Chapter 7

Uncertainty: Representation

Introduction to Artificial Intelligence

Chandra Gummaluru
University of Toronto

Version W22.1

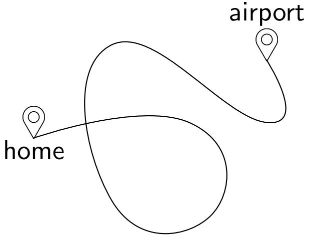
Acknowledgements

- The following is based on material developed by many individuals, including (but not limited to):
  - Sheila McIlraith
  - Bahar Aameri
  - Fahiem Bacchus
  - Sonya Allin

- Part of being an intelligent agent involves being able to make decisions even in uncertain situations.

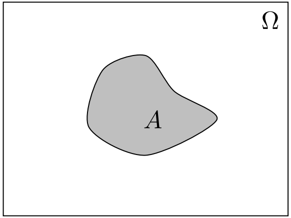
Example: Fastest Path to a Destination

- You have a flight today but are running very late. You need to find the fastest way to the airport. You could:
  - Take public transit
  - Drive yourself
  - Call a cab
- All of the options are reasonable, but it is difficult to tell which is the fastest since there are many factors that influence the travel time.
- We still need to act, but intelligently.



- Probability theory seeks to quantify uncertainty.
- For any experiment, the set of all possible outcomes, denoted Ω, is called its **sample space**.
- In general, Ω can be discrete or continuous, but we will focus only on the former.

- A subset of the sample space is called an **event**.
- An event, $A \subseteq \Omega$ is said to have occurred whenever any outcome $\omega \in A$ occurs.
- The set of all possible events is $\mathcal{P}(\Omega)$.
- The sample-space, Ω, is called the **certain event** since one of its outcomes must necessarily occur.
- The empty set, $\emptyset$, is called the **impossible event** since it has no outcomes.



- The probability of an event is the limit of its relative frequency as the experiment is repeated infinitely many times, i.e.,

$$\Pr\{A\} = \lim_{N \to \infty} \frac{N(A)}{N},$$

where $N(A)$ denotes the number of times A occurs.

- Actually repeating an experiment infinitely many times to calculate probabilities is not practical. Thus, we must manually assign them.

- A **probability measure** is a function, P, such that for each event A,

$$\Pr\{A\} = P(A).$$

The exact mapping is irrelevant, but it must satisfy a few axioms. In particular, $P : \mathcal{P}(\Omega) \to \mathbb{R}$ is a probability measure if and only if it is

1. **non-negative**: $P(A) \geq 0$, for any event, A
2. **additive**: $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
3. **normalized**: $P(\Omega) = 1, P(\emptyset) = 0$
4. **complimentary**: $P(\neg A) = 1 - P(A)$

- By definition, we have

$$
\begin{aligned}
\frac{\Pr\{A \text{ and } B\}}{\Pr\{B\}} &= \lim_{N\to\infty} \frac{N(A\cap B)}{N} \lim_{N\to\infty} \frac{N}{N(B)} \\
&= \lim_{N\to\infty} \frac{N(A\cap B)}{N} \frac{N}{N(B)} \\
&= \lim_{N\to\infty} \frac{N(A\cap B)}{N(B)} \\
&:= \Pr\{A \text{ given } B\}
\end{aligned}
$$

- As such, we may write

$$
P(A|B) = \frac{P(A\cap B)}{P(B)}, P(A) \neq 0,
$$

where $P(A|B)$ is the **conditional probability** of A, given B.

- Since $A \cap B = B \cap A$, it follows that

$$P(B \cap A) = P(A \cap B)$$

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

$$\Leftrightarrow P(B|A) = P(B) \underbrace{\frac{P(A|B)}{P(A)}}_{\text{Bayes' factor}}$$

- Bayes' rule allows us to compute $P(B|A)$, in terms of $P(B)$. In other words, it lets us update our beliefs given new information.

- Two events, A and B are **independent**, denoted $A \perp B$ (or $B \perp A$) iff

$$P(B|A) = P(B) \text{ or } P(A|B) = P(A).$$

- In other words, for independent events, knowledge of one does not influence the probability of the other.
- The events are **conditionally independent** given another event, C, denoted $AQ \perp B|C$ iff

$$P(B|A, C) = P(B|C) \text{ or } P(A|B, C) = P(A|C).$$

- In other words, for conditionally independent events, knowledge of one does not influence the probability of the other, provided the condition is satisfied.

- To reason about events more mathematically, it is useful to map them to a numerical space.

- Such a mapping is called a **random variable**.

- Since we are assuming Ω is discrete, we can simply use $\mathbb{N}$ as the numerical space.

- For any (discrete) random variable, $X : \Omega \to \mathbb{N}$, we define a **probability mass function**, $p_X : \mathbb{N} \to [0, 1]$ such that

$$p_X(x) := \Pr\{X = x\} = P(\{\omega \in \Omega : X(\omega) = x\}).$$

- We can similarly define a **joint probability mass function**, $p_{X,Y} : \mathbb{N}^2 \to [0, 1]$ between two random variables, X, Y such that

$$p_{X,Y}(x, y) := \Pr\{X = x \text{ and } Y = y\} = P(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}).$$

- We can compute the **conditional probability mass function**, $p_{X|Y} : \mathbb{N}^2 \to [0,1]$ between X and Y, as

$$p_{X|Y}(x,y) := \frac{p_{X,Y}(x,y)}{p_Y(y)}, p_Y(y) \neq 0.$$

- Given a joint probability mass function, $p_{X,Y}$, we can compute the (marginal) probability mass function as follows:

$$p_X(x) = \Pr\{X = x\}$$

$$= \sum_{\forall y} \Pr\{X = x \text{ and } Y = y\}$$

$$= \sum_{\forall y} p_{X,Y}(x,y).$$

- An uncertain situation can be modelled as a set of random, but related, variables.

- In general, these relationships can be incredibly complex.

Example: Modelling Uncertain Situations

- Let L represent whether we are late to the airport or not. Clearly, its value depends on a multitude of factors such as:
  - the mode of transit, M, that we choose
  - the weather conditions, W
  - the traffic conditions, T
  - the day of the week, D
  - whether it is rush hour or not, R
  - whether there are public transit delays or not, P
- However, the exact influence of these factors is unclear.
  E.g: Suppose we know there are public transit and traffic delays, but choose to drive. Does it matter whether it is rush hour or not?

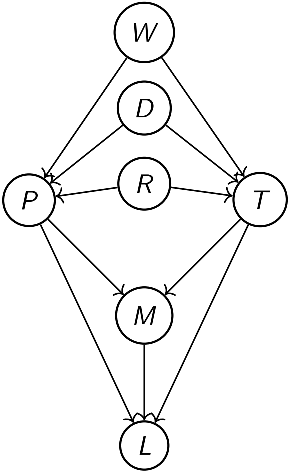
- We need a systematic way to represent dependence relationships.

- A **Bayesian network** is a directed graph, $(\mathcal{V}, \mathcal{E})$, in which the vertices represent the related random variables, and the edges represent the dependence relationships between them:
  - If $(V_1, V_2) \in \mathcal{E}$, where $V_1, V_2 \in \mathcal{V}$, then V_1 and V_2 are dependent.
  - However, the converse is not necessarily true.
  - The directions of the arcs is technically irrelevant since dependence is commutative.
  - The arcs typically indicate the direction of causality but need not be.
- **Example**: Causality in a Bayesian Network
  - Both networks below indicate that X_1 and X_2 are dependent, but the left suggests that X_1 causes X_2, while the right suggests that X_2 causes X_1.

$$X_1 \longrightarrow X_2 \qquad\qquad X_1 \longleftarrow X_2$$

Example: Modelling Uncertain Situations

- We can define a Bayesian network over $\mathcal{V} = \{L, M, W, T, D, R, P\}$.
- One possible network is shown on the right.
- The network says the following:
- W, D, and R all influence P and T, but are mutually independent.
- P and T influence M
- P, T, and M influence L.



- When expressed as a Bayesian network, we can graphically determine the conditional dependence relationships between any subset of the variables given knowledge of another subset of the variables.

- To develop it, we shall first consider the so-called "junction" network.

- A **junction**, $\mathcal{J} = (\{X_1, X_2, X_3\}, \{(e_1, e_2)\})$ is a Bayesian network that consists of three random variables, X_1, X_2, X_3, connected by two arcs, e_1 and e_2.

- It turns out that conditional mutual dependence relationships between the variables are directly tied to their causal relationships.

- Three types of causal relationships are possible.

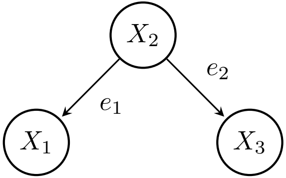
1. When one variable influences another, which in turn influences the third, as shown below, the relation is called a **causal chain**:

$$X_1 \xrightarrow{e_1} X_2 \xrightarrow{e_2} X_3$$

E.g: The number of courses you take influences how busy you are, which influences how tired you are.

In this case, X_1 and X_3 are dependent, but independent given X_2.

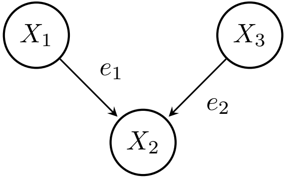
2. When one variable directly influence the other two, as shown below, the relation is referred to as a **common cause**:



E.g: Rain causes your hair to get wet and the road to be slippery.

In this case, X_1 and X_3 are dependent, but independent given X_2.

3. When one variable is directly influenced by the other two, as shown below, the relation is called a **common effect**:



E.g: Both rain and watering the garden can cause the grass to be wet.

In this case, X_1 and X_3 are independent, but dependent given X_2, or any of its successors.

- Let $\mathcal{B} = (\mathcal{V}, \mathcal{E})$ be some Bayesian network and $\mathcal{K} \subseteq \mathcal{V}$.

- A junction, $\mathcal{J} = (\{X_1, X_2, X_3\}, \{(e_1, e_2)\})$, of $\mathcal{B}$, is **blocked** under K if X_1 and X_2 are independent given K.

- Let a path, p, in $\mathcal{B}$ be an ordered list of junctions,
$$\mathcal{J}^{(i)} = \left(\left\{X_1^{(i)}, X_2^{(i)}, X_3^{(i)}\right\}, \left\{(e_1^{(i)}, e_2^{(i)})\right\}\right)$$
such that $X_1^{(i+1)} = X_2^{(i)}$ and $X_2^{(i+1)} = X_3^{(i)}$, or equivalently, $e_1^{(i+1)} = e_2^{(i)}$.

- The path, p, is **blocked** under K if $\mathcal{J}^{(i)}$ is blocked under K for some i.

- **Theorem**: Conditional Independence in Bayesian Networks
  - The variables, $V_1, V_2 \in \mathcal{V}$ are conditionally independent given K if and only if every path between V_1 and V_2 is blocked under K.

Example: Getting to the Airport (Dependence Separation)

- Suppose we know there are public transit and traffic delays, but choose to drive. Does it matter whether it is rush hour or not?
- In other words, given P, T, and M, are R and L dependent or not?
- Using the network on the right, we see that R and L are independent given P, T, and M.
- Thus, it is irrelevant whether we leave during rush hour or not.