# Bayesian Networks

In the Probability Theory lesson we have seen how the full joint distribution can be decomposed into smaller tables by using **CONDITIONAL INDEPENDENCE.**

Now we are going to see how to decompose a full joint distribution systematically by using BAYESIAN NETWORKS.

Bayesian networks can represent essentially any full joint probability distribution and in many cases can do so very concisely.

Text

Description automatically generated

Thus, a Bayesian Network is a DAG where each node represents a random variable X and the information contained in that node is the conditional probability distribution **P**(X | Parents(X)).

Diagram

Description automatically generatedIn other words, since every node X must contain the conditional probability distribution given its parents then this means that the parents must cause X. Remember that while causal knowledge is robust, diagnostic knowledge is not. Thus, when we create a Bayesian network, we need to reason in a causal way which means that the parent of each node X should be a cause of X and X should be the effect.

Recall the simple world described in Chapter 13, consisting of the variables Toothache, Cavity, Catch, and Weather . We argued that Weather is independent of the other variables; furthermore, we argued that Toothache and Catch are conditionally independent, given Cavity. These relationships are represented by the Bayesian network structure shown in Figure 14.1. Formally, the conditional independence of Toothache and Catch, given Cavity, is indicated by the absence of a link between Toothache and Catch. Intuitively, the network represents the fact that Cavity is a direct cause of Toothache and Catch, whereas no direct causal relationship exists between Toothache and Catch.

Now consider the following example, which is just a little more complex. You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes. (This example is due to Judea Pearl, a resident of Los Angeles—hence the acute interest in earthquakes.) You also have two neighbours, John and Mary, who have promised to call you at work when they hear the alarm. John nearly always calls when he hears the alarm, but sometimes confuses the telephone ringing with

Text

Description automatically generated

Diagram

Description automatically generatedAs we can see in FIG. 14.2, the Bayesian Network has been built from causes to consequences using causal knowledge instead diagnostic knowledge.

Given that a node X has k Boolean parents then it contains a conditional table of 2k rows. Given that X represents a Boolean random variable then the table will contain exactly 2k entries because each row by the axiom of probability must sum up to 1 and so if we know P(a|b,e) then it follows that

P(-a|b,e)= 1 – P(a|b,e) by the axiom of probability

## Semantic of Bayesian Networks

There are two equivalent ways to understand the semantics of Bayesian Networks:

1. A Bayesian Network represents the full joint probability distribution of all present random variables
2. A Bayesian Network is a collection of conditional independence statements

The two views are equivalent, but the first turns out to be helpful in understanding how to construct networks, whereas the second is helpful in designing inference procedures.

Proof Bayesian networks are equivalent to the full joint distribution

Text, letter

Description automatically generatedGiven that Bayesian networks are equivalent to Full Joint Distribution then they are sufficient to answer any query about a set of random variables.

A picture containing text

Description automatically generatedWe can express the equivalence between Bayesian networks and Full Joint Distributions in the following way:

A picture containing logo

Description automatically generated

Text, letter

Description automatically generated

Bayesian Networks do not contain redundant probability values.

It is impossible for the knowledge engineer or domain expert to create a Bayesian network that violates the axioms of probability.

In other words, a Bayesian Network is correct only when it is **ACYCLIC.­­­­­**

## Why using Bayesian Networks?

A Bayesian network can often be far more compact than the full joint distribution.

A close up of a newspaper

Description automatically generated

Therefore, Bayesian Networks let us deal with probabilistic problems which otherwise would be intractable if using the Full Joint Distribution.

In other words, the more links between nodes we add in a Bayesian Network, the more information we need to store. Therefore, we must find a trade-off between the number of links and the accuracy we want to achieve through the model.

Naturally, the more links there are, the more the Bayesian Network reflects the real world.

What is important when creating Bayesian Networks is that we always need to use causal relations and never diagnostic ones.

If we stick to a causal model, we end up having to specify fewer numbers, and the numbers will often be easier to come up with.

Diagram

Description automatically generated

Figure 14.3(b) shows a very bad node ordering: MaryCalls, JohnCalls, Earthquake, Burglary, Alarm. This network requires 31 distinct probabilities to be specified—exactly the same number as the full joint distribution. It is important to realize, however, that any of the three networks can represent exactly the same joint distribution.

Therefore, the choice of the node ordering is important as well. If we choose a bad node ordering then we might create a Bayesian Network that requires as much space in memory as a FULL JOINT DISTRIBUTION.

In order to prove that Bayesian Networks are equivalent to the Full Joint Distribution given that Parents(Xi) ⊆ {Xi-1,Xi-2,…,X1} I have used the Chain rule:

Text, Word

Description automatically generated

In my proof of equivalence between Bayesian Networks and Full Joint Distribution, I have used unconsciously conditional independence.

Given that we have 3 random variables A, B and C and given that B and A are conditionally independent given C then the following follows:

1. P(A **∧** B | C) = P(A|C) P(B|C)
2. P(A|C **∧** B) = P(A|C)

The proof is the following:

A picture containing text

Description automatically generatedTherefore, conditional independence and the Chain rule are the backbone of Bayesian Networks

## Bayesian Belief Update

A picture containing text

Description automatically generatedIn other words, Bayesian systems are important in AI because through them, an AI System can infer the probability that it is in a certain environment state.

Usually, such inference of the model (environment state) occurs through some observed data.

Thus, what probability systems based on Bayesian networks do is computing the probability that the environment is in a certain state given the observed data. In other words, the most common queries that such systems answer are **CONDITIONAL PROBABILTIES.**

Now we are going to analyse how Bayesian systems can infer probabilities.

The algorithm that we will analyse is the basic one called **Inference by enumeration** which exploits the probability axioms we have already studied.

It is important to notice that there are far better algorithms to inference probability rather than inference by enumeration.

## Inference by enumeration

Inference by enumeration exploits the properties that we have already shown.

The property we are talking about is the following:

A close up of a watch

Description automatically generated

Even though we do not have a full joint distribution we can compute the right hand side because as we have shown a Bayesian network is equivalent to the full joint distribution.

We can see what has already been said in the following example.

Text, letter

Description automatically generatedThe above formula represents how we can compute a conditional probability given a Bayesian network and it is called INFERENCE BY ENUMERATION.

Now we are going to prove that the time complexity of the INFERENCE BY ENUMERATION algorithm is O(n2n) assuming all the random variables are Boolean.

Given that we have a Bayesian network with n Boolean random variables then consequently we have n conditional probability tables.

Given that we want to know P(X) where X is a random variable of the Bayesian network then we need to follow the following procedure:

P(X) = a where a is the normalisation factor and z is any possible combination of the values of the remaining random variables. Since the remaining random variables are n-1 and Boolean then the number of possible combinations is 2n-1.

This means that the sum will contain 2n-1 terms. Each term is given by the product of all the conditional probability tables and these are n. Thus the time complexity is O(n2n-1). Additionally, we need to compute both values of X (X=true and X=false) thus we obtain 2O(n2n-1) = O(n2n).

As we can see exact inference in Bayesian networks using INFERENCE BY ENUMERATION is in the worst case exponential in the number of random variables present in the network.