| CS3243: Introduction to Artificial Intelligence | Fall 2020 |
|---|---|

# Lecture 10: October 28

*Lecturer: Prof. Kuldeep S. Meel* — *Scribe: Ang Zheng Yong*

## 10.1 Partially Observable Environments

So far, the environments which we have discussed are fully observable for the agent. However, this may not always be the case, and our model may only have a partially observable environment.

We have also discussed about stochastic models, where the actions taken by the agent would lead to outcomes determined by a probability distribution. For instance, consider the following state space:

| | | | |
|---|---|---|---|
| $s_3$ | $s_4$ | $s_5$ | $s_{10}$ |
| $s_2$ | $\times$ | $s_6$ | $s_9$ |
| $s_1$ | $s_{11}$ | $s_7$ | $s_8$ |

Previously, we have noted that given an action $A = direction$, our agent would go in the specified direction with probability 0.7, and go in the other directions with probabilities 0.1 each. For instance, if we tell our agent to head $RIGHT$ from state $s_1$, it will reach state $s_{11}$ with probability 0.7, $s_2$ with probability 0.1, and $s_1$ with probability 0.2.

Today, we will be mainly discussing about partially observable environments where variables can take on binary values.

### 10.1.1 Example 1: Jar with 2 coins

Suppose that there is a jar containing 2 coins, $c_{50}$ and $c_{90}$, where $P(heads) = 0.5$ for $c_{50}$ and $P(heads) = 0.9$ for $c_{90}$ respectively. In this problem, the agent is unable to see which coin is being picked, but it is able to observe our actions (i.e. tossing coins) and the outcome of the coin tosses.

Suppose further that we picked a coin, tossed it 6 times, and obtained the sequence $HTTHTH$, where $H$ represents that the outcome is *heads* and $T$ represents that the outcome is *tails*. The goal of the agent is to know which coin are we holding onto.

Before the coin was tossed, the agent perceives that the probability of us having $c_{50}$ and $c_{90}$ is 0.5 each. After the coin was tossed, the agent would update its perceived probability according to Bayes' Rule:

$$\begin{aligned}
P(c_{50}|toss) &= \frac{P(c_{50} \cap toss)}{P(toss)} \\
&= \frac{P(c_{50})P(toss|c_{50})}{P(toss)} \\
&= \frac{0.5 \times 0.5^6}{0.5 \times 0.5^6 + 0.5 \times 0.9^3 \times 0.1^3} \\
&= 0.95542
\end{aligned}$$

$$\begin{aligned}
P(c_{90}|toss) &= \frac{P(c_{90} \cap toss)}{P(toss)} \\
&= \frac{P(c_{90})P(toss|c_{90})}{P(toss)} \\
&= \frac{0.5 \times 0.9^3 \times 0.1^3}{0.5 \times 0.5^6 + 0.5 \times 0.9^3 \times 0.1^3} \\
&= 0.04458
\end{aligned}$$

### 10.1.2 Model classification

Initially, the agent has 2 models of the world:

- Model 1: $c_{50}$ is chosen.
- Model 2: $c_{90}$ is chosen.

After the coins were tossed, the agent could make use of the evidence (i.e. results of the coin tosses) to choose one of the models. This is known as **model classification**. The goal of the agent is to choose the model that is most capable of explaining the evidence.

In the aforementioned example, since $P(c_{50}|toss) > P(c_{90}|toss)$, the agent would likely conclude that $c_{50}$ was picked instead of $c_{90}$.

### 10.1.3 Example 2: Jar with 100 coins

Now, suppose instead that our jar contains 100 coins, where:

$$\begin{aligned}
c_1 &: P(H) = 0.01 \\
c_2 &: P(H) = 0.02 \\
&\vdots \\
c_{99} &: P(H) = 0.99 \\
c_{100} &: P(H) = 1
\end{aligned}$$

Similarly, we will pick one of the coins, toss the coin 6 times, and get the sequence $HTTHTH$. The goal of the agent is to figure out between $c_{50}$ and $c_{90}$, which of them is able to better explain the evidence (i.e. based outcome of the coin tosses, is $P(c_{50}|toss) > P(c_{90}|toss)$ or $P(c_{50}|toss) < P(c_{90}|toss)$)?

Recall from the previous example that

$$P(c_{50}|toss) = \frac{P(c_{50})P(toss|c_{50})}{P(toss)} \text{ and } P(c_{90}|toss) = \frac{P(c_{90})P(toss|c_{90})}{P(toss)}$$

Since the denominators are equivalent, we only need to compute the numerators of the fractions in order to compare the 2 probabilities. This would save us a significant amount of time, since

$$P(toss) = \sum_{i=1}^{100} \frac{1}{100} P(toss|c_i)P(c_i)$$

and as $i \to \infty$, the calculation will become cumbersome.

## 10.2 Revision for Probability

### 10.2.1 Axioms of probability

For some events $A$ and $B$,

1. $0 \leq P(A) \leq 1$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. $P(\mathit{True}) = 1;\ P(\mathit{False}) = 0$

### 10.2.2 Conditional probability

$P(A|B)$ is described as: "given the occurence of event $B$, what is the probability of event $A$ happening?" Mathematically,

$$P(A|B) = \frac{P(A,B)}{P(B)}$$
$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)}$$
$$\vdots$$

### 10.2.3 Independence of events

Two events $A$ and $B$ are said to be **independent** if $P(A|B) = P(A)$.

Given event $B$, event $A$ is said to be **conditionally independent** of event $C$ if $P(A|B,C) = P(A|B)$.

**Remark:** Prof has also briefly discussed the use of Venn diagrams to help visualize independent events, but it is currently Week 11 and it would take a rather large amount of time to draw them with the TikZ package D: This link provides an overview of Venn diagrams.

### 10.2.4 Chain rule

The chain rule is given by:

$$\begin{aligned}
P(x_1, x_2, \ldots, x_n) &= P(x_1|x_2, x_3, \ldots, x_n)P(x_2, x_3, \ldots, x_n) \\
&= P(x_1|x_2, x_3, \ldots, x_n)P(x_2|x_3, \ldots, x_n)P(x_3, \ldots, x_n) \\
&\vdots \\
&= P(x_1|x_2, x_3, \ldots, x_n)\ldots P(x_{n-1}|x_n)P(x_n)
\end{aligned}$$

## 10.3 Bayesian Networks

We will begin the discussion of Bayesian networks based on a simple example.

### 10.3.1 Example 3: Getting a job

It is tough to get our dream jobs. We will consider some of the factors which may influence our chance of getting into our dream job, including:

1. Grades: $G$ indicates that our grades are high, whereas $\bar{G}$ indicates that our grades are low.
2. Job interview: $I$ indicates that our interview went well, whereas $\bar{I}$ indiactes that our interview went horribly.
3. ERP: $E$ indicates that we had to pay high ERP charges, whereas $barE$ indicates that we did not have to pay ERP charges.

In order to come up with a model of the world, the agent would require data for these variables. We can use a survey for our data collection. The table below shows a sample dataset which we might have collected:

| $G$ | $E$ | $I$ | freq. | $Pr$ |
|---|---|---|---|---|
| $T$ | $T$ | $T$ | 160 | $P(G, E, I) = 160/600$ |
| $T$ | $T$ | $F$ | 60 | $P(G, E, \bar{I}) = 60/600$ |
| $T$ | $F$ | $T$ | 240 | $P(G, \bar{E}, I) = 240/600$ |
| $T$ | $F$ | $F$ | 40 | $P(G, \bar{E}, \bar{I}) = 40/600$ |
| $F$ | $T$ | $T$ | 10 | $P(\bar{G}, E, I) = 10/600$ |
| $F$ | $T$ | $F$ | 60 | $P(\bar{G}, E, \bar{I}) = 60/600$ |
| $F$ | $F$ | $T$ | 10 | $P(\bar{G}, \bar{E}, I) = 10/600$ |
| $F$ | $F$ | $F$ | 20 | $P(\bar{G}, \bar{E}, \bar{I}) = 20/600$ |

From this table, the agent would be able to figure out all kinds of probabilities involving the three variables, which may be of its interest. For example,

$$\begin{aligned} P(G) &= P(G, E, I) + P(G, E, \bar{I}) + P(G, \bar{E}, I) + P(G, \bar{E}, \bar{I}) \\ &= \frac{160 + 60 + 240 + 40}{600} \\ &= \frac{500}{600} \end{aligned}$$

However, if there are many variables, the table size may become very large. For instance, if we consider 2 additional variables:

- Job offer: $J$ indicates that we received an offer, $\bar{J}$ indicates that we did not receive an offer.
- Mood of the driver: $D$ indicates that the driver is happy, $\bar{D}$ indicates that the driver is unhappy.

Then, our table will need to have 32 entries (or 31, if we choose to compute one of the entries mathematically instead of storing it). In general, if we have $n$ variables, we would require a table with $O(2^n)$ entries. Is there any way for us to reduce the amount of space needed to store the entries?

### 10.3.2 Networks and Conditional Probability Tables (CPTs)

If we insist on taking note of the frequencies and probabiities of all the events, our table will definitely be large. However, we could aim for our agent to have a causal model of the world instead, i.e. if the outcome of an event $X$ is independent of the outcome of an event $Y$, then we do not need to store $P(X|..., Y)$, where "..." may represent any number of variables. This is because if events $X$ and $Y$ are conditionally independent of each other, then $P(X|..., Y) = P(X|...)$.

Hence, our agent may have the following model of the world:

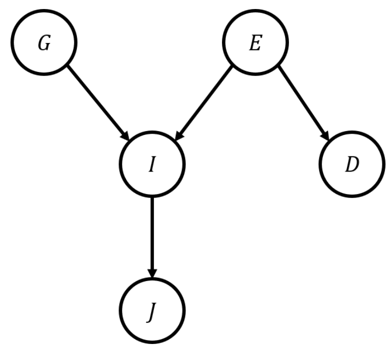


Figure 10.1: Bayesian network model for example 3

This directed acyclic graph (DAG) is also known as a **Bayesian/Inference/Belief Network**. In this graph, we are effectively making the following statements:

- The behavior of $I$ is solely determined by $G$ and $E$.
- The behavior of $J$ is solely determined by $I$.
- The behavior of $D$ is solely determined by $E$.

At each node, we can then construct the respective CPTs. For instance, at node $I$, our CPT would be:

| $G$ | $E$ | $Pr$ |
|---|---|---|
| $T$ | $T$ | $P(I\|G,E) = 160/200$ |
| $T$ | $F$ | $P(I\|G,\bar{E}) = 240/280$ |
| $F$ | $T$ | $P(I\|\bar{G},E) = 10/70$ |
| $F$ | $F$ | $P(I\|\bar{G},\bar{E}) = 10/30$ |

Similarly, we could construct the CPTs at nodes $D$ and $J$ as well, and be able to derive any probabilities that we may require.

### 10.3.3 Space consumption

Did our new model of the problem actually help with reducing the size of our table?

- Without adopting the aforementioned model, a network of $n$ nodes would require a table with $2^n$ entries.
- With the use of Bayesian networks, the number of entries is upper bounded by $n \times 2^p$, where $p$ represents the maximum number of parents that a node in the network can have.

In the previous example, we were able to reduce the number of entries from 32 to 10, merely by adopting the Bayesian network model. Admittedly, if there is a node with $n-1$ parents, then we would not be able to have a significant reduction of entries, but such networks are uncommon.

### 10.3.4 More on Bayesian networks

Bayesian networks have the following **axiom**:

> Each node in a Bayesian network is independent of its *nondescendants*, given its parents.

For instance, with respect to Figure 10.1,

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

However, note that this axiom does not apply to the descendants of nodes. For example,

$$P(I|J,G,E) = \frac{P(I,J,G,E)}{P(J,G,E)} \text{ is not necessarily equal to } P(I|G,E).$$

This highlights the idea that causation implies correlation, but correlation does not necessarily imply causation.

### 10.3.5 Computing probabilities

Suppose we want to compute $P(G,I,J,E,D)$. We can apply the chain rule which we have discussed in a previous section:

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

However, note that the probabilities on the right hand side of the equation cannot be obtained from the CPTs of the Bayesian network, and thus this computation would not be possible.

Instead, we may want to apply the chain rule in a different way, i.e. after rearranging the elements. Then,

$$\begin{aligned} P(G,I,J,E,D) &= P(J,I,G,D,E) \\ &= P(J|I,G,D,E)P(I|G,D,E)P(G|D,E)P(D|E)P(E) \\ &= P(J|I)P(I|G,E)P(G)P(D|E)P(E) && \text{from the axiom of BN} \end{aligned}$$

Now, the probabilities on the right hand side of the equation can be simply extracted from the CPTs of the Bayesian network, and thus the computation would be trivial.

Q: How do we figure out the arrangement/ordering of the variables?

When ordering the variables $X_{i+1}, X_{i+2}, \ldots, X_n$ in $P(X_i, X_{i+1}, X_{i+2}, \ldots, X_n)$, we would want to ensure that:

1. none of these variables is a descendant of the variable $X_i$, and
   - this is because if $X_{i+j}$ for some $j$ is a descendant of $X_i$, then $P(X_i | \ldots, X_{i+j})$ would be hard to compute (because $X_i$ and $X_{i+j}$ are not conditionally independent of each other), and we wouldn't want to have to compute such a probability.
2. it would also be nice to select a variable $X_i$ such that $X_{i+1}, \ldots, X_n$ contains as many parents as possible.

Since the network is a directed acyclic graph, it is possible for us to recursively choose the nodes without children and place them at the start of the ordering; this could be done easily using **topological sort**. Hence, for our previous example, we could pick the nodes in the order $J \to I \to G \to D \to E$, and this arrangement would enable us to compute the probability $P(G, I, J, E, D)$ easily via the chain rule.

In summary, the algorithm for computing probabilities is:

1. Rearrangement of variables.
2. Apply chain rule.
3. Lookup the conditional probabilities from the CPTs.