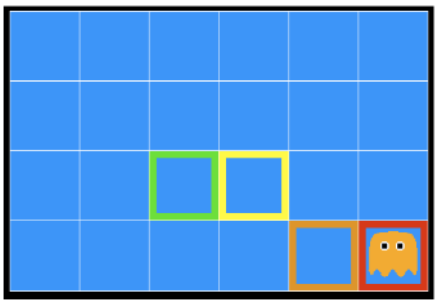
Uncertainty and Bayes Nets

Uncertainty

In this section we cover the background necessary to build agents who are capable of modeling and taking actions face of uncertainty about their environment. This knowledge will allow you to create a Pacman agent for the ghostbusters assignment that is able to track down ghost despite imperfect "sensor" readings, where given any individual sensor reading Pacman knows the probability distribution of that sensor reading given the ghosts potential distance.



In order to begin dealing with uncertainty, we first need a way to quantify and define uncertainty. A **random variable** represents some aspect of the world which has uncertainty, and is typically denoted with a single capital letter, e.g.

$S =$ is it sunny outside?
$T =$ what is the current temperature outside?

Every random variable has some domain which it's values will lie inside. For instance, the domain for S would be $\{true, false\}$, or since it is a boolean variable, another common notation would be $\{+s, -s\}$. The domain of T would be $\{0, \infty\}$ (assuming T is given in Kelvin).

A **probability distribution** for a random variable associates probabilities with outcomes or values that variable can take on, and can be represented as a table for discrete random variables. Unobserved random variables have distributions.

$$P(S)$$

S	P
sunny	0.5
cloudy	0.5

For any random variable X, denoted as an upper case letter, a lower case value probability is a single number, e.g. $P(S = sunny) = 0.5$

A common shorthand could express the same as $P(sunny) = 0.5$, this is valid if all domains are distinct. For a probability distribution table to be valid, the following must hold:

$$\forall x \; P(X = x) \geq 0$$

$$\sum_{x} P(X = x) = 1$$

An **event** is a set of outcomes. A **joint distribution** for some set of random variables X, Y, Z provides a real number probability for each possible assignment or event.

$$P(T, W)$$

T	W	P
hot	sunny	0.4
cold	sunny	0.1
hot	cloudy	0.2
cold	cloudy	0.3

A joint distribution must obey similar rules to a probability distribution, adapted to multiple random variables:

$$\forall x_1, x_2, ...x_n \; P(x_1, x_2, ...x_n) \geq 0$$

$$\sum_{x_1, x_2, ...x_n} P(X_1 = x_1, X_2 = x_2, ...X_n = x_n) = 1$$

A **marginal distribution** is a sub-table which eliminates other variables. The process of **marginalization** or summing out is the process of collapsing rows and adding their probabilities to produce a marginal distribution.

$P(T, W) =$

T	W	P
hot	sunny	0.4
cold	sunny	0.1
hot	cloudy	0.2
cold	cloudy	0.3

$P(t) = \sum_w P(t, w) \; \rightarrow$ P(T) =

T	P
hot	0.5
cold	0.5

P(w) = $\sum_t P(t, w) \; \rightarrow \; P(W) =$

W	P
sunny	0.6
cloudy	0.4

Below is the definition for the relation between conditional and joint probabilities. A **conditional probability** is the probability of an event occurring given that another event has already occurred.

$$P(a|b) = \frac{P(a,b)}{P(b)}$$

By rearranging, we can determine that when we have the conditional distribution but want the joint distribution, we can use the **product rule**.

$$P(y)P(x|y) = P(x,y)$$

We can say two variables are **independent** if $\forall x, y P(x,y) = P(x)P(y)$, or $\forall x, y P(x|y) = P(x)$.Models are always simplifications, which means they cannot account for every variable or interaction between variables. Independence is typically a simplifying, modeling assumption, meaning the empirical joint distributions are "close" to independent.

Unconditional independence is extremely rare. On the other hand **conditional independence**, a concept indicating that the occurrence or value of one random variable is independent of another given the knowledge of a third variable, is much more common. Two variables are conditionally independent if:

$$\forall x, y, z P(x,y|z) = P(x|z)P(y|z)$$

Independence assumptions can be useful for decreasing the space complexity of storing many variables. For instance, the joint distribution for n boolean variables (such as fair coin flips) would require a table with 2^n rows, but by assuming independence this could be reduced to n tables with 2 rows each.
The **chain rule** allows the decomposition of a complex joint probability into many conditional probabilities.

$$P(x_1, x_2, ...x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)...$$

Bayes Nets

Bayesian networks, or bayes nets, are probabilistic graphical models that represent and quantify the probabilistic relationships among a set of random variables using a directed acyclic graph and several conditional probability tables.

Nodes in the graph represent variables and edges represent direct influence. Variables can be observed or unobserved, and edges are directional and encode conditional independence. There is one node per variable, and one conditional probability table (CPT) for each node. A CPT is a collection of distributions over some variable, containing one entry for each combination of parent

values.

Bayes nets are able to implicitly encode joint distributions as a product of local conditional distributions. We can calculate the probability the network gives to any specific assignment by multiplying all the conditionals together:

$$P(x_1, x_2, ...x_n) = \prod_{i=1}^{n} P(x_1|parents(X_i))$$

Bayes nets assume conditional independence. This means that not every network can represent every joint distribution, because the particular topology enforces certain conditional independencies which may not hold well for certain joint distributions.

For a bayesian network to represent n variables, if m is the largest number of parent nodes of any node in the network, the space complexity will be $n * 2^m$. For the same number of variables to be represented by a table with no conditional independence assumptions, the space complexity will be 2^n. Since in practice $n >> m$, this means the space complexity of a bayesian network is far better.

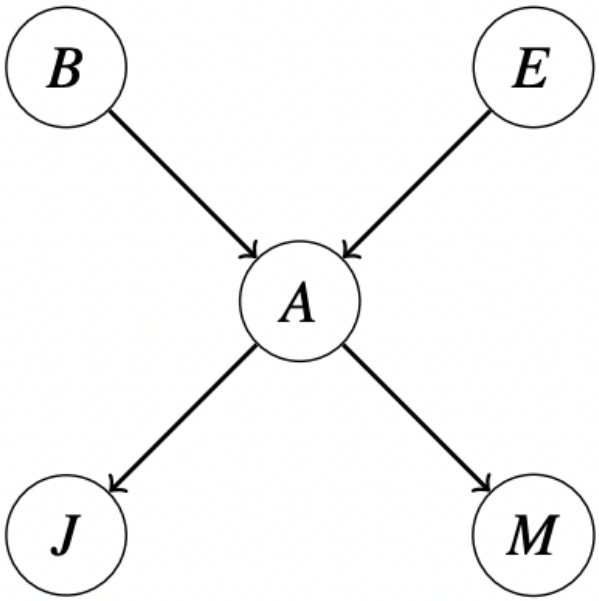
When we want to reason about a sequence of changing observations over time or space. In order to simplify the process of reasoning over changing observations, we have the **stationarity assumption**, which is the assumption that transition probabilities between states (in time, for example) are always the same (don't change over time). There is also the **markov assumption**, which is the assumption that each time step only depends on the previous time step and is independent of other past or future states.

Example of Bayes Net Representation

As an example of a Bayes Net, consider a model where we have five binary random variables described below:

- B: Burglary occurs.
- A: Alarm goes off.
- E: Earthquake occurs.
- J: John calls.
- M: Mary calls.

Assume the alarm can go off if either a burglary or an earthquake occurs, and that Mary and John will call if they hear the alarm. We can represent these dependencies with the graph shown below.



In this Bayes Net, we would store probability tables $P(B)$,$P(E)$,$P(A|B,E)$,$P(J|A)$ and $P(M|A)$. Given all of the CPTs for a graph, we can calculate the probability of a given assignment using the following rule:

$$P(x_1, x_2, ...x_n) = \prod_{i=1}^{n} P(x_1|parents(X_i))$$

For the alarm model above, we can actually calculate the probability of a joint probability as follows:

$$P(b, e, +a, +j, m) = P(b)P(e)P(+a|b, e)P(+j|+a)P(m|+a)$$

We will see how this relation holds in the next section.

As a reality check, it's important to internalize that Bayes Nets are only a type of model. Models attempt to capture the way the world works, but because they are always a simplification they are always wrong. However, with good modeling choices they can still be good enough approximations that they are useful for solving real problems in the real world.

In general, a good model may not account for every variable or even every interaction between variables. But by making modeling assumptions in the structure of the graph, we can produce incredibly efficient inference techniques that are often more practically useful than simple procedures like inference by enumeration.

Causal Chains

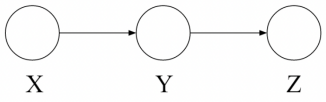


Figure 1: Causal Chain with no observations.

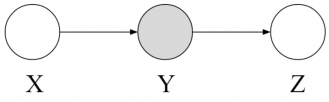


Figure 2: Causal Chain with *Y* observed.

Figure 1 is a configuration of three nodes known as a causal chain. It expresses the following representation of the joint distribution over X, Y, and Z:

$$P(x, y, z) = P(z|y)P(y|x)P(x)$$

It's important to note that X and Z are not guaranteed to be independent, as shown by the following counterexample:

$$P(y|x) = \begin{cases} 1 & x = y \\ 0 & else \end{cases}$$

$$P(z|y) = \begin{cases} 1 & z = y \\ 0 & else \end{cases}$$

However, we can make the statement that X is independent of Z | Y, as in Figure 2. Recall that this conditional independence:

$$P(X|Z, Y) = P(X|Y)$$

We can prove this statement as follows:

$$P(X|Z, y) = \frac{P(X, Z, y)}{P(Z, y)} = \frac{P(Z|y)P(y|X)P(X)}{\sum_x P(X, y, Z)} = \frac{P(Z|y)P(y|X)P(X)}{P(Z|y)\sum_x P(y|x)P(x)} = \frac{P(y|X)P(X)}{\sum_x P(y|x)P(x)}$$
$$= \frac{P(y|X)P(X)}{\sum_x P(y|x)P(x)} = \frac{P(y|X)P(X)}{P(y)} = P(X|y)$$

An analogous proof can be used to show the same thing for the case where X has multiple parents. To summarize, in the causal chain chain configuration X is independent of Z | Y.

Summary

Models are imperfect and make simplifying assumptions about the real world. If we are careful about these assumptions (like assumed independence or conditional independence) these can have minimal impact on accuracy but bring massive computational simplification. Bayes Nets are models that represent the probabilistic relationships among a set of random variables.

- We can describe the probabilistic profile of variables with distributions, and the relationships between them with joint and conditional distributions.

- Absolute v.s. conditional independence. Truly independent variables are extremely rare in practice. Conditional independence is very powerful, and much more common.
- There are two major problems with full joint distribution tables. First, unless there is a trivially small number of variables, the resulting table is far too large to explicitly represent. Secondly, it is hard to learn anything about more than a few variables at a time.
- Bayesian networks implicitly encode joint distributions, and with vastly lower space requirements.
- Not every bayesian network can represent every joint distribution.