CS3243: Introduction to Artificial Intelligence
Lecture Notes 10: Bayesian Networks

1. Coins

One of our assumptions we have made is that the agent is able to observe everything. However, that is not always the case. Instead, the agent needs to have a **probabilistic model** about the environment that it is in.

Suppose there is a jar containing two coins,

- C_{50}: $P(Head) = 0.5$
- C_{90}: $P(Head) = 0.9$

The idea is to repeatedly pick one of the coins and toss it. The agent does not know which coin I have picked, but it can see the output of the face of which the tossed coin lands. For example, the agent can see that we get the sequence $\{H, T, T, H, T, H\}$.

Before Toss	After Toss
$P(C_{50}) = 0.5, \quad P(C_{90}) = 0.5$	$P(C_{50} \mid Toss) > P(C_{90} \mid Toss)$

The conditional probabilities in the table above was computed using **Bayes' Theorem**.

$$P(C_{50} \mid Toss) = \frac{P(C_{50} \cap Toss)}{P(Toss)} = \frac{P(C_{50})P(Toss \mid C_{50})}{P(Toss)} = \frac{0.5(0.5^6)}{0.5(0.5^6) + 0.5(0.9^3)(0.1^3)}$$

$$P(C_{90} \mid Toss) = \frac{P(C_{90})P(Toss \mid C_{90})}{P(Toss)}$$

The agent has two models of the world. It gathers some evidence, and based on the evidence, it chooses the appropriate model which explains the evidence better.

- <u>Model 1:</u> C_{50} is chosen.
- <u>Model 2:</u> C_{90} is chosen.

This is known as **Model Classification**.

Now, suppose we have a jar containing a hundred coins,

- C_1: $P(Head) = 0.01$
- C_2: $P(Head) = 0.02$

  $\vdots$
- C_{99}: $P(Head) = 0.99$
- C_{100}: $P(Head) = 1$

Notice that the calculation of $P(Toss) = \sum_{i=1}^{100} \frac{1}{100}(P(Toss)C_i)$ becomes very cumbersome. However, we notice that $P(C_{50} \mid Toss)$ and $P(C_{90} \mid Toss)$ still has identical denominators. To compare these two terms, we would only have to compute the numerators.

2. Basics of Probability

Axioms of Probability:
$0 \leq P(a) \leq 1$
$P(a \cup b) = P(a) + P(b) - P(a \cap b)$
$P(True) = 1, P(False = 0)$

Conditional Probability:
$$P(a \mid b) = \frac{P(a \cap b)}{P(b)}$$
$$P(a \mid b, c) = \frac{P(a, b, c)}{P(b, c)}$$

Independence:
a and b are independent if $P(a \mid b) = P(a)$.

Conditional Independence:
Given b, a is conditionally independent of c, i.e. $P(a \mid b, c) = P(a \mid b)$.

Let us make a model of what students are concerned about. They are mostly worried about grades, and job interview. In our model, we declare the following parameters.

- Grades (G)
- Job Interview (I)
- ERP (E)

We then conduct the following data. To interpret the data, G means "grades are high" and $\bar{G}$ (*not* G) means "grades are low". E means "ERP was charged"

G	E	I	Frequency	Probability
T	T	T	160	$P(G, E, I) = \frac{160}{600}$
T	T	F	60	$P(G, E, \bar{I}) = \frac{60}{600}$
T	F	T	240	$P(G, \bar{E}, I) = \frac{240}{600}$
T	F	F	40	$P(G, \bar{E}, \bar{I}) = \frac{40}{600}$
F	T	T	10	$P(\bar{G}, E, I) = \frac{10}{600}$
F	T	F	60	$P(\bar{G}, E, \bar{I}) = \frac{60}{600}$
F	F	T	10	$P(\bar{G}, \bar{E}, I) = \frac{10}{600}$

F	F	F	20	$P(\bar{G}, \bar{E}, \bar{I}) = \frac{20}{600}$

Once you have calculated all these probabilities, suppose you want to find $P(G)$.

$$P(G) = P(G, E, I) + P(G, E, \bar{I}) + P(G, \bar{E}, I) + P(G, \bar{E}, \bar{I})$$

$$P(G) = P[(G, E, I) \cup (G, E, \bar{I}) \cup (G, \bar{E}, I) \cup (G, \bar{E}, \bar{I})]$$

Notice that $P(G, E, I \cap \ G, E, \bar{I}) = 0$.

Now, suppose you want to find $P(G \mid E)$. We will have to look at all the rows where E is true.

$$P(G \mid E) = \frac{60 + 160}{160 + 60 + 10 + 60}$$

Our approach so far was to draw the entire table and compute the probability values for every row. The problem with this, is that, the table would be huge.

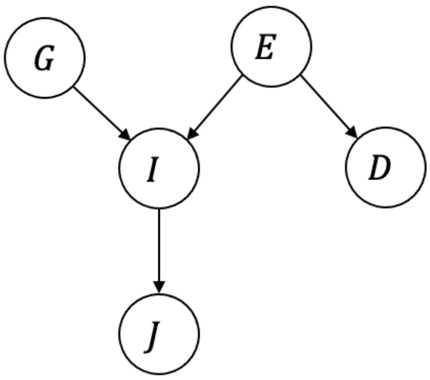
With 5 variables, we require 32 entries (or 31 entries if you exploit the property that all probabilities sum up to 1). If we have n variables, the space needed is $O(2^n)$.

3. Representation of Bayesian Networks

Let's say we have the following variables:

- Grades G
- ERP E
- Interview Performance I
- Job Offer J
- Driver Mood D

We can model all the **causal information** as a graph which is a statement of the world.



From the above graph, we can see that the behavior of a student in the interview is completely determined by the grades and ERP. We would need a smaller probability table to investigate the behavior of I. Such a table is known as a **Conditional Probability Table (CPT)**.

G	E	Probability
T	T	$P(I \mid G, E) = \frac{160}{220}$
T	F	$P(I \mid G, \bar{E}) = \frac{240}{280}$
F	T	$\cdots$
F	F	$\cdots$

The first row gives the probability that I is true given that G and E is true. We can also make use of the complement law to calculate the case where I is false,

$$P(\bar{I} \mid G, E) = 1 - P(I \mid G, E)$$

Such a graph is known as a **Bayesian Network**, also known as Inference Network or Belief Network. It is an acyclic directed graph.

4. Analysis

Suppose that we have a network of n nodes. If we enumerate all combinations of possibilities, the table generated would have 2^n entries.

Let the maximum number of parents for a node be q. The conditional probability table associated with that node will have 2^q entries. In total, the sum of the number of entries of all the conditional probability tables would be $n \times 2^q$. This is a huge saving from 2^n. In the case of 5 variables, we have gone from 32 entries to merely 10 entries.

Each node in a Bayesian network is **independent** of its non-descendants, given its parents. The equations below demonstrate this conditional independence.

$$P(I \mid G, E, D) = P(I \mid G, E)$$

$$P(D \mid G, E) = P(D \mid E)$$

As for variables that are descendants,

$$P(I \mid J, D, E) = \frac{P(I, J, D, E)}{P(J, D, E)}$$

Note that I is not independent of J.

Now, suppose that you want to calculate a particular probability value. We make use of the **Chain Rule**,

$$P(X_1, X_2, \cdots, X_n) = P(X_1 | X_2, \cdots, X_n) \times P(X_2 | X_3, \cdots, X_n) \times \cdots \times P(X_{n-1} | X_n) \times P(X_n)$$

$$P(G, I, J, E, D) = P(G|I, J, E, D) \times P(I|J, E, D) \times P(J|E, D) \times P(E|D) \times P(D)$$

We can make use of the conditional independence to shorten the chain,

$$P(G, I, J, E, D) = P(J, I, G, D, E)$$

$$P(J, I, G, D, E) = P(J|I) \times P(I|G, E) \times P(G) \times P(D|E) \times P(E)$$

The last problem is to find out how do we order the variable such that we exploit the conditional independence fully. We want to order the variable such that in a term $P(X_1|X_2, \cdots, X_n)$, none of the variables in $\{X_2, \cdots, X_n\}$ is a descendent of X_1, and it would be nice to have as many parents as possible.

Since the Bayesian Network is a Directed Acyclic Graph, such an ordering is easily obtained by a **topological sorting**, i.e. recursively choose notes without children and remove them.