Lecture 23 - Probability Theory

https://www.cs.ubc.ca/~jordon/teaching/cpsc322/2019w2/lectures/lecture23.pdf

Motivation/Key Points

AI agents (and humans) are not omniscient, and the problem is not only predicting the future or "remembering" the past.

Thus we ask: **Are agents all ignorant/uncertain to the same degree?**

Should an agent only act if it is certain of some relevant knowledge? And can the agent every be fully certain? For instance, not acting can often have implications... Therefore agents need to represent and reason about their ignorance/uncertainty.

Probability as a formal measure of uncertainty/ignorance

We can use probabilities to represent beliefs. For instance consider a simple dice roll. Suppose we want to know whether we rolled a 6. What would be $P(6)$? And in a new scenario, consider the fact that we know in advance that the number rolled is *even*:

- This evidence forces us to update our beliefs.
- What is the new conditional probability $P(6|even)$

Random Variables

A **random variable** is a *variable* (like the ones we have seen in CSP and Planning), but the agent can be uncertain about its value.

As usual:

- The **domain** of a random variable X is $dom(X)$ is the set of values that X can take
- values are mutually exclusive and exhaustive

A **tuple of random variables** $< X_1, ..., X_n >$ is a complex random variable with domain $< dom(X_1) \times ... \times dom(X_n) >$

An assignment $X = x$ means that X takes the value x.

A proposition is a Boolean formula made from assignments of values to variables.

Possible Worlds

A **possible world** specifies an assignment to each random variable. As usual, the possible worlds are mutually exclusive, and exhaustive.

$w \models X = x$ means that variable X was assigned value x in world w

Semantics of Probability

The belief of being in each possible world w can be expressed as a probability $\mu(w)$, and

$$\sum_{w \in W} \mu(w) = 1$$

Probability of a proposition

We ask what is the probability of a proposition f?

For any f, it is the probability of the worlds where f is **true**.

$$P(f) = \sum_{w \models f} \mu(w)$$

Probability Distributions

A probability distribution **P** on a random variable **X** is a function $dom(X) \rightarrow [0, 1]$ such that:

$$x \rightarrow P(X = x)$$

We can represent 3 kinds of beliefs:

1. complete certainty
2. perfect uncertainty
3. some reasonable guess

Joint Probability Distributions

When we have **multiple random variables**, their join distribution is a probability distribution over the variables' Cartesian product:

- Eg.: $P(< X_1, ..., X_n >)$
- We can think of a joint probability distribution over n variables as an n dimensional table
- Each entry indexed by $X_1 = x_1$, ..., $X_n = x_n$ corresponds to $P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n)$
- The sum of entries across the whole table is 1

Some remaining questions

Suppose you have the joint probability distribution of n variables:

- Can you compute the probability distribution for each variable?
- Can you compute the probability distribution for any combination of variables?
- Can you update these probabilities if you know something about some of the variables?
- Is there a downside to the joint probability distribution?

Lecture 24 - Conditional Probability

https://www.cs.ubc.ca/~jordon/teaching/cpsc322/2019w2/lectures/lecture24.pdf

Goals:

- Given a joint probability distribution (JPD), compute distributions over any subset of the variables
- Derive and use the formula to compute conditional probabilities $P(h|e)$
- Derive the **Chain Rule** and **Bayes' Rule**

Joint Distribution and Marginalization

Given a joint distribution, e.g. $P(X, Y, Z)$ we can compute distributions over any smaller sets of variables:

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

Conditioning (Conditional Probability)

We **model our environment** with a set of **random variables**. Assuming we have the **joint**, we can compute the probability of **any formula**.

Probabilistic conditioning specifies how to **revise beliefs based on new information**. You build a probabilistic model (for now the joint) taking all background information into account, which gives us the **prior probability**.

If evidence e is all the new information obtained subsequently, the conditional probability $P(h|e)$ of h given e is the posterior probability of h.

How can we compute P(h|e)

Q: What happens in terms of possible worlds if we know the value of a random var (or a set of random vars)?

A: Some worlds are **ruled out**. The others become **more likely**.

Suppose we have a given value of a **RV** that we know is true. Then we have that:

$$\mu_e(w) = \begin{cases} \frac{\mu(w)}{P(e)} & \text{if } w \models e \\ 0 & \text{otherwise} \end{cases}$$

Then as a result we have:

$$P(h|e) = \sum_{w \models h} \mu_e(w)$$

Thus , some semantics of conditional probability based on the above examples are used below, to derive the conditional probability of formula h given evidence e:

$$\begin{aligned} P(h|e) &= \sum_{w \models h} \mu_e(w) \\ &= \sum_{w \models h \wedge e} \frac{1}{P(e)} \mu(w) \\ &= \frac{1}{P(e)} \sum_{w \models h \wedge e} \mu(w) \\ &= \frac{P(h \wedge e)}{P(e)} \end{aligned}$$

Product Rule

The definition of conditional probability for random variables is:

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)}$$

The product rule gies a more intuitive alternative formulation:

$$P(X_1, X_2) = P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$$

And in general terms we have:

$$\begin{aligned} P(X_1, \ldots X_n) &= P(X_1, \ldots, X_t, X_{t+1}, \ldots, X_n) \\ &= P(X_{t+1}, \ldots, X_n|X_1, \ldots, X_t)P(X_1, \ldots, X_t) \end{aligned}$$

Chain Rule

The chain rule is derived though successive application of the product rule:

$$\begin{aligned} P(X_1, \ldots, X_{n-1}, X_n) &= P(X_1, \ldots, X_{n-1})P(X_n|X_1, \ldots, X_{n-1}) \\ &= P(X_1, \ldots, X_{n-2})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_n|X_1, \ldots, X_{n-1}) \\ &= \cdots = \\ &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \cdot \ldots \cdot P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_n|X_1, \ldots, X_{n-1}) \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

Bayes' Rule and Independence

We often use casual knowledge (forward from cause to evidence). For example:

- P(symptom|disease)
- P(light is off|status of switches and switch positions)
- P(alarm|fire)

In general: P(evidence e|hypothesis h)

And we would rather use evidential reasoning (backwards from evidence to cause). For example:

- P(disease|symptom)
- P(status of switches and switch positions|light is off)
- P(fire|alarm)

In general: P(hypothesis h|evidence e)

Baye's Rule

By definition we know that:

$$P(h|e) = \frac{P(h \wedge e)}{P(e)}$$

and:

$$P(e|h) = \frac{P(e \wedge h)}{P(h)}$$

We rearrange the terms to write:

$$P(h \wedge e) = P(h|e)P(e) \quad (1)$$
$$P(e \wedge h) = P(e|h)P(h) \quad (2)$$

but:

$$P(h \wedge e) = P(e \wedge h) \quad (3)$$

From (1), (2) and (3) we can derive **Bayes' rule**:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

Conditional probability (irrelevant evidence)

New evidence may be irrelevant, allowing simplification. For instance:

$$P(\text{cavity}|\text{toothache}, \text{sunny}) = P(\text{cavity}|\text{toothache})$$

We then say that Cavity is conditionally independent from Weather.

This kind of inference, sanctioned by domain knowledge, is crucial in probabilistic inference

Lecture 25 - Conditional Probability

https://www.cs.ubc.ca/~jordon/teaching/cpsc322/2019w2/lectures/lecture25.pdf

Marginal Independence

Q: Do we always have to revise our beliefs?

NO, not when your knowledge of $\mathbf{Y}$'s value doesn't affect your belief in the value of $\mathbf{X}$

> **Def:** A *RV* $\mathbf{X}$ is said to be marginally independent of random variable $\mathbf{Y}$ if for all $x_i \in \text{dom}(\mathbf{X})$ and all $y_k \in \text{dom}(\mathbf{Y})$:
>
> $$P(X = x_i | Y = y_k) = P(X = x_i)$$
>
> Consequently, $\mathbf{X}$ and $\mathbf{Y}$ are said to be marginally independent if:
>
> $$P(X|Y) = P(X) \quad \text{or} \quad P(Y|X) = P(Y) \quad \text{or} \quad P(X,Y) = P(X)P(Y)$$

Conditional Independence

With marginal independence, for n independent random vars we move from $O(d^n)$ to $O(nd)$ space complexity:

$$P(x_1, \ldots, x_n) = P(x_1) \cdot \ldots \cdot P(x_n)$$

Absolute independence is powerful **but** when you model a **particular domain**, it is **rare**.

We often run into cases with hundreds of variables, few of which are independent. What should be done in that case?

We look for weaker forms of independence.

For example in the toothache example, we may have the following statements that are true:

1. $P(\text{Catch}|\text{Toothache}, \text{Cavity}) = P(\text{Catch}|\text{Cavity})$
   And the following are equivalent statements:
2. $P(\text{Toothache}|\text{Catch}, \text{Cavity}) = P(\text{Toothache}|\text{Cavity})$
3. $P(\text{Toothache}, \text{Catch}|\text{Cavity}) = P(\text{Toothache}|\text{Cavity})P(\text{Catch}|\text{Cavity})$

A proof follows:

$$\begin{aligned}
(1) &\implies P(X|Y,Z) = P(X|Z) \\
&\implies \frac{P(X,Y,Z)}{P(Y,Z)} = \frac{P(X,Z)}{P(Z)} \quad (\implies (2)) \\
&\implies \frac{P(X,Y,Z)}{P(X,Z)} = \frac{P(Y,Z)}{P(Z)} \\
&\implies P(Y|X,Z) = P(Y|Z)
\end{aligned}$$

Additionally we have:

$$
\begin{aligned}
(3) \implies P(X,Y|Z) &= \frac{P(X,Y,Z)}{P(Z)} \\
&= \frac{P(Y,Z)P(X,Z)}{P(Z)} \cdot \frac{1}{P(Z)} \quad \text{(from (2) above)} \\
&= \frac{P(Y,Z)}{P(Z)} \cdot \frac{P(X,Z)}{P(Z)} \\
&= P(Y|Z) \cdot P(X|Z)
\end{aligned}
$$

Conditional Independence: Formal Definition

Sometimes, two variables might not be marginally independent. However, they *become* independent after we observe some third variable.

> **Def:** A random variable **X** is said to be conditionally independent of random variable **Y** given a random variable **Z** if, for all $x_i \in \text{dom}(\mathbf{X})$, $y_k \in \text{dom}(\mathbf{Y})$ and $z_m \in \text{dom}(\mathbf{Z})$:
>
> $$P(X = x_1|Y = y_k, Z = z_m) = P(X = x_1|Y = y_k)$$
>
> In other words, the knowledge of**Y**'s value does not affect the belief in the value of **X**, given a value of **Z**.

Side note: storing distributions

Joint Probability Distribution (JPD): has $O(d^n)$ values:

- But they have to sum to 1,so do we need to store all of them?
- How many do we need to store?

Conditional Probability Table (CPT): has $O(d^n)$ values

- But each row has to sum to 1, so do we need to store all of them?
- How many do we need to store?

Conditional independence: Use

We can write out the join distributions using the chain rule

$$
\begin{aligned}
P(\text{Cavity}, \text{Catch}, \text{Toothache}) &= P(\text{Toothache}|\text{Catch}, \text{Cavity})P(\text{Catch}|\text{Cavity})P(\text{Cavity}) \\
&= P(\text{Toothache}|\text{Cavity})P(\text{Catch}|\text{Cavity})P(\text{Cavity})
\end{aligned}
$$

So how many probabilities do we need to write out?

The use of conditional independence often reduces the size of the representation of the joint distribution from exponential in n to linear in n, where n is the number of variables

Conditional independence is our **most basic** and **robust** form of **knowledge** about **uncertain environments**