence and independence (or relevance and irrelevance) from a causal structure.

Dependence and independence depends on what you know (and do not know).
1. C and G are d-connected.
2. C and E are d-separated.
3. C and E are d-connected given evidence on G.
4. A and G are d-separated given evidence on D and E.
5. A and G are d-connected given evidence on D.
Causal networks
- Serial/diverging/converging connections
- Transmission of evidence in causal networks
- Explaining away (intercausal reasoning)
Summary

- Causal networks
  - Serial/diverging/converging connections
  - Transmission of evidence in causal networks
  - Explaining away (intercausal reasoning)

- Dependence and independence
  - d-separation in causal networks
Outline

1. Introduction
2. Contents of Course
3. Causal Networks and Relevance Analysis
4. Bayesian Probability Theory
- Axioms of probability theory
- Probability calculus
  - Fundamental rule
  - Bayes’ rule
  - The chain rule
  - Combination and marginalization
- Conditional independence
- Evidence
The probability of an event, \( a \), is denoted \( P(a) \). Probabilities obey the following axioms:

1. \( 0 \leq P(a) \leq 1 \), with \( P(a) = 1 \) if \( a \) is certain.
The probability of an event, $a$, is denoted $P(a)$. Probabilities obey the following axioms:

1. $0 \leq P(a) \leq 1$, with $P(a) = 1$ if $a$ is certain.
2. If events $a$ and $b$ are mutually exclusive, then

$$P(a \text{ or } b) \equiv P(a \lor b) = P(a) + P(b).$$

In general, if events $a_1, a_2, \ldots$ are pairwise incompatible, then

$$P \left( \bigcup_i a_i \right) = P(a_1) + P(a_2) + \cdots = \sum_i P(a_i).$$
Axioms of Probability

The probability of an event, $a$, is denoted $P(a)$. Probabilities obey the following axioms:

1. $0 \leq P(a) \leq 1$, with $P(a) = 1$ if $a$ is certain.
2. If events $a$ and $b$ are mutually exclusive, then

\[ P(a \text{ or } b) = P(a \lor b) = P(a) + P(b). \]

In general, if events $a_1, a_2, \ldots$ are pairwise incompatible, then

\[ P \left( \bigcup_i a_i \right) = P(a_1) + P(a_2) + \cdots = \sum_i P(a_i). \]

3. Joint probability: $P(a \text{ and } b) \equiv P(a, b) = P(b \mid a)P(a)$. 

Reykjavik University, April/May 2005, BNs and DGs: Bayesian Probability Theory
The basic concept in the Bayesian treatment of uncertainty in causal networks is *conditional probability*. 
The basic concept in the Bayesian treatment of uncertainty in causal networks is *conditional probability*.

Every probability is conditioned on a context. For example,

\[ P(\text{six}) = \frac{1}{6} \equiv P(\text{six} | \text{SymmetrixDie}) = \frac{1}{6} \]
The basic concept in the Bayesian treatment of uncertainty in causal networks is *conditional probability*.

Every probability is conditioned on a context. For example,

\[
\text{“} P(\text{six}) = \frac{1}{6} \text{” } \equiv \text{“} P(\text{six} \mid \text{SymmetrixDie}) = \frac{1}{6} \text{”}
\]

In general, given the event \( b \), the conditional probability of the event \( a \) is \( x \):

\[
P(a \mid b) = x.
\]

It is *not* “whenever \( b \) we have \( P(a) = x \)”. 
The basic concept in the Bayesian treatment of uncertainty in causal networks is \emph{conditional probability}.

Every probability is conditioned on a context. For example,

\[ P(\text{six}) = \frac{1}{6} \equiv P(\text{six} | \text{SymmetrixDie}) = \frac{1}{6} \]

In general, given the event \( b \), the conditional probability of the event \( a \) is \( x \): \( P(a | b) = x \).

It is \emph{not} “whenever \( b \) we have \( P(a) = x \)”.

\textbf{Conditional Probability} \\
If \( b \) is true and everything else known is irrelevant for \( a \), then the probability of \( a \) is \( x \).
A Justification of Axiom 3 — The Fundamental Rule

\( C \): a set of cats

\( B \): the subset of brown cats (\( m \))

\( A \): the subset of Abyssinians (\( i \) of them are brown)

\[
f(A|B, C) = \frac{i}{m}, \quad f(B|C) = \frac{m}{n}
\]
A Justification of Axiom 3 — The Fundamental Rule

\( C \): a set of cats

\( B \): the subset of brown cats (\( m \))

\( A \): the subset of Abyssinians (\( i \) of them are brown)

\[
f(A|B, C) = \frac{i}{m}, \quad f(B|C) = \frac{m}{n}
\]

\[
f(A, B|C) = \frac{i}{n} = \frac{i}{m} \cdot \frac{m}{n} = f(A|B, C) \cdot f(B|C).
\]
A discrete random variable, $A$, has a set of exhaustive and mutually exclusive states, $\text{dom}(A) = \{a_1, \ldots, a_n\}$. 
A discrete random variable, $A$, has a set of exhaustive and mutually exclusive states, $\text{dom}(A) = \{a_1, \ldots, a_n\}$.

In this context, an event is an assignment of values to a set of variables and

$$P(A = a_1 \lor \cdots \lor A = a_n) = P(A = a_1) + \cdots + P(A = a_n) = \sum_{i=1}^{n} P(A = a_i) = 1.$$
Discrete Random Variables

- A discrete random variable, $A$, has a set of exhaustive and mutually exclusive states, $\text{dom}(A) = \{a_1, \ldots, a_n\}$.
- In this context, an event is an assignment of values to a set of variables and

$$P(A = a_1 \lor \cdots \lor A = a_n) = P(A = a_1) + \cdots + P(A = a_n)$$

$$= \sum_{i=1}^{n} P(A = a_i) = 1.$$ 

- Capital letters will denote a variable, or a set of variables, and lower case letters will denote states (values) of variables.
- Example: $R = r, B = \neg b$ (Rain? = Raining, BirdsOnRoof = No) is an event.
If $X$ is a variable with states $x_1, \ldots, x_n$, then $P(X)$ denotes a probability distribution over these states:

$$P(X) = (P(X = x_1), \ldots, P(X = x_n)),$$

where

$$P(X = x_i) \geq 0 \text{ and } \sum_{i=1}^{n} P(X = x_i) = 1.$$

$P(X | \text{pa}(X))$ consists of one $P(X)$ for each configuration of the parents, $\text{pa}(X)$, of $X$.

\begin{table}
\begin{tabular}{ c | c c c c }
\hline
 & $B = n$ & $B = n$ & $B = y$ & $B = y$ \\
\hline
$A = n$ & 0.999 & 0.1 & 0.05 & 0.01 \\
$A = y$ & 0.001 & 0.9 & 0.95 & 0.99 \\
\hline
\end{tabular}
\end{table}
Rule of Total Probability

\[ P(A) = P(A, B = b_1) + \cdots + P(A, B = b_n) \]
\[ = P(A | B = b_1)P(B = b_1) + \cdots + P(A | B = b_n)P(B = b_n). \]
Rule of Total Probability

\[
P(A) = P(A, B = b_1) + \cdots + P(A, B = b_n) \\
= P(A | B = b_1)P(B = b_1) + \cdots + P(A | B = b_n)P(B = b_n).
\]

Computing \(P(A)\) from \(P(A, B)\) using the rule of total probability is often called *marginalization*, and is written compactly as

\[
P(A) = \sum_i P(A, B = b_i),
\]

or even shorter as

\[
P(A) = \sum_B P(A, B).
\]
The fundamental rule of probability calculus on variables:

\[ P(X, Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X). \]
The fundamental rule of probability calculus on variables:

\[ P(X, Y) = P(X | Y)P(Y) = P(Y | X)P(X). \]

Bayes’ rule:

\[
P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)} = \frac{P(X | Y)P(Y)}{P(X | Y = y_1)P(Y = y_1) + \cdots + P(X | Y = y_n)P(Y = y_n)}.
\]
Graphically, inference using Bayes’ rule corresponds to reversing arrows:

\[
P(A, B) = P(A)P(B \mid A) \quad \quad P(A, B) = P(B)P(A \mid B)
\]

\[
P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(A)P(B \mid A)}{\sum_{a \in \text{dom}(A)} P(A = a)P(B \mid A = a)}
\]
Let \( V = \{X_1, \ldots, X_n\} \) be a set of variables
Let \( P(V) \) denote the joint probability distribution over \( V \)
Using the Fundamental Rule, \( P(V) \) can be written as

\[
P(V) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\]

Thus, any joint can be represented as a product of conditionals, e.g.,

\[
P(X_1, X_2, X_3) = P(X_3 | X_1, X_2)P(X_2, X_1)
\]

\[
= P(X_3 | X_1, X_2)P(X_2 | X_1)P(X_1)
\]
Let $V = \{A, B, C, D\}$. Then $P(V)$ factorizes as

Let $V = \{A, B, C, D\}$. Then $P(V)$ factorizes as

\[
\]

\[
= P(A|B, C, D)P(B|C, D)P(C, D)
\]
Let \( V = \{A, B, C, D\} \). Then \( P(V) \) factorizes as

\[
\]
Let $V = \{A, B, C, D\}$. Then $P(V)$ factorizes as

$$
$$

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

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

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

(1)

(2)
Let \( V = \{A, B, C, D\} \). Then \( P(V) \) factorizes as

\[
\]
\[
= P(A | B, C, D)P(B | C, D)P(C, D)
\]
\[
= P(A | B, C, D)P(B | C, D)P(C | D)P(D)
\]
\[
= P(B | A, C, D)P(D | A, C)P(C | A)P(A)
\]
\[
= \ldots
\]

(1)

(2)

etc.
Combination and Marginalization

- *Combination* of probability distributions is multiplication.
Combination and Marginalization

- **Combination** of probability distributions is multiplication.
- Using the Rule of Total Probability, from $P(X, Y)$ the *marginal* probability distribution $P(X)$ can be computed:

$$P(x) = P(X = x) = \sum_{y \in \text{dom}(Y)} P(X = x, Y = y).$$
Combination and Marginalization

- *Combination* of probability distributions is multiplication.
- Using the Rule of Total Probability, from $P(X, Y)$ the *marginal* probability distribution $P(X)$ can be computed:

$$P(x) = P(X = x) = \sum_{y \in \text{dom}(Y)} P(X = x, Y = y).$$

- A variable $Y$ is *marginalized out* of $P(X, Y)$ as

$$P(X) = \sum_{y \in \text{dom}(Y)} P(X, Y = y) \text{ or short: } P(X) = \sum_Y P(X, Y).$$
Combination and Marginalization

- **Combination** of probability distributions is multiplication.
- Using the Rule of Total Probability, from $P(X, Y)$ the *marginal* probability distribution $P(X)$ can be computed:

$$P(x) = P(X = x) = \sum_{y \in \text{dom}(Y)} P(X = x, Y = y).$$

- A variable $Y$ is *marginalized out* of $P(X, Y)$ as

$$P(X) = \sum_{y \in \text{dom}(Y)} P(X, Y = y) \text{ or short: } P(X) = \sum_{Y} P(X, Y).$$

- The unity rule:

$$\sum_{X} P(X \mid \text{pa}(X)) = 1_{\text{pa}(X)}$$

Reykjavik University, April/May 2005, BNs and DGs: Bayesian Probability Theory
Combination:

<table>
<thead>
<tr>
<th></th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

$\times \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

<table>
<thead>
<tr>
<th></th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>0.12</td>
<td>0.14</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0.15</td>
<td>0.42</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0.03</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Combination and Marginalization — Example

Combination:

<table>
<thead>
<tr>
<th></th>
<th>$b_1$</th>
<th>$b_2$</th>
<th></th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>0.4</td>
<td>0.2</td>
<td>$a_1$</td>
<td>0.12</td>
<td>0.14</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0.5</td>
<td>0.6</td>
<td>$a_2$</td>
<td>0.15</td>
<td>0.42</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0.1</td>
<td>0.2</td>
<td>$a_3$</td>
<td>0.03</td>
<td>0.14</td>
</tr>
</tbody>
</table>

$\times \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

Marginalization:

$P(A) = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$

$P(A) = \begin{bmatrix} 0.12 + 0.14 \\ 0.15 + 0.42 \\ 0.03 + 0.14 \end{bmatrix} = (0.26, 0.57, 0.17)$
A variable $X$ is independent of $Y$ given $Z$ if

\[ P(x_i | y_j, z_k) = P(x_i | z_k), \quad \forall i, j, k \]

Shorthand:

\[ P(X | Y, Z) = P(X | Z) \]

Notice that the definition is symmetric

Notation: $X \perp \perp Y | Z$
A variable $X$ is *independent* of $Y$ *given* $Z$ if

$$P(x_i | y_j, z_k) = P(x_i | z_k), \ \forall i, j, k$$

**Shorthand:**

$$P(X | Y, Z) = P(X | Z)$$

Notice that the definition is symmetric

**Notation:** $X \perp \!\!\!\!\!\!\!\perp Y | Z$

Under $X \perp \!\!\!\!\!\!\!\perp Y | Z$ the Fundamental Rule reduces to:

$$P(X, Y | Z) = P(X | Y, Z)P(Y | Z)$$

$$= P(X | Z)P(Y | Z)$$
Graphical Representations of $X \independent Y \mid Z$

$X$ and $Y$ are conditionally independent given $Z$: 

$$P(X, Y, Z) = P(X \mid Y, Z)P(Y \mid Z)P(Z)$$

$$= P(X \mid Z)P(Y \mid Z)P(Z)$$
Graphical Representations of $X \perp \perp Y \mid Z$

$X$ and $Y$ are conditionally independent given $Z$:

\[
P(X, Y, Z) = P(X \mid Y, Z)P(Y \mid Z)P(Z) = P(X \mid Z)P(Y \mid Z)P(Z)
\]

\[
P(X, Y, Z) = P(X)P(Y \mid X, Z)P(Z \mid X) = P(X)P(Y \mid Z)P(Z \mid X)
\]
X and Y are conditionally independent given Z:

\[
\]

\[
P(X, Y, Z) = P(X)P(Y|X, Z)P(Z|X) = P(X)P(Y|Z)P(Z|X)
\]

\[
\]
Definition 1

An instantiation of a variable $X$ is an observation on the exact state of $X$.

$$f(X) = (0, \ldots, 0, 1, 0, \ldots, 0)$$
Definition 1
An instantiation of a variable $X$ is an observation on the exact state of $X$.

\[ f(X) = (0, \ldots, 0, 1, 0, \ldots, 0) \]

Definition 2
Let $X$ be a variable with $n$ states. An evidence function on $X$ is an $n$-dimensional table of zeros and ones.

\[ f(X) = (0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0) \]
Definition 1
An instantiation of a variable $X$ is an observation on the exact state of $X$.

$$f(X) = (0, \ldots, 0, 1, 0, \ldots, 0)$$

Definition 2
Let $X$ be a variable with $n$ states. An evidence function on $X$ is an $n$-dimensional table of zeros and ones.

$$f(X) = (0, \ldots, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0)$$

Definition 3
Let $X$ be a variable with $n$ states. An evidence function (likelihood evidence) on $X$ is an $n$-dimensional table of non-negative numbers.

$$f(X) = (0, \ldots, 0, 2, 0, \ldots, 0, 1, 0, \ldots, 0)$$
Let $U$ be a set of variables and let $\{X_1, \ldots, X_n\}$ be the subset of $U$.

Given a set of evidence $\varepsilon = \{\varepsilon_1, \ldots, \varepsilon_m\}$ we want to compute $P(X_i | \varepsilon)$, for all $i$.

This can be done via

$$P(X_i | \varepsilon) = \frac{\sum_{X \in U \setminus \{x_i\}} P(U, \varepsilon)}{\sum_{U} P(U, \varepsilon)} = \frac{P(X_i, \varepsilon)}{P(\varepsilon)}.$$ 

This requires the full joint $P(U)$. 

Summary

- Axioms of probability theory.
- Conditional probabilities.
- The fundamental rule and Bayes’ rule.
- Probability calculus.
  - Fundamental Rule.
  - Bayes’ Rule.
  - Combination and marginalization.
  - The chain rule.
- Conditional independence.
- Evidence.