Probability

"The mind is a neural computer, fitted by natural selection with combinatorial algorithms for causal and probabilistic reasoning about plants, animals, objects, and people. It is driven by goal states that served biological fitness in ancestral environments, such as food, sex, safety, parenthood, friendship, status and knowledge."

Steven Pinker, How the Mind Works, 1997

Learning Objectives

At the end of the class you should be able to:

- justify the use and semantics of probability
- know how to compute marginals and apply Bayes' theorem
- build a belief network for a domain
- next time: predict the inferences for a belief network
- next time: explain the predictions of a causal model

Using Uncertain Knowledge

- Agents don't have complete knowledge about the world.
- Agents need to make decisions based on their uncertainty.
- It isn't enough to assume what the world is like.
- An agent needs to reason about its uncertainty.
- When an agent makes an action under uncertainty, it is gambling $\Longrightarrow$ probability.

Why Probability?

- There is lots of uncertainty about the world, but agents still need to act.
- Predictions are needed to decide what to do:
  - definitive predictions: you will be run over tomorrow
  - point probabilities: probability you will be run over tomorrow is 0.002
  - probability ranges: you will be run over with probability in range [0.001,0.34]
- Acting is gambling: agents who don't use probabilities will lose to those who do
- Probabilities can be learned from data.
- Bayes' rule specifies how to combine data and prior knowledge.

Probability

- Bayesian probability is used to express an agent's measure of belief in some proposition
  — **subjective probability.**
- An agent's belief depends on its prior assumptions AND what the agent observes.
- Assume uncertainty is **epistemological**, pertaining to an agent's knowledge of the world, rather than **ontological**, or how the world is

Numerical Measures of Belief

- Belief in proposition, f, can be measured in terms of a number between 0 and 1 — this is the **probability of f.**
  - The probability f is 0 means that f is believed to be definitely false.
  - The probability f is 1 means that f is believed to be definitely true.
  - f has a probability between 0 and 1, means the agent is ignorant of its truth value.
- Using 0 and 1 is purely a convention.
- Probability is a measure of an agent's ignorance.
- Probability is *not* a measure of degree of truth.

Random Variables

- A **random variable** is a term that can take one of a number of different values (not really random)
- The **range** of a variable X, written $range(X)$, is the set of values X can take.
- A tuple of random variables $\langle X_1, \ldots, X_n \rangle$ has $range(X_1) \times \cdots \times range(X_n)$.
  Often the tuple is written as $X_1, \ldots, X_n$.
- Assignment $X = x$ means variable X has value x.
- A **proposition** is a Boolean formula made from assignments of values to variables.

Possible World Semantics

- A **possible world** specifies an assignment of one value to each random variable.
- A **random variable** can be a function from possible worlds into the range of the random variable.
- $\omega \models X = x$
  means variable X is assigned value x in world ω.
  Note: This is not the same $\models$ as in KB.
- Logical connectives have their standard meaning:

  $\omega \models \alpha \wedge \beta$ if $\omega \models \alpha$ and $\omega \models \beta$

  $\omega \models \alpha \vee \beta$ if $\omega \models \alpha$ or $\omega \models \beta$

  $\omega \models \neg\alpha$ if $\omega \not\models \alpha$
- Let Ω be the set of all possible worlds.

Semantics of Probability: general case

A probability measure over the worlds is a function μ from sets of worlds into the non-negative real numbers. We define $\mu(\omega)$ for some sets $\omega \subseteq \Omega$ satisfying:

- $\mu(\omega) \geq 0$
- $\mu(\Omega) = 1$
- $\mu(\omega_1 \cup \omega_2) = \mu(\omega_1) + \mu(\omega_2)$ if $\omega_1 \cap \omega_2 = \{\}$.
  Or sometimes σ-additivity:

$$\mu(\bigcup_i \omega_i) = \sum_i \mu(\omega_i) \text{ if } \omega_i \cap \omega_j = \{\} \text{ for } i \neq j$$

Semantics of Probability

For a finite number of possible worlds:

- The **probability** of proposition α, written $P(\alpha)$, is the measure of the set of possible worlds in which α is true: $P(\alpha) = \mu(\{\omega : \omega \models \alpha\})$

$$P(\alpha) = \sum_{\omega \models \alpha} \mu(\omega)$$

Probability Distributions

- Probability distribution, $P(X)$, over a random variable X is a function with $range(X) \to [0,1]$ from the domain of X into the real numbers such that, given a value $x \in dom(X)$

$$x \mapsto P(X = x).$$

- This is written as: $P(X)$ is the probability of the proposition $X = x$
- This also includes the case where we have tuples of variables. E.g., $P(X, Y, Z)$ means $P(\langle X, Y, Z \rangle)$.

Probability density functions

- When $range(X)$ is continuous sometimes we need a probability density function...
- A probability density function, p, is a function from reals into non-negative reals that integrates to 1.
- Probability that a real-valued random variable X has value between a and b is given by

$$P(a \leq X \leq b) = \int_a^b p(X)dX$$

Axioms of Probability: finite case

Three axioms define what follows from a set of probabilities:

Axiom 1 $0 \leq P(\alpha)$ for any proposition α.

Axiom 2 $P(true) = 1$ or $P(\tau) = 1$ where τ is any tautology (obviously true)

Axiom 3 If a and b cannot both be true, $P(\alpha \vee \beta) = P(a) + P(b)$

- If there are a finite number of discrete random variables, Axioms 1, 2, and 3 are sound and complete with respect to the semantics.

Axioms of probability cont..

The following hold for all propositions α and β:

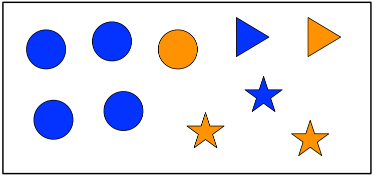
- Negation of a proposition:
  $P(\neg\alpha) = 1 - P(\alpha)$
- Logically equivalent propositions have the same probability:
  If $\alpha \leftrightarrow \beta$, then $P(\alpha) = P(\beta)$
- Reasoning by cases:
  $P(\alpha) = P(\alpha \wedge \beta) + P(\alpha \wedge \neg\beta)$
- If V is a random variable with domain D, then, for all propositions α:
  $P(\alpha) = \sum_{d \in D} P(\alpha \wedge V = d)$
- Disjunction for non-exclusive propositions:
  $P(\alpha \vee \beta) = P(\alpha) + P(\beta) - P(\alpha \wedge \beta)$

Conditioning

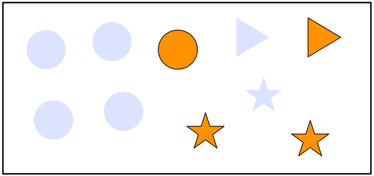
- Probabilistic conditioning specifies how to revise beliefs based on new information.
- An agent builds a probabilistic model taking all background information into account. This gives the **prior probability.**
- All other information must be conditioned on.
- **evidence** e is all information newly obtained
- h is a hypothesis or proposition
- **conditional probability** of h given e is stated as $P(h|e)$, defined as **posterior probability** of h.

Conditioning

Possible Worlds:
What is the probability of star?:



Observe *Color* = *orange*:
What is the probability of star given orange?



Conditional probabilities: joint distribution

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega)$$

Conditional probabilities: joint distribution

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	.108	.012	.072	.008
$\neg$ *cavity*	.016	.064	.144	.576

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Conditional probabilities: joint distribution

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$

$P(cavity \vee toothache) = 0.108+0.012+0.072+0.008+0.016+0.064 = 0.28$

Conditional probabilities: joint distribution

Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	.108	.012	.072	.008
$\neg$ *cavity*	.016	.064	.144	.576

Can also compute conditional probabilities:

$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Conditional probabilities: joint distribution

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

Denominator can be viewed as a normalization constant α

$$\begin{aligned}
&\mathbf{P}(Cavity|toothache) = \alpha\,\mathbf{P}(Cavity, toothache)\\
&\quad = \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]\\
&\quad = \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]\\
&\quad = \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}$$

General idea: compute distribution on query variable
by fixing evidence variables and summing over hidden variables

α AKA c on next slide (c is better, since α is already used in stats)

Semantics of Conditional Probability

- Evidence e rules out possible worlds incompatible with e.
- Evidence e induces a new measure, μ_e, over possible worlds

$$\mu_e(S) = \begin{cases} c \times \mu(S) & \text{if } \omega \models e \text{ for all } \omega \in S \\ 0 & \text{if } \omega \not\models e \text{ for all } \omega \in S \end{cases}$$

  We can show $c = \frac{1}{P(e)}$.
- The conditional probability of formula h given evidence e is

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

Chain rule intro

If independent: $P(A \land B) = P(A)P(B)$
This is true whether or not A and B are independent:
$P(A \land B) = P(A|B)P(B) = P(B|A)P(A)$

Generalize to the chain rule

$$
\begin{aligned}
& P(f_1 \wedge f_2 \wedge \ldots \wedge f_n) \\
= \quad & P(f_n | f_1 \wedge \cdots \wedge f_{n-1}) \times \\
& P(f_1 \wedge \cdots \wedge f_{n-1}) \\
= \quad & P(f_n | f_1 \wedge \cdots \wedge f_{n-1}) \times \\
& P(f_{n-1} | f_1 \wedge \cdots \wedge f_{n-2}) \times \\
& P(f_1 \wedge \cdots \wedge f_{n-2}) \\
= \quad & P(f_n | f_1 \wedge \cdots \wedge f_{n-1}) \times \\
& P(f_{n-1} | f_1 \wedge \cdots \wedge f_{n-2}) \\
& \times \cdots \times P(f_3 | f_1 \wedge f_2) \times P(f_2 | f_1) \times P(f_1) \\
= \quad & \prod_{i=1}^{n} P(f_i | f_1 \wedge \cdots \wedge f_{i-1})
\end{aligned}
$$

Baye's theorem

The chain rule and commutativity of conjunction ($h \wedge e$ is equivalent to $e \wedge h$) gives us:

$$\begin{aligned} P(h \wedge e) &= P(h|e) \times P(e) \\ &= P(e|h) \times P(h). \end{aligned}$$

If $P(e) \neq 0$, divide the right hand sides by $P(e)$:

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

This is **Bayes' theorem.**

Background knowledge, k, is often implicit:
$P(h|e \wedge k) = \frac{P(e|h \wedge k) \times P(h|k)}{P(e|k)}$

Why is Bayes' theorem interesting?

- Often you have causal knowledge:
  $P(symptom \mid disease)$
  $P(light\ is\ off \mid status\ of\ switches\ and\ switch\ positions)$
  $P(alarm \mid fire)$
- and want to do evidential reasoning:
  $P(disease \mid symptom)$
  $P(status\ of\ switches \mid light\ is\ off\ and\ switch\ positions)$
  $P(fire \mid alarm)$.

Bayes theorem illustrated

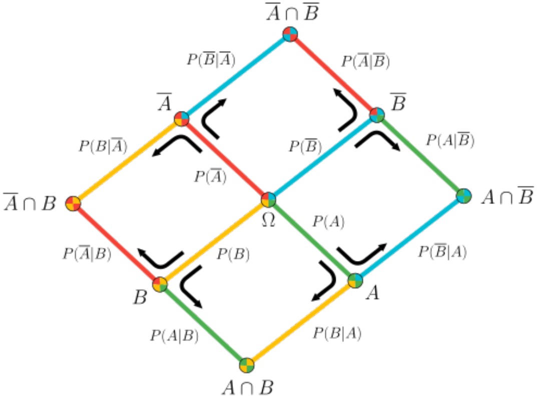
Relative size	Case B	Case $\bar{B}$	Total
Condition A	w	x	$w+x$
Condition $\bar{A}$	y	z	$y+z$
Total	$w+y$	$x+z$	$w+x+y+z$

$$P(A|B) \times P(B) = \frac{w}{w+y} \times \frac{w+y}{w+x+y+z} = \frac{w}{w+x+y+z}$$

$$P(B|A) \times P(A) = \frac{w}{w+x} \times \frac{w+x}{w+x+y+z} = \frac{w}{w+x+y+z}$$

Values w, x, y and z give relative weights of each corresponding condition and case. Figures denote the of the table involved in each metric, with probability being the fraction of each figure that is shaded. Shows that $P(A|B)P(B) = P(B|A)P(A)$

Bayes theorem illustrated



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

For each sub-diamond, product of opposing color-pairs are equal, and specify the process needed to derive the missing value.

Expected values

- A variable's expected value is the variable's weighted average value, where its value in each possible world is weighted by the measure of the possible world.
- Suppose V is a random variable whose domain is numerical, and ω is a possible world. Define $V(\omega)$ to be the value v in the domain of V such that $\omega \models V = v$. That is, we are treating a random variable as a function on worlds.
- The expected value of numerical variable V is: $E(V) = \sum_{\omega \in \Omega} V(\omega)\mu(\omega)$

Simple independence

In probability theory, two events are independent if the occurrence of one does not affect the probability of the other. Similarly, two random variables are independent if the realization of one does not affect the probability distribution of the other:

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

Conditional independence

Random variable X is conditionally independent of random variable Y **given** random variable Z:

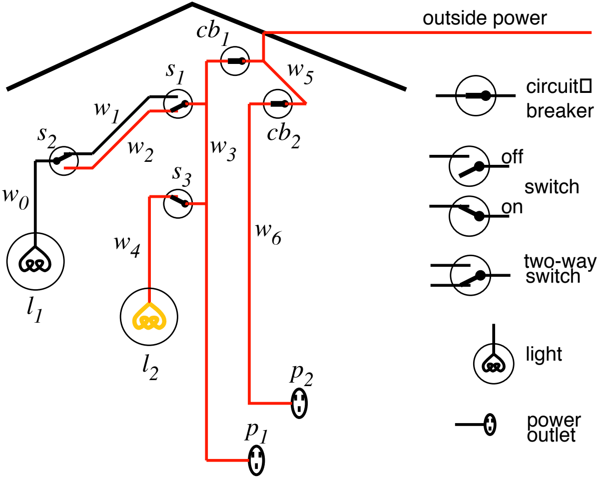
$$P(X|YZ) = P(X|Z)$$

i.e. for all $x_i \in dom(X)$, $y_j \in dom(Y)$, $y_k \in dom(Y)$ and $z_m \in dom(Z)$,

$$\begin{aligned} & P(X = x_i | Y = y_j \wedge Z = z_m) \\ & \quad = \quad P(X = x_i | Y = y_k \wedge Z = z_m) \\ & \quad = \quad P(X = x_i | Z = z_m). \end{aligned}$$

That is, knowledge of Y's value doesn't affect the belief in the value of X, given a value of Z.

Example domain (diagnostic assistant)

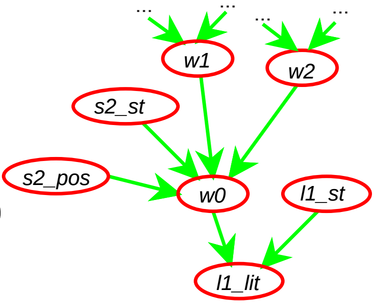


Examples of conditional independence?

- Whether there is someone in a room is independent of whether a light $l2$ is lit given what?
  The position of switch $s3$.
- Whether light $l1$ is lit is independent of the position of light switch $s2$ given what?
  Whether there is power in wire w_0.
- Every other variable may be independent of whether light $l1$ is lit given whether there is power in wire w_0 and the status of light $l1$ (if it's *ok*, or if not, how it's broken).

Idea of belief networks

- $l1$ is lit ($L1\_lit$) depends only on the status of the light ($L1\_st$) and whether there is power in wire $w0$.
- In a belief network, $W0$ and $L1\_st$ are **parents** of $L1\_lit$.



- $W0$ depends only on whether there is power in $w1$, whether there is power in $w2$, the position of switch $s2$ ($S2\_pos$), and the status of switch $s2$ ($S2\_st$).

Belief networks

- Order the variables of interest: $X_1, \ldots, X_n$ via congruence with chain rule: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})$
- The **parents** $parents(X_i)$ of X_i are those predecessors of X_i that render X_i independent of the other predecessors. That is:
  - $parents(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ and
  - $P(X_i | parents(X_i)) = P(X_i | X_1, \ldots, X_{i-1})$
- So $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))$
- A **belief network** is a graph: the nodes are random variables; there is an arc from the parents of each node into that node.

Example: fire alarm belief network

Variables:

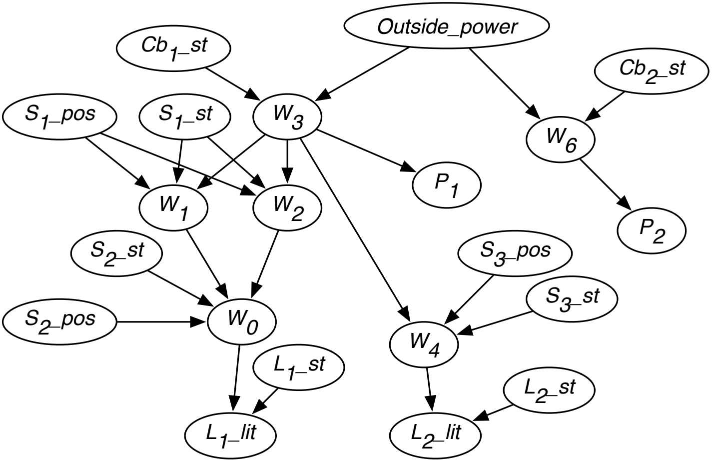
- Fire: there is a fire in the building
- Tampering: someone has been tampering with the fire alarm
- Smoke: what appears to be smoke is coming from an upstairs window
- Alarm: the fire alarm goes off
- Leaving: people are leaving the building *en masse*.
- Report: a colleague says that people are leaving the building *en masse*. (A noisy sensor for leaving.)

Components of a belief network

A belief network consists of:

- a directed acyclic graph with nodes labeled with random variables
- a domain for each random variable
- a set of conditional probability tables for each variable given its parents (including prior probabilities for nodes with no parents).

Example belief network



Example belief network (continued)

The belief network also specifies:

- The domain of the variables:
  - $W_0, \ldots, W_6$ have domain $\{live, dead\}$
  - S_1_pos, S_2_pos, and S_3_pos have domain $\{up, down\}$
  - S_1_st has $\{ok, upside\_down, short, intermittent, broken\}$.
- Conditional probabilities, including:
  - $P(W_1 = live | s_1\_pos = up \wedge S_1\_st = ok \wedge W_3 = live)$
  - $P(W_1 = live | s_1\_pos = up \wedge S_1\_st = ok \wedge W_3 = dead)$
  - $P(S_1\_pos = up)$
  - $P(S_1\_st = upside\_down)$

Belief network summary

- A belief network is a directed acyclic graph (DAG) where nodes are random variables.
- The **parents** of a node *n* are those variables on which *n* directly depends.
- A belief network is automatically acyclic by construction.
- A belief network is a graphical representation of dependence and independence:
  - A variable is independent of its **non-descendants** given its parents.
- Belief networks model causal systems, hypotheses, and interventions; if someone were to artificially force a variable to have a particular value, the variable's descendants, but no other nodes, would be affected.

Constructing belief networks

To represent a domain in a belief network, you need to consider:

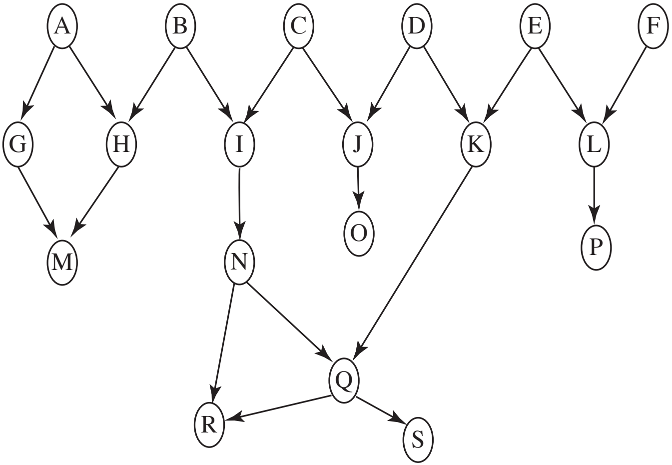
- What are the relevant variables?
  - What will you observe?
  - What would you like to find out (query)?
  - What other features make the model simpler?
  - Hidden or latent variables
- What values should these variables take?
- What is the relationship between them? This should be expressed in terms of a directed graph, representing how each variable is generated from its predecessors.
- How does the value of each variable depend on its parents? This is expressed in terms of the conditional probabilities.

Using belief networks

The power network can be used in a number of ways:

- Conditioning on the status of the switches and circuit breakers, whether there is outside power and the position of the switches, you can simulate the lighting.
- Given values for the switches, the outside power, and whether the lights are lit, you can determine the posterior probability that each switch or circuit breaker is *ok* or not.
- Given some switch positions and some outputs and some intermediate values, you can determine the probability of any other variable in the network.

Understanding independence: example



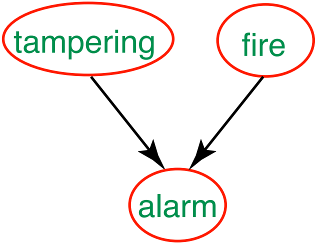
Understanding independence: questions

- On which given probabilities does $P(N)$ depend?
- If you were to observe a value for B, which variables' probabilities will change?
- If you were to observe a value for N, which variables' probabilities will change?
- Suppose you had observed a value for M; if you were to then observe a value for N, which variables' probabilities will change?
- Suppose you had observed B and Q; which variables' probabilities will change when you observe N?

What variables are affected by observing?

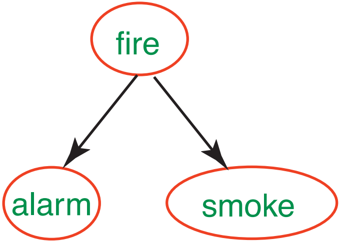
- If you observe variable $\overline{Y}$, the variables whose posterior probability is different from their prior are:
  - The ancestors of $\overline{Y}$ and
  - their descendants.
- Intuitively (if you have a causal belief network):
  - You do **abduction** to possible causes and
  - **prediction** from the causes.

Common descendants



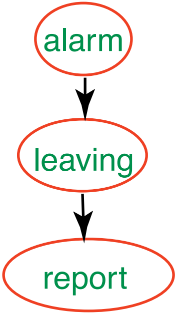
- *tampering* and *fire* are independent
- *tampering* and *fire* are dependent given *alarm*
- Intuitively, *tampering* can **explain away** *fire*

Common ancestors



- *alarm* and *smoke* are dependent
- *alarm* and *smoke* are independent given *fire*
- Intuitively, *fire* can **explain** *alarm* and *smoke*; learning one can affect the other by changing your belief in *fire*.

Chain



- *alarm* and *report* are dependent
- *alarm* and *report* are independent given *leaving*
- Intuitively, the only way that the *alarm* affects *report* is by affecting *leaving*.

For next time...

The most common probabilistic inference task is to compute the posterior distribution of a query variable given some evidence. Unfortunately, even the problem of estimating the posterior probability in a belief network within an absolute error (of less than 0.5), or within a constant multiplicative factor, is NP-hard, so general efficient implementations will not be available.