7. Uncertain knowledge

By: Alex S.

Apr, 2024

1. Acting under uncertainty

Using propositional logic does not always work as intended. Consider the example:

$$\forall p \quad \text{symptoms}(p, \text{toothache}) \rightarrow \text{disease}(p, \text{cavity})$$

However, it is not always the case that toothache is received from a cavity, let's describe other problems as well:

$$\forall p \quad \text{symptoms}(p, \text{toothache}) \rightarrow \text{disease}(p, \text{cavity})$$
$$\vee \, \text{disease}(p, \text{gum problems}) \vee \ldots$$

There are several problems:

1. There can be lots of different problems, and we might miss some condition.
2. We cannot use disjunction for inference. We need conjugate form.
3. We are lazy to describe everything.

Therefore we use **decision theory** to describe rational decisions for the agent:

Decision theory = probability theory + utility theory

With this theory, the agent is rational when it chooses the action that gives the best result which is the arithmetic average of all available actions.

2. Probability theory

- Described by **random variables**, e.g. X
- Discrete random variable: $P(\text{Weather}) = (0.7; 0.3)$ or $P(\text{Weather} = \text{sunny}) = 0.7$ and

$P(\text{Weather} = \text{rainy}) = 0.3$. These types always have a **domain** - the set of possible values, in this case Weather $\in \{\text{sunny}, \text{rainy}\}$ or Dice $\in \{1...6\}$

- Boolean random variable: $P(\text{cavity})$ or $P(\neg \text{ cavity})$. It has a domain of $\{\text{true}, \text{false}\}$
- Continuous random variables cannot be defined with the vector of values. Instead it uses a **probability density function**:

$$P(\text{Temp} = x) = \text{Uniform}_{[18,26]}(x)$$

$$= \begin{cases} \frac{1}{8} & \text{if } 18 \leq x \leq 26 \\ 0 & \text{otherwise} \end{cases}$$

It means that in 100% the temperature is in the interval of 18 - 26 Celsius.

2.1. Syntax:
- $P(A)$ is a **prior** or **unconditional probability**
- $P(A|B)$ is a **conditional** or **posterior** probability where A is given by B(evidence). In the agent world B might be a measurement from the sensor.
- **Conditional probability** product rule:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}, \text{if } P(B) > 0$$

$$P(A \wedge B) = P(A|B) \cdot P(B)$$
$$P(A \wedge B) = P(B|A) \cdot P(A)$$

2.2. Axioms
- **Probability model**:

$0 \leq P(\omega) \leq 1$ for every ω and $\sum_{\omega \in \Omega} P(\omega) = 1$, where ω is some world, but Ω is a set of all possible worlds called the **sample space**.

- $P(\text{true}) = 1, P(\text{false}) = 0$
- **Kolmogorov's axiom**: $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$, we can derive $P(A \wedge \neg A) = P(A) + P(\neg A) = 1$

2.3. A full joint distribution example

	Toothache	$\neg$ Toothache	$\sum$
Cavity	0.04	0.06	0.1
$\neg$ Cavity	0.01	0.89	0.9
$\sum$	0.05	0.95	

- $P(\text{Cavity}) = P(\text{Toothache}) + P(\neg \text{ Toothache}) = 0.1$

It is also called a **marginal probability**. General form:

$$P(Y) = \sum_{z \in Z} P(Y, z)$$

In the example;

$$P(\text{Cavity}) = \sum_{z \in \{\text{Toothache}\}} P(\text{Cavity}, z)$$

- $$P(\text{Cavity} \vee \text{Toothache}) =$$
  $$= P(\text{Cavity}) + P(\text{Toothache}) -$$
  $$- P(\text{Cavity} \wedge \text{Toothache}) =$$
  $$= 0.1 + 0.05 - 0.04 = 0.11$$

- $$P(\text{Cavity} \mid \text{Toothache}) =$$
  $$= \frac{P(\text{Cavity} \wedge \text{Toothache})}{P(\text{Toothache})} =$$
  $$= \frac{0.04}{0.05} = \frac{4}{5} = 0.8$$

3. Bayes' rule and its use

Previously we defined a **product rule**, from which Bayes' rule is derived:

$$P(A \wedge B) = P(A|B) \cdot P(B)$$

$$P(A \wedge B) = P(B|A) \cdot P(A)$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

- It is used for diagnostic laws
- It is also used for probabilistic inference mechanisms

Bayes' rule is used for **uncertain knowledge representation** - *Bayes' network*:

3.1. Bayes' network

- It is a directed acyclic graph(DAG)
- The node is a random variable
- The node has a distribution table
- Edge X to Y represents that X influences Y

An example:

- An alarm system in the house
- It can be activated if earthquake happens
- It is activated if burglary happens
- John or Mary can call if they supposedly heard the alarm

The graph can be represented as in the image Figure 1. The graph represents a knowledge base.
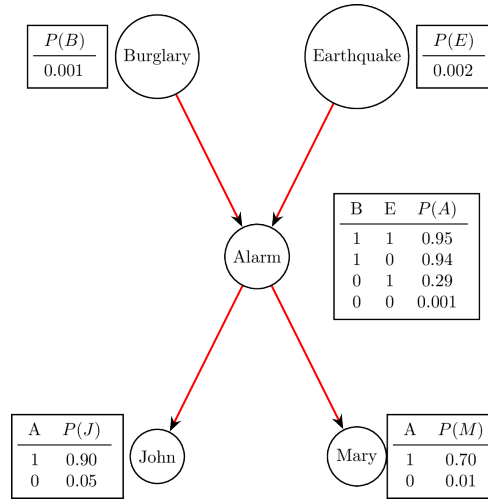

Figure 1: Bayes' network example

Now we can use it for querying probabilities:

$$P(X_i \mid \text{parents}(X_i))$$

The value of the entry can be calculated as this:

$$P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n) = P(x_1, \ldots, x_n) =$$

$$\Pi_{i=1}^{n} P(x_i \mid \text{parents}(x_i))$$

Examples:

$$P(A, J, M, \neg B, \neg E) = P(A \wedge J \wedge M \wedge \neg B \wedge \neg E) =$$

$$= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg E) \cdot P(\neg B) =$$

$$= 0.9 \cdot 0.7 \cdot 0.001 \cdot 0.999 \cdot 0.998 = 0.000628$$

$$\bullet \qquad P(A \wedge B \wedge E) =$$

$$= P(A \mid B \wedge E) \cdot P(B) \cdot P(E) =$$

$$= 0.95 \cdot 0.001 \cdot 0.002 = 0.0000019$$

$$\bullet \qquad P(J, \neg M, \neg A, \neg B, \neg E) =$$

$$= P(J \mid \neg A) \cdot P(\neg M \mid \neg A) \cdot P(\neg A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) =$$

$$= 0.05 \cdot 0.99 \cdot 0.999 \cdot 0.999 \cdot 0.998 = 0.0493$$

4. Chain rule

The Bayesian network is a correct representation of the domain only if each node is conditionally independent of its other predecessors in the node ordering, given its parents:

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1}, \ldots, x_1)$$

$$P(x_1, \ldots, x_n) = P(x_n | x_{n-1}, \ldots, x_1) \cdot$$

$$\cdot P(x_{n-1} | x_{n-2}, \ldots, x_1) \cdot \ldots \cdot P(x_2 | x_1) \cdot P(x_1) =$$

$$= \Pi_{i=1}^{n} P(x_i | x_{i-1}, \ldots, x_1)$$

Every node in the network is equal to:

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

$$\text{Parents}(X_i) \subseteq \{X_{i-1}, \ldots, X_1\}$$

5. Conditional independence relations in Bayesian networks

We have provided a "numerical" semantics for Bayesian networks in terms of the representation of the full joint distribution, as in equation above.

Now let's define topological semantics that specifies the conditional independence relationships encoded by the graph structure. From which later we can derive numerical semantics.

1. A node X is conditionally independent of its non-descendants(e.g. Z_{ij}s), given its parents (the U_is shown in the dashed area).
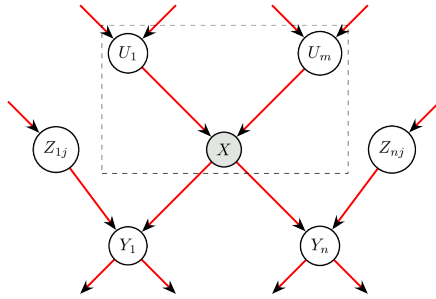
Figure 2: X - conditionally independent node, U_i - parents, $Z_{ij}s$ - non-descents, Y_i - descents

2. A node X is conditionally independent of all other nodes in the network given its **Markov blanket** (the dashed area).
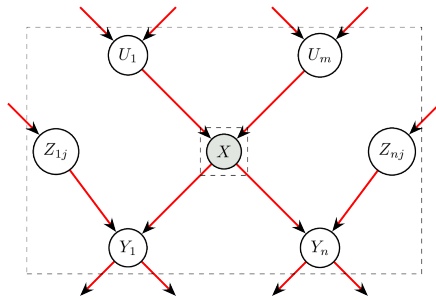

Figure 3: X - conditionally independent node, U_i - parents, $Z_{ij}s$ - non-descents, Y_i - descents

5.1. D-separation

There is also a general topological criterion called **d-separation** for deciding whether a set of nodes X is conditionally independent of another set Y, given a third set Z.

A set of nodes E(evidence) d-separate X and Y, if every non-directed path $X \to Y$ is blocked by the E set(There might be other nodes in the path as well). The path is blocked by the set E, if there is a node z, for which one out of 3 conditions is met:
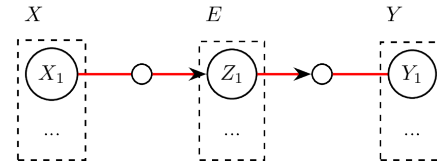
1. $z \in E \land (z$ has one input and output$)$

Figure 4: Z_1 blocks the path to the Y_1 for X_1
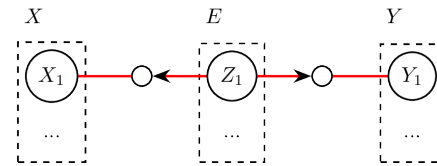
2. $z \in E \land (z$ has two outputs$)$

Figure 5: Z_1 blocks the path to the Y_1 for X_1

3. $z \land \text{children}(z) \notin E \land (z$ has two inputs$)$
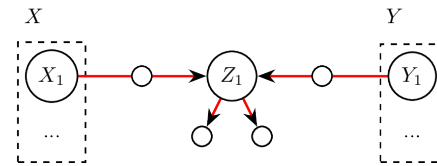

Figure 6: $Z_1 \notin E$ blocks the path to the Y_1 for X_1
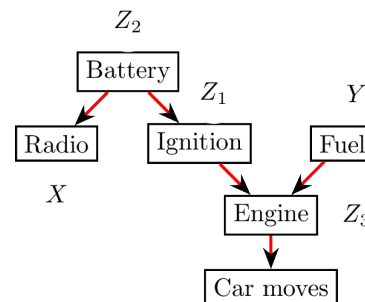
5.2. D-separation example

Figure 7: An example of d-separation in the Bayesian network

1. Evidence Z_1 blocks $X \to Y$ in a way described in the first D-separation rule.
2. Evidence Z_2 blocks $X \to Y$ in a way described in the second D-separation rule.
3. If there is no evidence at all, then X is independent from Y as in the third D-separation rule.
4. With evidence Z_3 it turns out that X is dependent on Y, because we have an evidence and there is no rule for a separation.

6. Inference types

- Every node in the network can be used for querying and getting a proof.

In all images Q is query, E is evidence

1. **Diagnostic inference** - from conclusions to consequences.

Figure 8: The query is $P(B|J)$, B - burglary, J - John

2. **Causal inference** - from consequences to conclusions.

Figure 9: The query is $P(J|B)$, B - burglary, J - John calls

3. **Midcausal inference** - from common conclusions get consequences.

Figure 10: The query is $P(B|A \wedge K)$, B - burglary, A - Alarm, K - earthquake

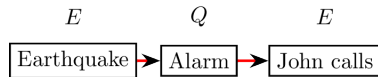4. **Mixed inference** - two or 3 previous inference combination.

Figure 11: The query is $P(A|J \wedge \neg K)$, J - John calls, A - Alarm, K - earthquake. The combination of $1 + 2$ inference
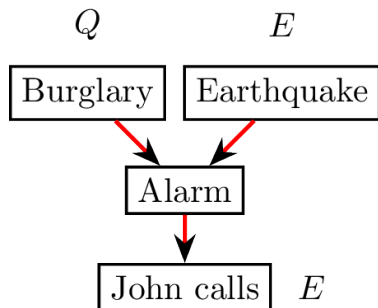

Figure 12: The query is $P(B|K \wedge J)$, J - John calls, B - Burglary, K - earthquake. The combination of $1 + 3$ inference

7. Probability system modelling steps

1. Knowledge engineer should decide what factors should be modelled and which could not directly affect the model
2. It should be decided about the random variable dictionary - is it discrete, continuous or boolean.
3. Need to decide the Bayesian network's topology. Find the qualitative or quantitative relations between random variables.

4. Encode the specific cases.
5. Test the network by querying. In some cases a **sensitivity analysis** can occur which helps to check how robust the values are.