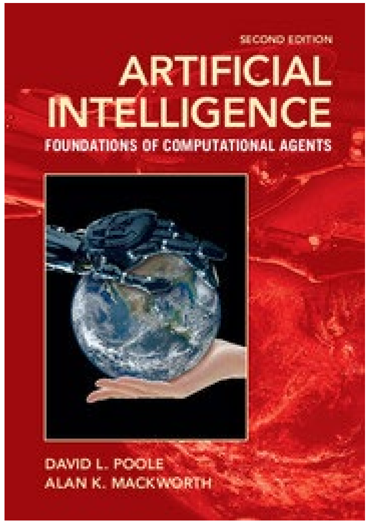
Chapter 8
Reasoning Under Uncertainty – Part I

Main Textbook: ***Artificial Intelligence Foundations of Computational Agents***, 2nd Edition, David L. Poole and Alan K Mackworth, Cambridge University Press, 2018.



<u>Reference Textbook</u>: ***Artificial Intelligence: A Guide to Intelligence Systems***, Michael Negnevitsky, 3rd Edition, 2011, Addison Wesley, ISBN 978-1408225745

Introduction

- In some situations, the agents are forced to make decisions based on **incomplete information**.
- Even when an agent senses the world to find out more information, it rarely finds out the **exact state** of the world
  - For example, *a doctor does not know exactly what is happening inside a patient, a teacher does not know precisely what a student understands*, etc.
- This topic considers **reasoning with the uncertainty** that arises whenever an agent is not omniscient
  - describes how **probability theory** can represent the world by making appropriate independent assumptions and shows how to reason.

Probability

- Reasoning with **uncertainty** has been famous in **probability theory** and **decision theory**.
- When an agent makes decisions and is **uncertain** about the outcomes of its actions, it is gambling on the outcomes
  - We have learned the **probability** of tossing ***coins*** and ***rolling dice***.
- In general**,** the **probability is a calculus for belief** designed for making decisions
  - **Probability theory** is the study of **how knowledge affects belief** and is measured in terms of a number between **0** and **1**
    - The probability of **belief *α*** is **0** means that ***α*** is ***false***, and the probability of ***α*** is **one** means that ***α*** is ***true***.

Probability

- **Probability** is a measure of **belief,** and the **belief** needs to be updated when **new evidence** is observed.
- If an agent's probability of belief ***α*** is greater than **0** and less than **1**, this does not mean that ***α*** is **true** to some degree.
- The view of probability as a **measure of belief** is known **as Bayesian probability** or **subjective probability**.
  – **Uncertainty** in the world is **epistemological** – about an **agent's beliefs** of the world, rather than **ontological** – how the world is.

Semantics of Probability

- **Probability theory** is built on the foundation of **worlds** and **variables**
  - **Variables** could be described in terms of **worlds**: *a* ***variable*** *is a function from worlds into the domain of the variable.*
- **Variables** will be written starting with an **uppercase letter**.
- Each variable has a **domain** which is the set of **values**
  - A **Boolean variable** is a variable with the **domain {*true*, *false*}**.
  - A **discrete variable** has a domain with **a finite** set
    - For example, a world could contain ***symptoms***, ***diseases,*** and ***test results***.
    - We might be able to answer questions about the **probability** that a patient with a particular combination of symptoms may come into the hospital again soon.

Semantics of Probability

- We first define a **probability** over **finite sets of worlds** with **finite variables** and use this to define the probability of **propositions**.
- A **probability measure** is a function ***P*** from a set of worlds ***w*** into the non-negative real numbers such that,

$$\sum_{w \in \Omega} p(w) = 1$$

  - The **probability of proposition *α***, written ***P*(*α*)**, is the **sum of the probabilities** of possible worlds in which ***α*** is **true**.
  - where ***Ω*** is the set of all possible worlds.
  - The use of 1 as the probability of the set of all of the worlds $\{w_1, w_2, \ldots w_n\}$ is just by convention.

Semantics of Probability

Example 8.2 Consider the ten worlds of Figure 8.1, with Boolean variable *Filled*, and with variable *Shape* with domain $\{circle, triangle, star\}$. Each world is defined by its shape, whether it's filled and its position. Suppose the probability of each of these 10 worlds is 0.1, and any other worlds have probability 0. Then $P(Shape{=}circle) = 0.5$ and $P(Filled{=}false) = 0.4$.

$P(Shape{=}circle \wedge Filled{=}false) = 0.1$

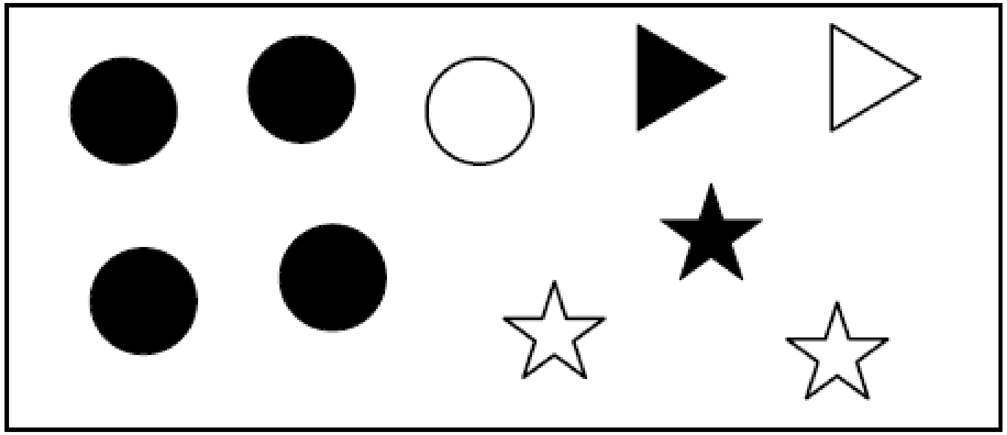


Figure 8.1: Ten worlds described by variables *Filled* and *Shape*

Semantics of Probability

- If ***X*** is a **random variable**, a **probability distribution**, ***P*(*X*)**, over ***X*** is a function from the domain of ***X*** into the real numbers, given a value ***x*** ∈ **domain(*X*)**, ***P*(*x*)** is the **probability** of the proposition ***X*** = ***x***.
- A **probability distribution** over a ***set of variables*** is a function from the values of those variables into a probability
  – For example**, *P*(*X*, *Y*)** is a probability distribution over ***X*** and ***Y*** such that ***P***(***X***=***x***, ***Y***=***y***), where ***x*** ∈ **domain(*X*)** and ***y*** ∈ **domain(*Y*)**, has the value **P(*X*=*x* ∧ *Y*=*y*)**, where ***X*=*x* ∧ *Y*=*y*** is a proposition and ***P*** is the function on propositions defined above.
- If $(\boldsymbol{X_1} \ldots \boldsymbol{X_n})$ are **random variables**, the probability distribution over all worlds, $\boldsymbol{P}(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$, is called the **joint probability distribution**.

Axioms for Probability

- An **axiomatic definition** specifies axioms.
- Suppose ***P*** is a function from propositions into real numbers that satisfies the following **three axioms of probability**:
  - **Axiom1:** **0 ≤ *P*(*α*)** for any proposition ***α***. That is, the **belief** in any proposition **cannot be negative**.
  - **Axiom2:** ***P***(***τ***) **= 1** if ***τ*** is a **tautology (*τ*** is **true** in all possible worlds) its probability is **1**.
  - **Axiom3:** ***P***(***α*** ∨ ***β***) = ***P***(***α***) **+ *P***(***β***) if ***α*** and ***β*** are **contradictory propositions**; In other words, if two propositions **cannot both be true** (they are ***mutually exclusive***), the probability of their disjunction is the sum of their probabilities.
- These axioms form a **sound** and **complete** axiomatization of the meaning of probability.

Axioms for Probability

- **Proposition 8.1:** If there are a finite number of discrete random variables, Axioms 1, 2, and 3 are **sound** and **complete** with respect to the semantics.
- **Proposition 8.2:**The following hold for all propositions ***α*** and ***β***
  - (a) **Negation of a proposition: $P(\neg \alpha) = 1 - P(\alpha)$.**
  - (b) If $\alpha \leftrightarrow \beta$, then **$P(\alpha) = P(\beta)$.** That is, **logically equivalent propositions have the same probability.**
  - (c) **Reasoning by cases: $P(\alpha) = P(\alpha \wedge \beta) + P(\alpha \wedge \neg \beta)$.**
  - (d) If ***V*** is a **random variable** with domain ***D***, then, for all propositions ***α***,

$$p(\alpha) = \sum_{d \in D} p(\alpha \wedge V = d)$$

  - (e) **Disjunction for non-*mutually exclusive*** propositions: **$P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$.**

Axioms for Probability

- **<u>Proof. (a)</u>:** The propositions $\alpha \vee \neg \alpha$ and $\neg(\alpha \wedge \neg \alpha)$ are **tautologies**. Therefore, $\mathbf{1} = P(\alpha \vee \neg \alpha) = P(\alpha) + P(\neg \alpha)$. Rearranging gives the desired result.
- **<u>Proof. (b)</u>:** If $\alpha \leftrightarrow \beta$, then $\alpha \vee \neg \beta$ is a **tautology**, so $P(\alpha \vee \neg \beta) = 1$. α and $\neg \beta$ are **contradictory** statements, so **Axiom3** gives $P(\alpha \vee \neg \beta) = P(\alpha) + P(\neg \beta)$. Using part (a), $P(\neg \beta) = 1 - P(\beta)$. Thus, $P(\alpha) + 1 - P(\beta) = 1$, and so $P(\alpha) = P(\beta)$.
- **<u>Proof. (c)</u>:** The proposition $\alpha \leftrightarrow ((\alpha \wedge \beta) \vee (\alpha \wedge \neg \beta))$ and $\neg((\alpha \wedge \beta) \wedge (\alpha \wedge \neg \beta))$ are **tautologies**. Thus, $P(\alpha) = P((\alpha \wedge \beta) \vee (\alpha \wedge \neg \beta)) = P(\alpha \wedge \beta) + P(\alpha \wedge \neg \beta)$.
- **<u>Proof. (d)</u>:** The proof is analogous to the proof of proposition **(c)**.
- **<u>Proof. (e)</u>:** $(\alpha \vee \beta) \leftrightarrow ((\alpha \wedge \neg \beta) \vee \beta)$ is a **tautology**. Thus, $P(\alpha \vee \beta) = P((\alpha \wedge \neg \beta) \vee \beta) = P(\alpha \wedge \neg \beta) + P(\beta)$. Proposition **(c)** shows $P(\alpha \wedge \neg \beta) = P(\alpha) - P(\alpha \wedge \beta)$. Thus, $P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$.

Conditional Probability

- **Probability** is a **measure of belief**.
- The **measure of belief** in proposition ***h*** given proposition **e** is called the **conditional probability** of ***h*** given ***e***, written ***P*(*h*|*e*)**.
  - If **evidence**(***e***) then **hypothesis**(***h***), then ***P*(*h*|*e*) is the probability of *h* in the presence of *e***
  - The proposition **e** representing the **conjunction** of the agent's observations of the world is called **evidence**.
  - Given **evidence e**, the conditional probability ***P*(*h*|e)** is the agent's **posterior probability** of ***h***.
- The **probability *P*(*h*)** is the **prior probability** of ***h*** and is the same as ***P*(*h*|true)**
  - The **evidence** used for the **posterior probability** is ***everything*** the agent observes about a particular situation.

Semantics of Conditional Probability

- **Evidence *e***, where ***e*** is a **proposition,** will rule out all possible worlds incompatible with ***e*** (the proposition ***e*** selects all the possible worlds in which ***e*** is **true**).
  - As in the definition of probability, first define **the conditional probability** over worlds, then use this to define a probability over propositions.
- **Evidence *e*** induces a probability $\boldsymbol{P(w|e)}$ of **world *w*** given **e**. A world where **e is false** has **conditional probability 0**, and the remaining worlds are normalized so that the probabilities of the worlds sum to **1**:

$$P(w|e) = \begin{cases} c \times P(w) & \text{if } e \text{ is } \textit{true} \text{ in world } w \\ 0 & \text{if } e \text{ is } \textit{false} \text{ in world } w \end{cases}$$

  - where ***c* is a constant** (depends on ***e***) that ensures the posterior probability of all worlds sums to **1**.

Semantics of Conditional Probability

where c is a constant (that depends on e) that ensures the posterior probability of all worlds sums to 1.

For $P(w \mid e)$ to be a probability measure over worlds for each e:

$$
\begin{aligned}
1 &= \sum_{w} P(w \mid e) \\
&= \sum_{w\,:\,e \text{ is true in } w} P(w \mid e) + \sum_{w\,:\,e \text{ is false in } w} P(w \mid e) \\
&= \sum_{w\,:\,e \text{ is true in } w} c * P(w) + 0 \\
&= c * P(e)
\end{aligned}
$$

Therefore, $c = 1/P(e)$. Thus, the conditional probability is only defined if $P(e) > 0$. This is reasonable, as if $P(e) = 0$, e is impossible.

Semantics of Conditional Probability

The conditional probability of $\boxed{\text{proposition } h}$ given $\boxed{\text{evidence } e}$ is the sum of the conditional probabilities of the possible worlds in which $\boxed{h \text{ is true}}$. That is,

$$
\begin{aligned}
&= \sum_{w\,:\,h \text{ is true in } w} P(w \mid e) \\
&= \sum_{w\,:\,h \wedge e \text{ is true in } w} P(w \mid e) + \sum_{w\,:\,\neg h \wedge e \text{ is true in } w} P(w \mid e) \\
&= \sum_{w\,:\,h \wedge e \text{ is true in } w} \frac{1}{P(e)} * P(w) + 0
\end{aligned}
$$

$$
\boxed{P(h \mid e) = \frac{P(h \wedge e)}{P(e)}.}
$$

Semantics of Conditional Probability

Example 8.5 As in Example 8.2, consider the worlds of Figure 8.1 (page 346), each with probability 0.1. Given the evidence *Filled*=*false*, only 4 worlds have a non-zero posterior probability. $P(Shape{=}circle \mid Filled{=}false) = 0.25$ and $P(Shape{=}star \mid Filled{=}false) = 0.5$.

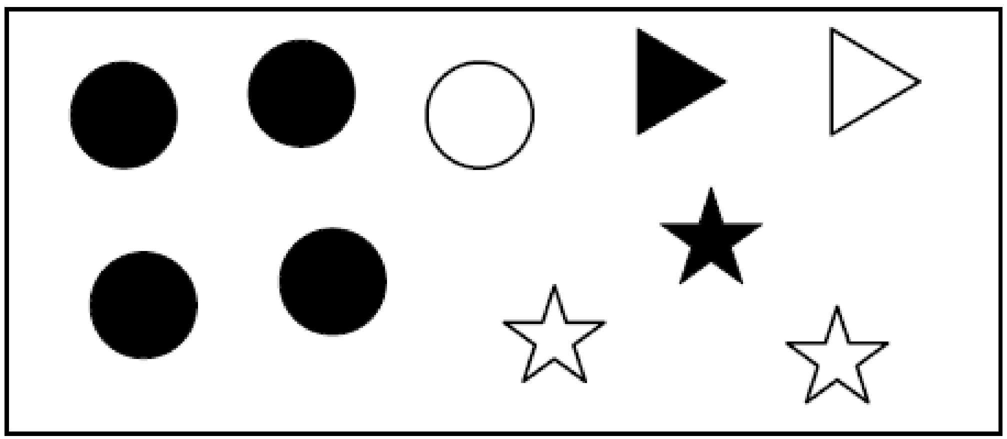


Figure 8.1: Ten worlds described by variables *Filled* and *Shape*

Semantics of Conditional Probability

Proposition 8.3. *(Chain rule) For any propositions* $\alpha_1, \ldots, \alpha_n$:

$$
\begin{aligned}
P(\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n) = & P(\alpha_1) * \\
& P(\alpha_2 \mid \alpha_1) * \\
& P(\alpha_3 \mid \alpha_1 \wedge \alpha_2) * \\
& \vdots \\
& P(\alpha_n \mid \alpha_1 \wedge \cdots \wedge \alpha_{n-1}) \\
= & \boxed{\prod_{i=1}^{n} P(\alpha_i \mid \alpha_1 \wedge \cdots \wedge \alpha_{i-1}),}
\end{aligned}
$$

where the right-hand side is assumed to be zero if any of the products are zero (even if some of them are undefined).

Axioms of probability: Summary

1. **$0 \leq P(A) \leq 1$** ; for every $A \subseteq S$ (S is the given finite sample space)
2. **Boundary:** $P(\phi) = 0$ and $P(S) = 1$
3. **Monotonic:** if $A \subseteq B \subseteq S$, then $p(A) \leq p(B)$
4. **Inclusion–exclusion:**

   If ***A*** and ***B*** are **mutually inclusive**, $P(A \vee B) = P(A)+P(B)-P(A \wedge B)$

   If ***A*** and ***B*** are **mutually exclusive**, then $P(A \vee B) = P(A) + P(B)$

5. **Intersection:** $P(A \wedge B) = P(A)^* P(B|A)$, where A and B are *true*
6. **Negation**: $P(\sim A) = 1 - P(A)$, this is a generalization of the fact that A is true if and only if $\sim A$ is false and vice versa
7. **Equivalence:** If $A \equiv B$, then $P(A) = P(B)$

   (Assume A and B are two propositions)

Bayes' Rule

- Let ***A*** and ***B*** are two events in the world, suppose that ***A*** and ***B*** are **<u>not mutually exclusive</u>** (they **occur conditionally**):
  - The probability that event ***A*** will occur if event ***B*** occurs is called the **conditional probability**
  - **Conditional probability** is denoted as ***p(A|B)***, interpreted as '***conditional probability of event A occurring given that event B has occurred***'

$$p(A|B) = \frac{\text{the no. of times } A \text{ and } B \text{ occur}}{\text{the no. of times } B \text{ occur}} \qquad (1)$$

  - The no. of times ***A*** and ***B*** can occur, or the probability that both ***A*** and ***B*** will occur is called the **joint probability** of ***A*** and ***B*** *is* represented as $p(A \cap B)$

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \qquad (2)$$

$$p(B|A) = \frac{p(B \cap A)}{p(A)} \qquad (3)$$

Bayes' Rule

- From (**2**) and (**3**) we get

  $$p(A \cap B) = p(A|B) \times p(B) \text{ and } p(B \cap A) = p(B|A) \times p(A) \quad \textbf{(4)}$$

  Joint probability is commutative, thus

  $$p(A \cap B) = p(B \cap A), \text{ therefore}$$

  $$p(A|B) \times p(B) = p(B|A) \times p(A) \quad \textbf{(5)}$$

  Eq. **(5)** yields the following equation:

  $$p(A|B) = \frac{p(B|A) \times p(A)}{p(B)} \quad \textbf{(6)}$$

- The Eq. **(6)** is known as **Bayes' rule**

  Where:
  - ***p(A|B)*** is the conditional probability that event ***A*** occurs given that event ***B*** has occurred
  - ***p(B|A)*** is the conditional probability of event ***B*** occurring given that event ***A*** has occurred
  - ***p(A)*** is the prior probability of event ***A***
  - ***p(B)*** is the prior probability of event ***B***

Bayes' Rule: Proof

- Assume that the event ***A*** is dependent upon event ***B*** (the **joint probability** of events ***A*** and ***B*** is shown in **Figure 3.1**)
  - The event ***A*** depends on a number of events $\boldsymbol{B_1}$, $\boldsymbol{B_2}$,…., $\boldsymbol{B_n}$

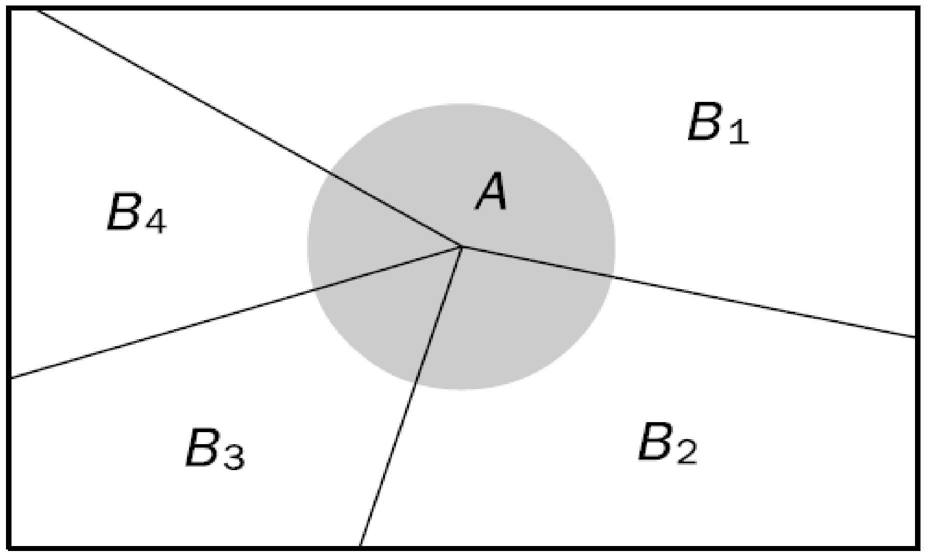


Figure 3.1 The joint probability

Bayes' Rule: Proof

- From **(4)** based on **Figure 3.1**

$$p(A \cap B_1) = p(A|B_1) \times p(B_1)$$
$$p(A \cap B_2) = p(A|B_2) \times p(B_2)$$
$$\vdots \qquad\qquad \vdots$$
$$p(A \cap B_n) = p(A|B_n) \times p(B_n)$$

We combined the above sequences to get ***p*(*A*)** and is given as:

$$\boxed{\sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(A \mid B_i) * p(B_i) = p(A)} \qquad (\mathbf{7})$$

Equation (**7**) is called **law of total probability.**

From **figure 3.1**, if the occurrence of event ***A*** depends on **only two mutually exclusive events**, ***B*** *and* **¬*B*,** then (**7**) becomes

$$\boxed{p(A) = p(A|B) \times p(B) + p(A|\neg B) \times p(\neg B)} \qquad (\mathbf{8})$$

Bayes' Rule: Proof

Similarly, if the occurrence of ***B*** depends on events, ***A*** and $\neg$ ***A*,** then

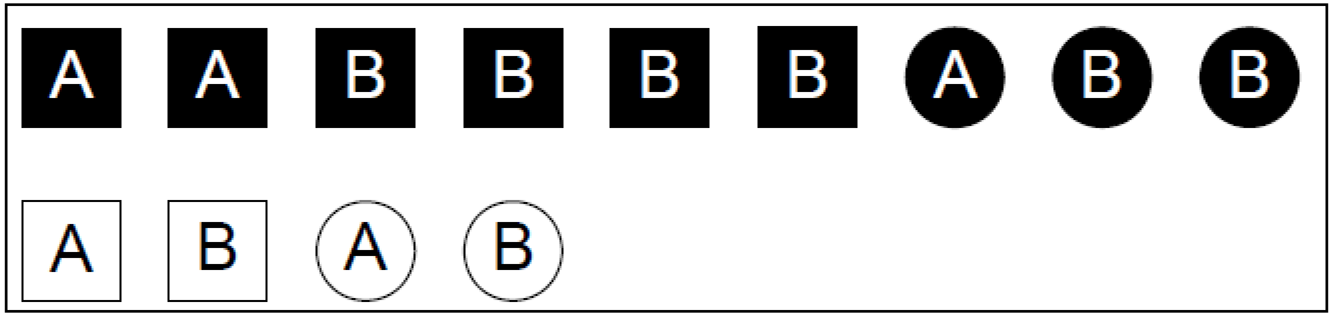
$$p(B) = p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A) \qquad (9)$$

Substitute (**9**) into **Baye's rule** (**6**) to yield

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)} \qquad (10)$$

- Eq. (**10**) provides the background for the **application of probability theory to mange uncertainty**.

Bayes' Rule



Let ***S*** be the set of all objects in the Figure shown above. Let **Black** be the set of all **black objects**, White be the set of all **white objects**, **Square** be the set of all **square objects**, ***A*** be the set of all objects containing an ***A***, and ***B*** be the set of all objects containing a ***B***. We then have that

$\boldsymbol{P}(\mathbf{A|White}) = 2/4 = 1/2$, $\boldsymbol{P}(\mathbf{Black}) = 9/13$, $\boldsymbol{P}(\mathbf{A|Black}) = 3/9 = 1/3$, $\boldsymbol{P}(\mathbf{White}) = 4/13$, $\boldsymbol{P}(\mathbf{A|Square}) = 3/8$, $\boldsymbol{P}(\mathbf{A|Square} \cap \mathbf{Black}) = 2/6 = 1/3$, $\boldsymbol{P}(\mathbf{A|Square} \cap \mathbf{White}) = 1/2$

- From the above values, ***P*(A)** can be <u>Calculated based on Eq. **(8):**</u>

$\boldsymbol{P}(\mathbf{A}) = \boldsymbol{P}(\mathbf{A|Black}) \times \boldsymbol{P}(\mathbf{Black}) + \boldsymbol{P}(\mathbf{A|White}) \times \boldsymbol{P}(\mathbf{White})$

$= (1/3 \times 9/13) + (1/2 \times 4/13)$

$= \mathbf{5/13}$

Bayes' Rule

- An agent updates its **belief** using **probability** when it observes a **new evidence**
  - The **new evidence** is conjoined to the **old evidence** to form complete evidence.
- **Bayes' rule** specifies how an agent should **update its belief** based on new evidence:
  - Suppose an agent has a **current belief *h*** (called a **hypothesis**) based on **evidence *k***, can be given by $P(h|k)$, and can subsequently add a **new evidence *e*** in the observation: $P(h|(e \wedge k))$.
  - **Bayes' rule** tells us how to update the agent's **belief** in **hypothesis *h*** as **new evidence** arrives.

Bayes' Rule

Proposition 8.4. *(**Bayes' rule**) As long as* $P(e \mid k) \neq 0$,

$$\boxed{P(h \mid e \wedge k) = \frac{P(e \mid h \wedge k) * P(h \mid k)}{P(e \mid k)}.}$$

This is often written with the background knowledge k implicit. In this case, if $P(e) \neq 0$, then

$$\boxed{P(h \mid e) = \frac{P(e \mid h) * P(h)}{P(e)}.}$$

$P(e \mid h)$ is the **likelihood** and $P(h)$ is the **prior probability** of the hypothesis h. Bayes' rule states that the **posterior probability** is proportional to the likelihood times the prior.

Bayes' Rule

Proof. The commutativity of conjunction means that $h \wedge e$ is equivalent to $e \wedge h$, and so they have the same probability given k. Using the rule for multiplication in two different ways,

1). $P(h \wedge e) = P(h|e) * P(e)$ After adding event ***k*** on (***h*** $\wedge$ **e**):

$$\boxed{P(h \wedge e \mid k) = P(h \mid e \wedge k) * P(e \mid k)}$$

2). $P(e \wedge h) = P(e|h) * P(h)$ After adding event ***k*** on (**e** $\wedge$ ***h***):

$$\boxed{P(e \wedge h \mid k) = P(e \mid h \wedge k) * P(h \mid k)}$$

Bayes' Rule

<u>**Example 8.8**</u> Suppose an agent has information about the reliability of fire alarms. It may know how likely it is that an alarm will work if there is a fire. To determine the probability that there is a fire, given that there is an alarm, Bayes' rule gives:

$$\boxed{P(fire \mid alarm) = \frac{P(alarm \mid fire) * P(fire)}{P(alarm)}}$$

$$\boxed{= \frac{P(alarm \mid fire) * P(fire)}{P(alarm \mid fire) * P(fire) + P(alarm \mid \neg fire) * P(\neg fire)}}$$

where $\boxed{P(alarm \mid fire)}$ is the probability that the alarm worked, assuming that there was a fire. It is a measure of the alarm's reliability. The expression $\boxed{P(fire)}$ is the probability of a fire given no other information. It is a measure of how fire-prone the building is. $\boxed{P(alarm)}$ is the probability of the alarm sounding, given no other information. $\boxed{P(fire \mid alarm)}$ is more difficult to directly represent because it depends, for example, on how much vandalism there is in the neighborhood.

Bayes' Rule

- Suppose that the rules in a **Knowledge Base (KB)** are represented in the following form:

IF	**Evidence *E*** is ***True***
THEN	**Hypothesis *H*** is ***True*** { *with a probability of* ***p*** }

– What if event ***E*** has occurred but **do not know whether event *H* has occurred**? Can we compute the probability that **event *H* has occurred as well**?

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E|H) \times p(H) + p(E|\neg H) \times p(\neg H)} \qquad (11)$$

- Where $p(H|E)$ is the probability of the hypothesis ***H*** in the presence of evidence ***E***

Bayes' Rule

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E|H) \times p(H) + p(E|\neg H) \times p(\neg H)}$$

- $p(H)$ is the prior probability of hypothesis H being **true**
- $p(E|H)$ is the probability that hypothesis H being **true** will result E
- $p(\neg H)$ is the prior probability of hypothesis H being **false**
- $p(E|\neg H)$ is the probability of finding evidence E even when hypothesis H is f**alse**

- In a **knowledge based system**, the probabilities required to solve a problem are provided by **experts**
  - An e**xpert** determines the prior probabilities $p(H)$ and $p(\neg H)$ and observing evidence E if H is true, $p(E|H)$ and if hypothesis H is false, $p(E|\neg H)$
  - The system computes $p(H|E)$ for H in the light of evidence E
- $p(H|E)$ is the **posterior probability** of H upon evidence E

Bayes' Rule

- Generalize the **Eq.(11)** with multiple hypotheses $\boldsymbol{H_1}$, $\boldsymbol{H_2}$,…$\boldsymbol{H_m}$ and multiple evidences $\boldsymbol{E_1}$, $\boldsymbol{E_2}$,…,$\boldsymbol{E_n}$ (but the *hypotheses* and *evidences* must be **mutually exclusive**)
- Single evidence $\boldsymbol{E}$ and multiple hypotheses $\boldsymbol{H_1}$, $\boldsymbol{H_2}$,…$\boldsymbol{H_m}$ follow:

$$p(H_i|E) = \frac{p(E|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E|H_k) \times p(H_k)} \qquad \textbf{(12)}$$

- Multiple evidences $\boldsymbol{E_1}$, $\boldsymbol{E_2}$,…,$\boldsymbol{E_n}$ and multiple hypothesis $\boldsymbol{H_1}$, $\boldsymbol{H_2}$,…$\boldsymbol{H_m}$ follow:

$$p(H_i|E_1E_2\ldots E_n) = \frac{p(E_1E_2\ldots E_n|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E_1E_2\ldots E_n|H_k) \times p(H_k)} \qquad \textbf{(13)}$$

Bayes' Rule

- An application of Eq. **(13)** requires to obtain the **conditional probabilities** of all possible combinations of **evidences** for all hypothesis;

$$p(H_i|E_1E_2\ldots E_n) = \frac{p(E_1|H_i) \times p(E_2|H_i) \times \ldots \times p(E_n|H_i) \times p(H_i)}{\sum_{k=1}^{m} p(E_1|H_k) \times p(E_2|H_k) \times \ldots \times p(E_n|H_k) \times p(H_k)} \quad \textbf{(14)}$$

- How does an **ES** compute all posterior probabilities and finally rank potentially true hypothesis?
  - Suppose an ES, given three conditionally independent evidences $\boldsymbol{E_1}$, $\boldsymbol{E_2}$, and $\boldsymbol{E_3}$ creates three mutually exclusive hypothesis $\boldsymbol{H_1}$, $\boldsymbol{H_2}$,and $\boldsymbol{H_3}$ and provides prior probabilities for these hypothesis- $\boldsymbol{p(H_1)}$, $\boldsymbol{p(H_2)}$ and $\boldsymbol{p(H_3)}$ respectively
- **Table 3.2** illustrates the prior and conditional probabilities provided by the **expert**

Bayes' Rule

Table 3.2 The prior and conditional probabilities

	Hypothesis		
Probability	$i = 1$	$i = 2$	$i = 3$
$p(H_i)$	0.40	0.35	0.25
$p(E_1 \mid H_i)$	0.3	0.8	0.5
$p(E_2 \mid H_i)$	0.9	0.0	0.7
$p(E_3 \mid H_i)$	0.6	0.7	0.9

Bayes' Rule

- Assume that we first observe evidence ***E*$_3$**, based on **Eq.(14)**:

$$p(H_i|E_3) = \frac{p(E_3|H_i) \times p(H_i)}{\sum_{k=1}^{3} p(E_3|H_k) \times p(H_k)}, \quad i = 1,2,3$$

$$\boxed{\sum_{k=1}^{3} p(E_3|H_k) \times p(H_k) = p(E3|H1) \times p(H1) + p(E3|H2) \times p(H2) + p(E3|H3) \times p(H3)}$$

Thus,

$$p(H_1|E_3) = \frac{0.6 \times 0.40}{0.6 \times 0.40 + 0.7 \times 0.35 + 0.9 \times 0.25} = 0.34$$

$$p(H_2|E_3) = \frac{0.7 \times 0.35}{0.6 \times 0.40 + 0.7 \times 0.35 + 0.9 \times 0.25} = 0.34$$

$$p(H_3|E_3) = \frac{0.9 \times 0.25}{0.6 \times 0.40 + 0.7 \times 0.35 + 0.9 \times 0.25} = 0.32$$

Bayes' Rule

- Suppose now that we observe evidence $\boldsymbol{E_1}$ along with $\boldsymbol{E_3}$, the posterior probabilities are calculated as:

$$p(H_i|E_1E_3) = \frac{p(E_1|H_i) \times p(E_3|H_i) \times p(H_i)}{\sum_{k=1}^{3} p(E_1|H_k) \times p(E_3|H_k) \times p(H_k)}, \qquad i = 1, 2, 3$$

Hence,

$$p(H_1|E_1E_3) = \frac{0.3 \times 0.6 \times 0.40}{0.3 \times 0.6 \times 0.40 + 0.8 \times 0.7 \times 0.35 + 0.5 \times 0.9 \times 0.25} = 0.19$$

$$p(H_2|E_1E_3) = \frac{0.8 \times 0.7 \times 0.35}{0.3 \times 0.6 \times 0.40 + 0.8 \times 0.7 \times 0.35 + 0.5 \times 0.9 \times 0.25} = 0.52$$

$$p(H_3|E_1E_3) = \frac{0.5 \times 0.9 \times 0.25}{0.3 \times 0.6 \times 0.40 + 0.8 \times 0.7 \times 0.35 + 0.5 \times 0.9 \times 0.25} = 0.29$$

Hypothesis H_2 is now considered as the most likely one, while belief in hypothesis H_1 has decreased dramatically.

Bayes' Rule

- Along with E_1 and E_3 observing evidence E_2 as well, the ES calculates the final posterior probabilities for all hypotheses:

$$p(H_i|E_1E_2E_3) = \frac{p(E_1|H_i) \times p(E_2|H_i) \times p(E_3|H_i) \times p(H_i)}{\sum_{k=1}^{3} p(E_1|H_k) \times p(E_2|H_k) \times p(E_3|H_k) \times p(H_k)}, \qquad i = 1,2,3$$

Thus,

$$p(H_1|E_1E_2E_3) = \frac{0.3 \times 0.9 \times 0.6 \times 0.40}{0.3 \times 0.9 \times 0.6 \times 0.40 + 0.8 \times 0.0 \times 0.7 \times 0.35 + 0.5 \times 0.7 \times 0.9 \times 0.25}$$
$$= 0.45$$

$$p(H_2|E_1E_2E_3) = \frac{0.8 \times 0.0 \times 0.7 \times 0.35}{0.3 \times 0.9 \times 0.6 \times 0.40 + 0.8 \times 0.0 \times 0.7 \times 0.35 + 0.5 \times 0.7 \times 0.9 \times 0.25}$$
$$= 0$$

$$p(H_3|E_1E_2E_3) = \frac{0.5 \times 0.7 \times 0.9 \times 0.25}{0.3 \times 0.9 \times 0.6 \times 0.40 + 0.8 \times 0.0 \times 0.7 \times 0.35 + 0.5 \times 0.7 \times 0.9 \times 0.25}$$
$$= 0.55$$

Bayes' Rule

- Although the initial ranking provided by the expert was H_1, H_2, and H_3, only hypothesis H_1 and H_3 remain under consideration after all evidences (E_1, E_2 and E_3) were observed.
- Hypothesis H_2 can now be **completely abandoned**.
- **Note:** The hypothesis H_3 is considered more likely than hypothesis H_1.

Exercises

1. Prove $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$ {where A and B are ***not mutually exclusive*** *events*}.
2. Prove that $p(A \rightarrow B) = p(\sim A) + p(B) - p(\sim A \wedge B)$ { where A and B are ***not mutually exclusive*** *events*}.
3. Show, $P(A| B \wedge A) = 1$
4. Consider an incandescent bulb manufacturing unit. Here machines **M1**, **M2** and **M3** make **20%**, **30%** and **50%** of the total bulbs. Of their output, let's assume that **2%, 3%** and **5%** are **defective**. A bulb is drawn at random and is found defective. What is the probability that the bulb is made by machine **M1** or **M2** or **M3**.

Exercises

- Suppose there is a disease randomly found in **one-half of one percent** of the general population. A certain clinical blood test is **99%** effective in detecting the **presence** of this disease; i.e., it will yield an **accurate Positive** result in **99%** of the cases where the disease is **actually present**. But it also yields **false Positive** results in **5%** of the cases where the disease is **not present**. Show the probabilities;
  - a) **The probability that the disease will be present in any particular person.**
  - **b) The probability that the disease will not be present in any particular person.**
  - **c) The probability that the test will yield a Positive result if the disease is present.**
  - **d) The probability that the test will yield a Negative result if the disease is present.**
  - **e) The probability that the test will yield a Positive result if the disease is not present.**
  - **f) The probability that the test will yield a Negative result if the disease is not present.**
  - **g) The probability that the disease is present if the test result is Positive (i.e., the probability that a Positive test result will be a true positive).**
  - **h) The probability that the disease is not present if the test result is Positive (i.e., the probability that a Positive test result will be a False Positive).**
  - **i) The probability that the disease is absent if the test result is Negative (i.e., the probability that a Negative test result will be a true negative).**

The Joint Probability Distribution

- **Probability distribution** gives values for all possible assignments:
  - For example, weather is one of <*sunny*, *rain*, *cloudy*, *snow*>
  - *P*(weather) = <0.72, 0.1, 0.08, 0.1> (normalized, i.e., sums to 1).
- **Joint probability distribution** for a set of variables gives values for each possible assignment to all the variables
  - **What is the probability of having either a cavity or a Toothache?**

– $\boldsymbol{P(Cavity \vee Toothache)}$ **?**

	Toothache=true	Toothache = false
Cavity=true	0.04	0.06
Cavity=false	0.01	0.89

$= P(Cavity \wedge Toothache) + P(Cavity \wedge \neg Toothache) + P(\neg Cavity \wedge Toothache)$

$= 0.04 + 0.06 + 0.01$

$= 0.11$

Then, find $P(Cavity)$? $P(Toothache)$? $P(Cavity|Toothache)$? $P(\sim Cavity|Toothache)$

Information

- The **information theory** discusses how to represent **information** using **bits**.
- For ***x*** **∈ domain(*X*)**, it is possible to build a code that, to identify ***x*** uses $-\log_2 P(x)$ bits. The expected number of bits to transmit a value for ***X*** is then

$$H(X) = \sum_{x \in domain(X)} -P(X=x) * \log_2 P(X=x)$$

- This is the **information content** or **entropy** of random variable ***X***.
  - **<u>Note:</u>** ***H*** is a function of the variable ***X***, not a function of the values of the variable.
  - Thus, for a variable ***X***, the **entropy** ***H*(*X*)** is a number.

Information

The entropy of X given the observation $Y = y$ is

$$\boxed{H(X \mid Y{=}y) = \sum_x -P(X{=}x \mid Y{=}y) * \log_2 P(X{=}x \mid Y{=}y).}$$

Before observing Y, the expectation over Y:

$$\boxed{H(X \mid Y) = \sum_y P(Y{=}y) * \sum_x -P(X{=}x \mid (Y{=}y) * \log_2 P(X{=}x \mid (Y{=}y)}$$

is called **conditional entropy** of X given Y.

For a test that determines the value of Y, the **information gain** from this test is $\boxed{H(X) - H(X \mid Y)}$, which is the number of bits used to describe X minus the expected number of bits to describe X after learning Y. The information gain is never negative.

Information

Example 8.11 Suppose spinning a wheel in a game can produces a number in the set $\{1,2,\ldots,8\}$, each with equal probability. Let S be the outcome of a spin. Then $H(S) = -\sum_{i=1}^{8} \frac{1}{8} * \log_2 \frac{1}{8} = \boxed{3 \text{ bits.}}$

Suppose there is a sensor G that detects whether the outcome is greater than 6. $G{=}true$ if $H > 6$. Then $H(S \mid G) = -0.25 \log_2 \frac{1}{2} - 0.75 \log_2 \frac{1}{6} = \boxed{2.19.}$ The information gain of G is thus $3 - 2.19 = \boxed{0.81 \text{ bits.}}$ A fraction of a bit makes sense in that it is possible to design a code that uses 219 bits to predict 100 outcomes.

For an "even" sensor E, where $E{=}true$ if H is even, $H(S \mid E) = -0.5 \log_2 \frac{1}{4} - 0.5 \log_2 \frac{1}{4} = \boxed{2.}$ The information gain of E is thus 1 bit.

Information

- The notion of information is used for a number of tasks:
  - In **diagnosis**, an agent could choose a test that provides the most information.
  - In **decision tree learning**, information theory provides a useful criterion for choosing which property to split on: split on the property that provides the greatest information gain.
  - In **Bayesian learning**, information theory provides a basis for deciding which is the best model given some data.

Independence

- The **axioms** of probability are very weak and provide few constraints on allowable conditional probabilities.
- A helpful way to **limit the required information** is to assume that each **variable only directly depends** on a few other variables.
- This uses assumptions of **conditional independence**.
- As long as the value of ***P*(*h*|*e*)** is not **0** or **1**, the value of ***P*(*h*|*e*) does not constrain** the value of ***P*(*h*|*f*∧*e*)**.
- This latter probability could have any value in the range **[0, 1]**. It is **1** when ***f*** implies ***h***, and it is **0** if ***f*** implies ¬ ***h***.
- Certain knowledge in ***P*(*h*|*e*)** = ***P*(*h*|*f*∧*e*)** specifies ***f* is irrelevant (*f* is 1)** to the probability of ***h*** given that **e** is observed.

Independence

Random variable X is **conditionally independent** of random variable Y **given** a set of random variables Zs if

$$\boxed{P(X \mid Y, Zs) = P(X \mid Zs)}$$

whenever the probabilities are well defined. This means that for all $x \in domain(X)$, for all $y \in domain(Y)$, and for all $z \in domain(Zs)$, if $P(Y = y \wedge Zs = z) > 0$,

$$\boxed{P(X = x \mid Y = y \wedge Zs = z) = P(X = x \mid Zs = z).}$$

That is, given a value of each variable in Zs, knowing Y's value does not affect the belief in the value of X.

Belief Networks

- A **belief network** is a directed model of **conditional dependence** among a set of **random variables**.
- The **conditional independence** in a **belief network** takes in an **ordering of the variables**, and results in a **directed graph**.
- To define a **belief network** on a set of ***random variables***, {$\boldsymbol{X_1, \ldots, X_n}$}, first select a total ordering of the variables, say, $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$
  - The **chain rule** shows (see **proposition 8.3**) how to decompose a **conjunction** into **conditional probabilities**:

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

  - Define the **parents** of random variable $\boldsymbol{X_i}$, written **parents(**$\boldsymbol{X_i}$**)**, to be a minimal set of predecessors of $\boldsymbol{X_i}$ in the total ordering such that the other predecessors of $\boldsymbol{X_i}$ **are conditionally independent** of $\boldsymbol{X_i}$ given **parents(**$\boldsymbol{X_i}$**)**.

Chain Rule

Proposition 8.3. *(Chain rule) For any propositions* $\alpha_1, \ldots, \alpha_n$:

$$\begin{aligned} P(\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n) = & P(\alpha_1) * \\ & P(\alpha_2 \mid \alpha_1) * \\ & P(\alpha_3 \mid \alpha_1 \wedge \alpha_2) * \\ & \vdots \\ & P(\alpha_n \mid \alpha_1 \wedge \cdots \wedge \alpha_{n-1}) \\ = & \boxed{\prod_{i=1}^{n} P(\alpha_i \mid \alpha_1 \wedge \cdots \wedge \alpha_{i-1}),} \end{aligned}$$

where the right-hand side is assumed to be zero if any of the products are zero (even if some of them are undefined).

Belief Networks

- Thus X_i probabilistically depends on each of its **parents**, but is **independent** of its **other predecessors**.
- That is, **parents(X_i) ⊆ {$X_1, . . . , X_{i-1}$}** such that **(X_i | $X_1, . . . , X_{i-1}$) = P(X_i | parents(X_i)).** Putting the **chain rule** and the definition of parents together gives: $$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid parent(X_i))$$
  - The **probability** over all of the variables, **$P(X_1, X_2, . . . , X_n)$**, is called the **joint probability distribution**.
- A **belief network** defines a factorization of **the joint probability distribution** into a **product of conditional probabilities:**
  - A **belief network**, also called a **Bayesian network**, is an **acyclic directed graph (DAG)**, where the **nodes are random variables**.
  - a set of **conditional probability distributions** giving **P(X | parents(X))** for each variable ***X***.

Belief Networks

- **<u>Example 8.13</u>** Consider the **four variables** with the ordering:{***Intelligent*, *Works_hard*, *Answers*, *Grade***}. Consider the variables in order. ***Intelligent*** does not have any **parents**; thus **parents(Intelligent) = {}**. Similarly, ***Works_hard*** *is* independent of ***Intelligent***, and it has **no parents**. The corresponding **belief network** is given **in Figure 8.2**. The variable ***Answers*** has two parents, ***Intelligent*** and ***Works_hard***, so ***parents*(*Answers*) =** {***Intelligent*, *Works_hard***}**.**
- ***Grade*** is independent of ***Intelligent*** and ***Works_hard*** *and has a parent*:

  ***parents*(*Grade*) = {*Answers*}.**
- This graph defines the decomposition of **the joint distribution**:

$$P(Intelligent, Works\_hard, Answers, Grade) = P(Intelligent) * P(Works\_hard) * P(Answers|Intelligent \wedge Works\_hard) * P(Grade|Answers)$$

  – The **domains** of the **variables** are simple, for example, the domain of ***Answers*** may be {***insightful, clear, superficial, vacuous***}

Belief Networks

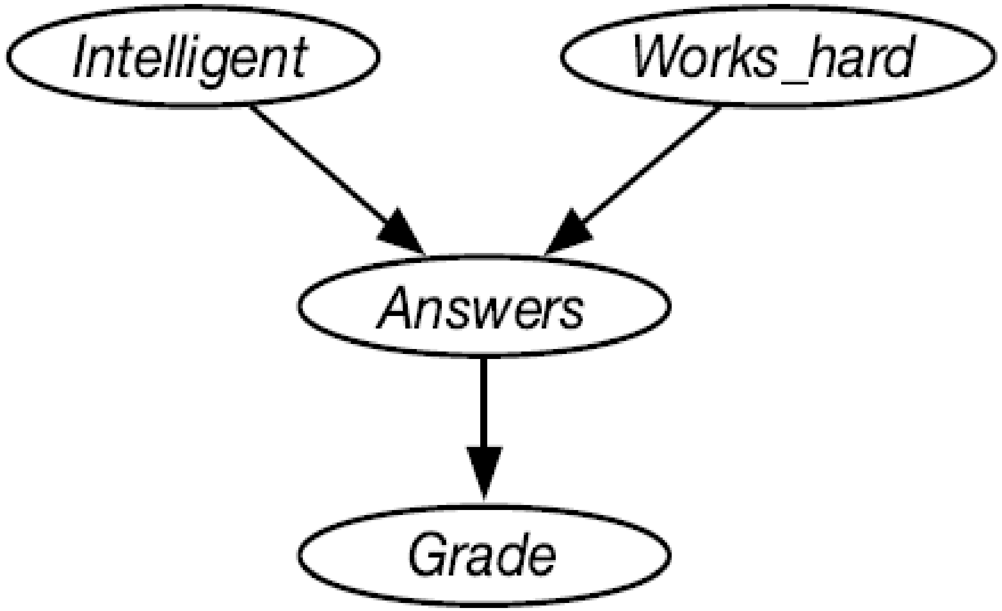


Figure 8.2: Belief network for exam answering of Example 8.13

- A **belief network** specifies a **joint probability distribution** from which arbitrary conditional probabilities can be derived.

Observations and Queries

- **Example 8.14:** Before there are any observations, the distribution over intelligence is ***P*(*Intelligent*)**, which is provided as part of the network.
- To determine the distribution over grades, ***P*(*Grade*)**, requires **inference**.
- If a grade of **A** is observed, the **posterior distribution** of ***Intelligent*** is given by: ***P*(*Intelligent*|*Grade*=A);** *intelligent in the presence of grade A*
- If it was observed that ***Works_hard*** is **false**, the **posterior distribution** of ***Intelligent*** is:

$$P(Intelligent|Grade{=}A \wedge Works\_hard = false).$$

  - Although ***Intelligent*** and ***Works_hard*** are **independent** as per the given observations, they depend on the ***grade***.
- This might explain why some people claim **they did not work hard** to get a **good grade**; it increases the probability that they are **intelligent**.

Constructing Belief Networks

- To represent a domain in a **belief network**, the designer of a network must consider the following questions:
  - **What are the relevant variables?** What the agent may observe in the domain. What information is the agent interested in knowing the posterior probability of? What are the other hidden variables?
  - **What values should these variables take?** For each variable, the designer should specify what it means to take each value in its domain.
  - **What is the relationship between the variables?** This should be expressed by adding arcs in the graph to define the parent relation.
  - **How does the distribution of a variable depend on its parents?** This is expressed in terms of ***conditional probability distributions***.

Example 8.15 Suppose you want to use the diagnostic assistant to diagnose whether there is a fire in a building and whether there has been some tampering with equipment based on noisy sensor information and possibly conflicting explanations of what could be going on. The agent receives a report from Sam about whether everyone is leaving the building. Suppose Sam's report is noisy: Sam sometimes reports leaving when there is no exodus (a false positive), and sometimes does not report when everyone is leaving (a false negative). Suppose the leaving only depends on the fire alarm going off. Either tampering or fire could affect the alarm. Whether there is smoke only depends on whether there is fire. Suppose we use the following variables in the following order:

- *Tampering* is true when there is tampering with the alarm.
- *Fire* is true when there is a fire.
- *Alarm* is true when the alarm sounds.
- *Smoke* is true when there is smoke.
- *Leaving* is true if there are many people leaving the building at once.
- *Report* is true if Sam reports people leaving. *Report* is false if there is no report of leaving.

Assume the following conditional independencies:

- *Fire* is conditionally independent of *Tampering* (given no other information).
- *Alarm* depends on both *Fire* and *Tampering*. That is, we are making no independence assumptions about how *Alarm* depends on its predecessors given this variable ordering.
- *Smoke* depends only on *Fire* and is conditionally independent of *Tampering* and *Alarm* given whether there is a *Fire*.
- *Leaving* only depends on *Alarm* and not directly on *Fire* or *Tampering* or *Smoke*. That is, *Leaving* is conditionally independent of the other variables given *Alarm*.
- *Report* only directly depends on *Leaving*.

The belief network of Figure 8.3 (on the next page) expresses these dependencies. This network represents the factorization

$$\boxed{\begin{aligned} & P(Tampering, Fire, Alarm, Smoke, Leaving, Report) \\ & \quad = P(Tampering) * P(Fire) * P(Alarm \mid Tampering, Fire) \\ & \qquad * P(Smoke \mid Fire) * P(Leaving \mid Alarm) * P(Report \mid Leaving). \end{aligned}}$$

Note that the alarm is not a smoke alarm, which would affected by the smoke, and not directly by the fire, but rather is a heat alarm that is directly affected by the fire. This is made explicit in the model in that the $\boxed{Alarm}$ is independent of $\boxed{Smoke}$ given $\boxed{Fire.}$

We also must define the domain of each variable. Assume that the variables are Boolean; that is, they have $\boxed{\text{domain } \{true, false\}}$. We use the lowercase variant of the variable to represent the true value and use negation for the false value. Thus, for example, $\boxed{Tampering = true}$ is written as $\boxed{tampering}$, and $\boxed{Tampering = false}$ is written as $\boxed{\neg tampering.}$

Constructing Belief Networks

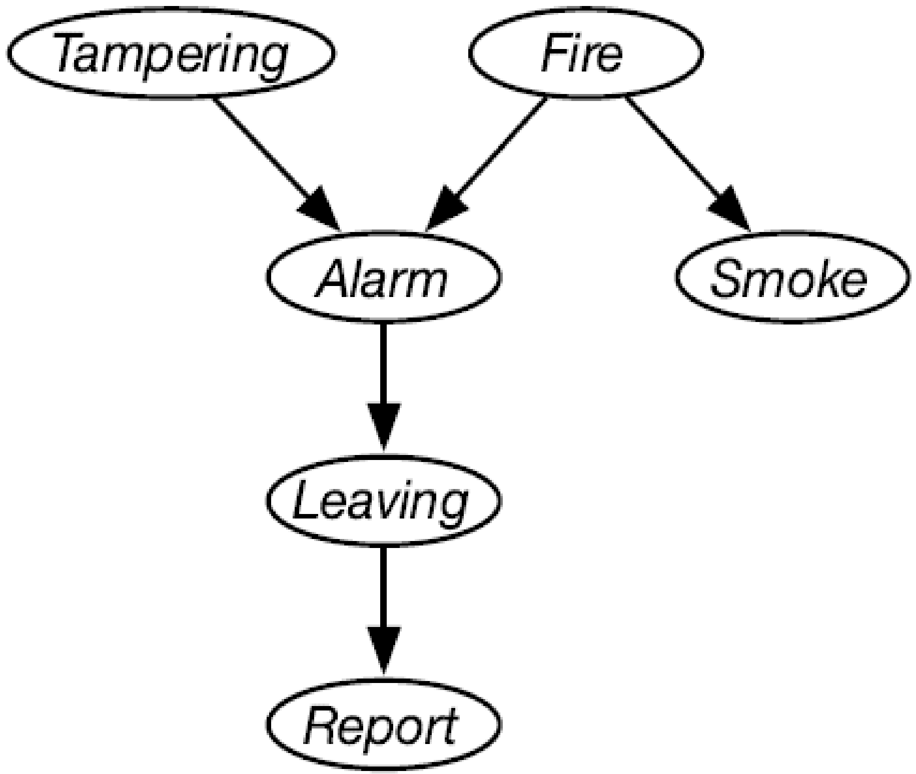


Figure 8.3: Belief network for report of leaving of Example 8.15

The examples that follow assume the following conditional probabilities:

$$P(tampering) = 0.02$$
$$P(fire) = 0.01$$
$$P(alarm \mid fire \wedge tampering) = 0.5$$
$$P(alarm \mid fire \wedge \neg tampering) = 0.99$$
$$P(alarm \mid \neg fire \wedge tampering) = 0.85$$
$$P(alarm \mid \neg fire \wedge \neg tampering) = 0.0001$$

$$P(smoke \mid fire) = 0.9$$
$$P(smoke \mid \neg fire) = 0.01$$
$$P(leaving \mid alarm) = 0.88$$
$$P(leaving \mid \neg alarm) = 0.001$$
$$P(report \mid leaving) = 0.75$$
$$P(report \mid \neg leaving) = 0.01$$

Before any evidence arrives, the probability is given by the priors. The following probabilities follow from the model (all of the numbers here are to about three decimal places):

$$P(tampering) = 0.02 \qquad P(report) = 0.028$$
$$P(fire) = 0.01 \qquad P(smoke) = 0.0189$$

Observing a report gives the following:

$$P(\textit{tampering} \mid \textit{report}) = 0.399$$
$$P(\textit{fire} \mid \textit{report}) = 0.2305$$
$$P(\textit{smoke} \mid \textit{report}) = 0.215$$

As expected, the probabilities of both *tampering* and *fire* are increased by the report. Because the probability of *fire* is increased, so is the probability of *smoke*. Suppose instead that *smoke* alone was observed:

$$P(\textit{tampering} \mid \textit{smoke}) = 0.02$$
$$P(\textit{fire} \mid \textit{smoke}) = 0.476$$
$$P(\textit{report} \mid \textit{smoke}) = 0.320$$

Note that the probability of *tampering* is not affected by observing *smoke*; however, the probabilities of *report* and *fire* are increased.

Suppose that both *report* and *smoke* were observed:

$$P(tampering \mid report \wedge smoke) = 0.0284$$
$$P(fire \mid report \wedge smoke) = 0.964$$

Observing both makes *fire* even more likely. However, in the context of the *report*, the presence of *smoke* makes *tampering* less likely. This is because the *report* is **explained away** by *fire*, which is now more likely.

Suppose instead that *report*, but not *smoke*, was observed:

$$P(tampering \mid report \wedge \neg smoke) = 0.501$$
$$P(fire \mid report \wedge \neg smoke) = 0.0294$$

In the context of the *report*, *fire* becomes much less likely and so the probability of *tampering* increases to explain the *report*.

Constructing Belief Networks

- **Consider the following situation**

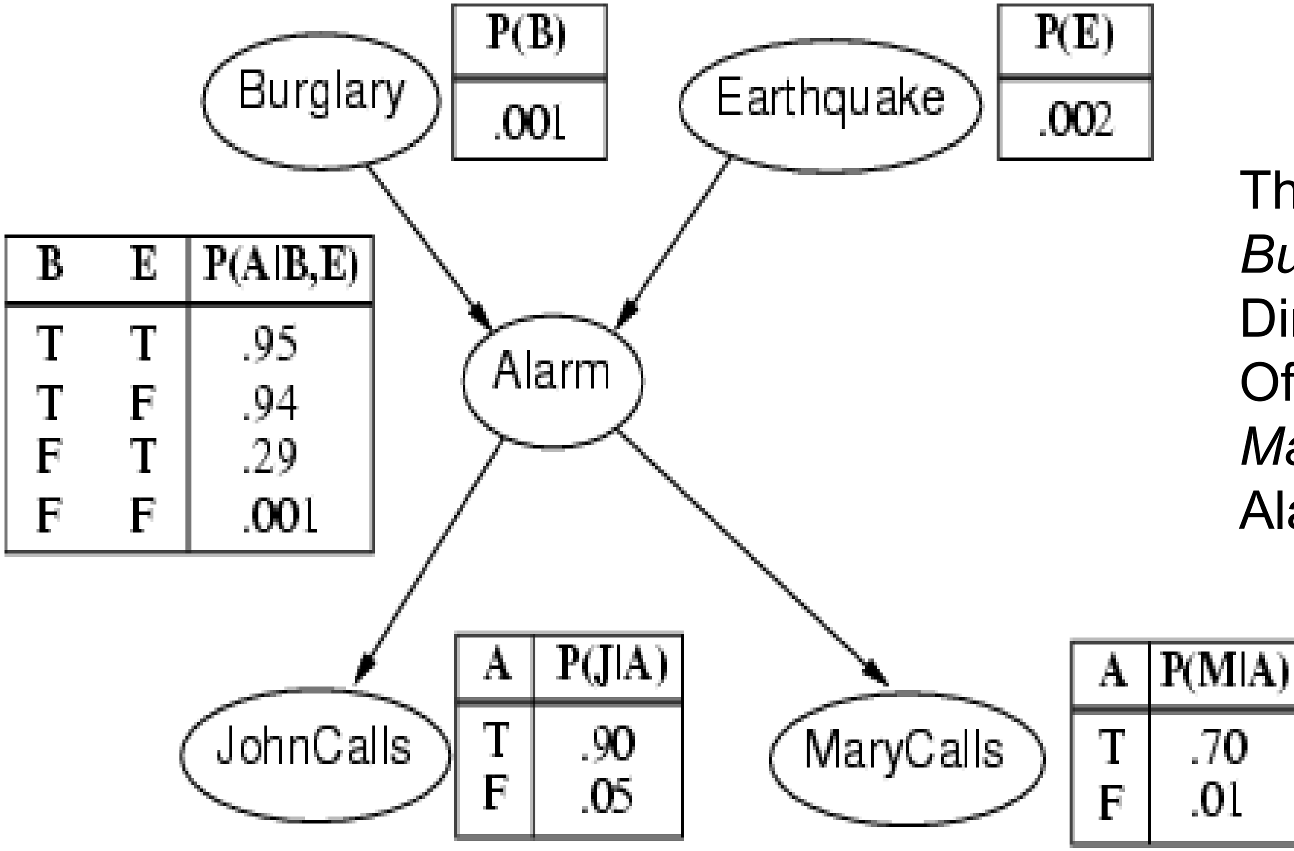
  You have a new burglar alarm installed at home. It is reliable at detecting a **burglary** but also responds to minor earthquakes. You have two neighbors, *John* and *Marry*, who have promised to call you at work when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then. On the other hand, Marry likes rather loud music and sometimes misses the alarm. Given the evidence of who has or has not called, we would like to estimate the **probability of a burglary**.
- In the case of a **burglary network**, **burglary** and **earthquakes** directly affect the probability of the **alarm**, but John and **Marry's call** depends only on the alarm; that is, the network represents that they do not perceive any burglaries directly, and they do not feel the minor earthquakes.

<u>Notice that</u> the network does not have nodes corresponding to Marry currently listening to loud music or the telephone ringing and confusing John. The probabilities summarize a potentially infinite set of circumstances in which the alarm might fail, John or Marry might fail to call, etc.

Constructing Belief Networks

- "*I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call* ". Sometimes it's set off by minor earthquakes. Is there a burglar?
  - **Variables**: *Burglary*(***B***), *Earthquake*(***E***), *Alarm*(***A***)*, JohnCalls*(***J***), *MaryCalls*(***M***)
- Network topology reflects "causal" knowledge:
  - A *burglar* can set the alarm off
  - An *earthquake* can set the alarm off
  - The alarm can cause *Mary to call*
  - The alarm can cause *John to call*

Constructing Belief Networks



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

The topology shows that *Burglary* and *Earthquake* Directly affect the probability Of the alarm. But *JohnCall* and *MaryCall* depend on the Alarm.

- In the **CPT**(**conditional probability table**), letters *B*, *E*, *A*, *J* and *M* stand for Burglary, Earthquake, Alarm, JohnCalls and MaryCalls respectively

Constructing Belief Networks

- Once we have specified the topology, we need to specify the CPT for each node. For example, **the CPT for the random variable *Alarm* might look like this:**

B	*E*	*P*(*A*|*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

Constructing Belief Networks

- There are two ways to understand the semantics of Bayesian networks,
  - To see the network as a representation of **JPD (joint probability distribution)**
  - To view the network as an encoding of a **collection of conditional independence** statements
- A generic entry in the joint is the probability of ***conjunction*** of particular assignments to each variable, such as

  $P(X_i = x_1 \wedge x_2 \wedge \ldots x_n) = \prod_{i=1-n} P(X_i | Parents(X_i))$

- Calculate the probability of the event that the *alarm* has sounded, but neither a *burglary* nor an *earthquake* has occurred, and both John and Mary call done:
  - $\boldsymbol{P(J \wedge M \wedge A \wedge \sim B \wedge \sim E) = P(J|A) * P(M|A) * P(A \mid \sim B \wedge \sim E) * P(\sim B)\, P(\sim E)}$

    $= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = \mathbf{0.00062}$

Constructing Belief Networks

- **Example 8.16** Consider the problem of diagnosing why someone is ***sneezing*** and perhaps has a ***fever***. Sneezing could be because of **influenza** or because of **hay fever**. They are not independent, but are correlated due to the **season**. Suppose hay fever depends on the season because it depends on the amount of **pollen**, which in turn depends on the **season**. The agent does not get to observe **sneezing** directly, but rather observed just the "***Achoo***" ***sound***. Suppose fever depends directly on ***influenza***. These dependency considerations lead to the **belief network** of **Figure 8.4**.

Constructing Belief Networks

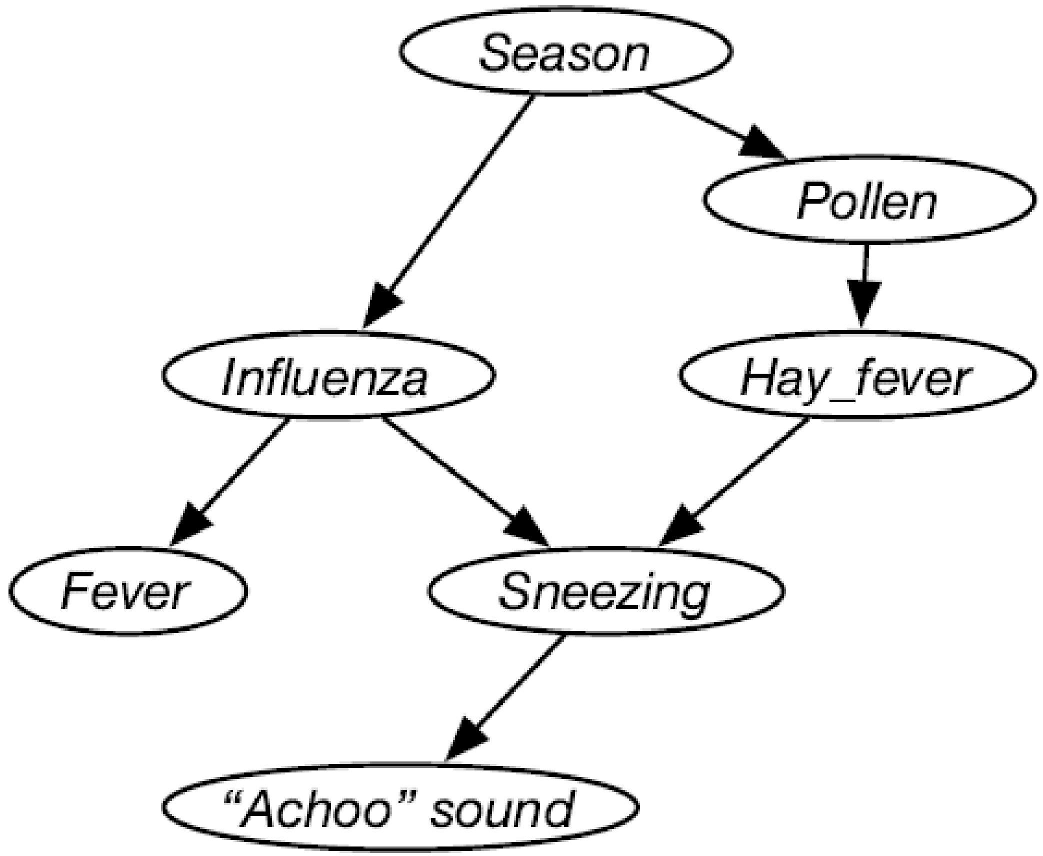
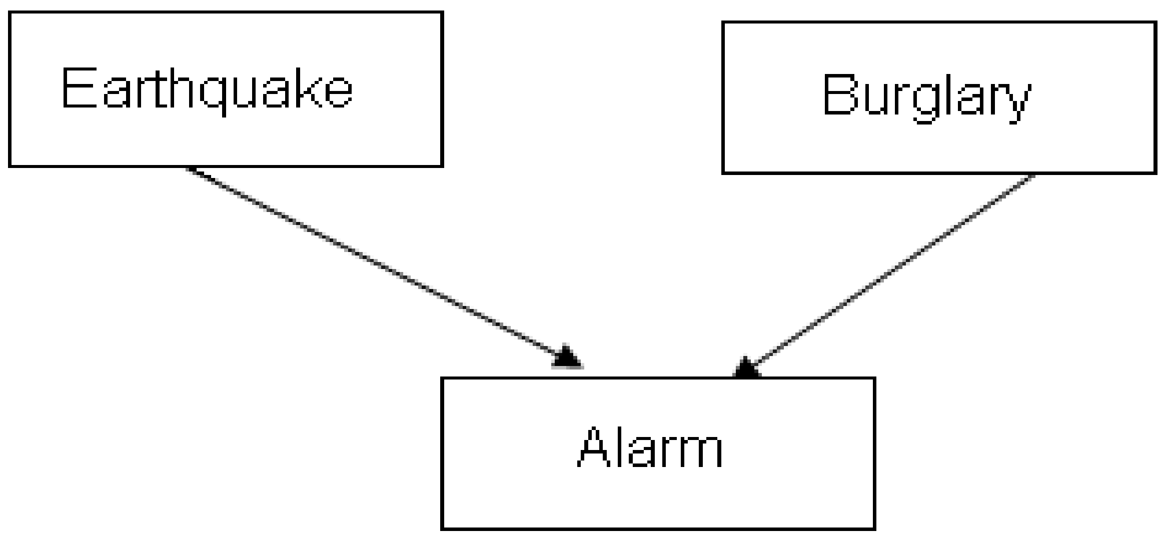


Figure 8.4: Belief network for Example 8.16

Exercise 1

- Consider the following **three variables** Bayesian network:



Exercise 1 cont

- The **joint distribution** over the three variables Earthquake (*E*), Burglary (*B*) and Alarm (*A*) are shown as:

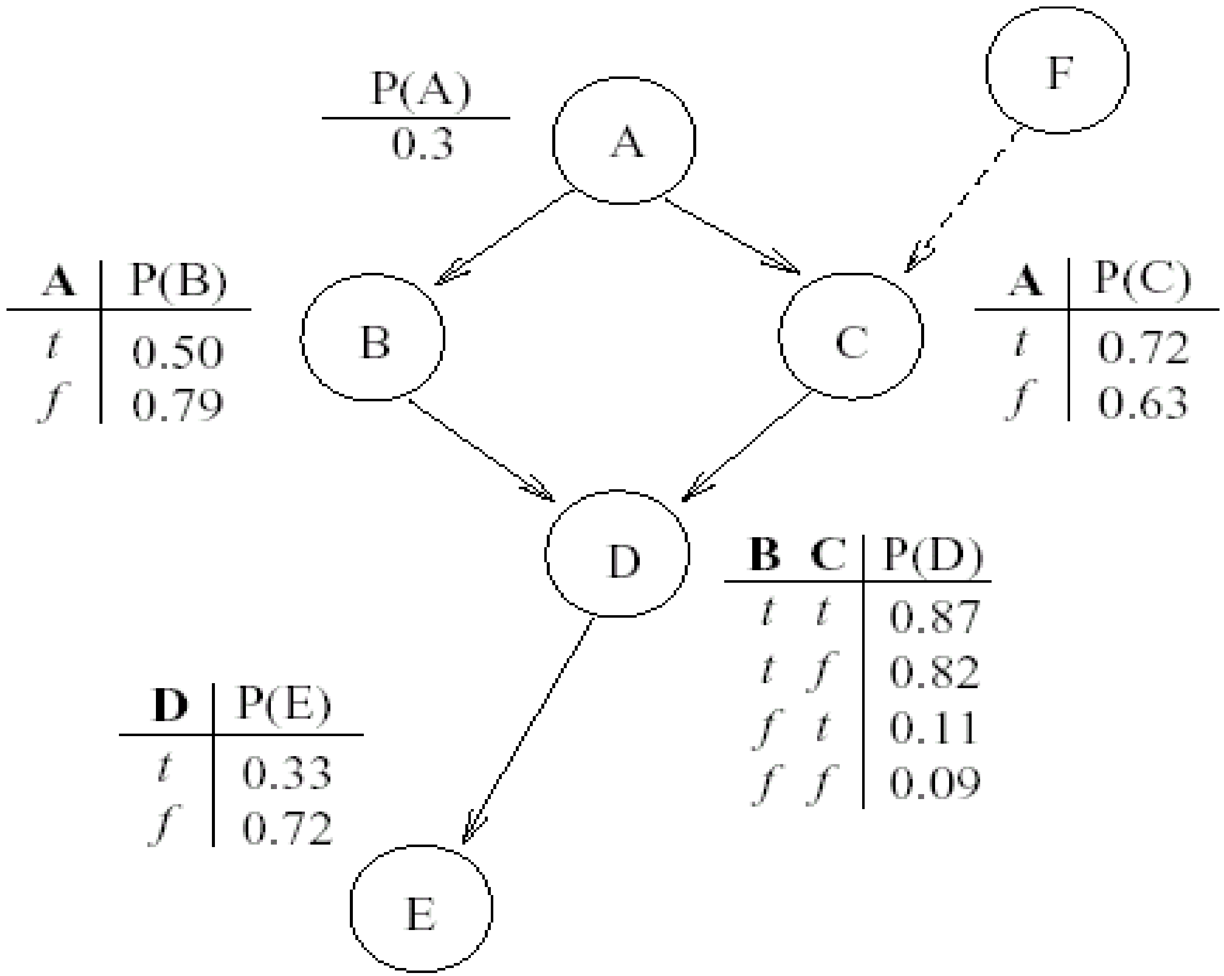
E^1	B^1	A^1	0.0270
E^1	B^1	A^0	0.0030
E^1	B^0	A^1	0.1620
E^1	B^0	A^0	0.1080
E^0	B^1	A^1	0.0140
E^0	B^1	A^0	0.0560
E^0	B^0	A^1	0.0063
E^0	B^0	A^0	0.6237

Exercise 1 cont

- Based on the network, compute $P(A^1|E^1)$
- Based on the network, compute $P(A^1| E^0, B^1)$
- Based on the network, compute $P(A^0| E^1, B^1)$
- By direct computation of the joint probability distribution, show $P(E^1|A^1,B^1) < P(E^1|A^1)$.

Exercise 2

Given the following network , answer the following questions about **conditional independence** (node F is a newly added node for question e)



P(A)
0.3

A	P(B)
t	0.50
f	0.79

A	P(C)
t	0.72
f	0.63

B	C	P(D)
t	*t*	0.87
t	*f*	0.82
f	*t*	0.11
f	*f*	0.09

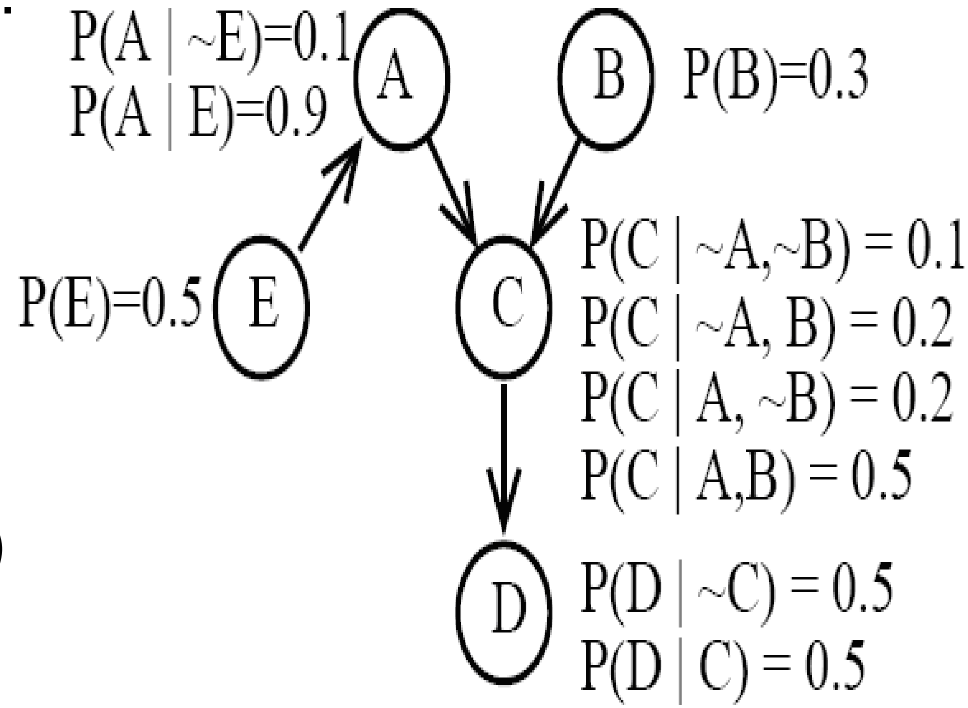
D	P(E)
t	0.33
f	0.72

Exercise 2 cont

a. Are A and E conditionally independent given D?
b. Are A and E conditionally independent given B?
c. Are B and C conditionally independent given A?
d. Are B and C conditionally independent given A and E?
e. If we add new node F to the network, are B and F conditionally independent given C?

Exercise 3

- Consider the following Bayesian Network shown below containing four Boolean random variables.



- Calculate the following:
  1) $P(A \wedge C \wedge D \wedge \sim B \wedge \sim E)$
  2) $P(\sim D | C \wedge E)$
  3) $P(B|D)$
  4) Show an **expression** for how to compute $P(D)$ in terms of values in the **full joint distribution**