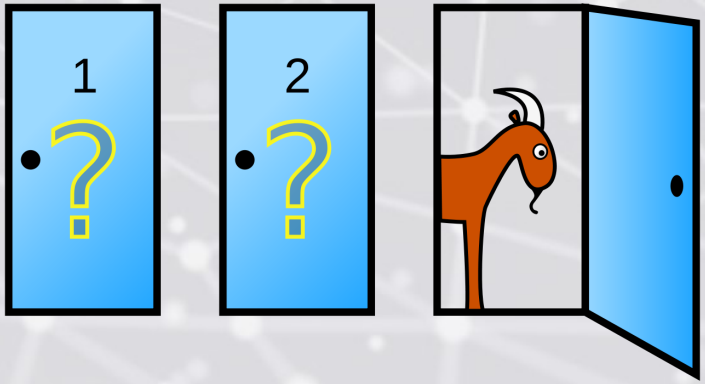
Artificial Intelligence
Chapter 13: Quantifying Uncertainty

Digression: The Monty Hall Problem

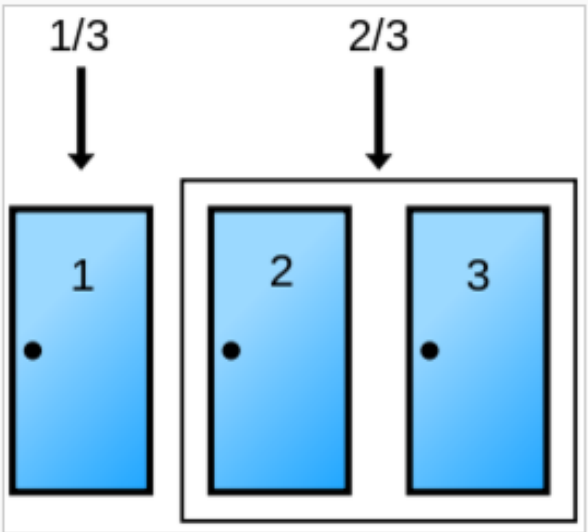
- Suppose you're on a game show, and you're given the choice of three doors:

Behind one door is a car; behind the others, goats.

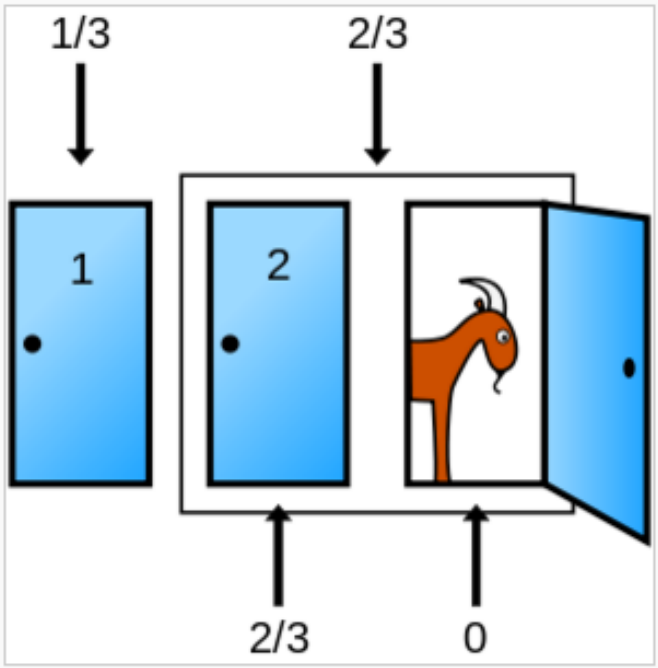
You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



Digression: The Monty Hall Problem

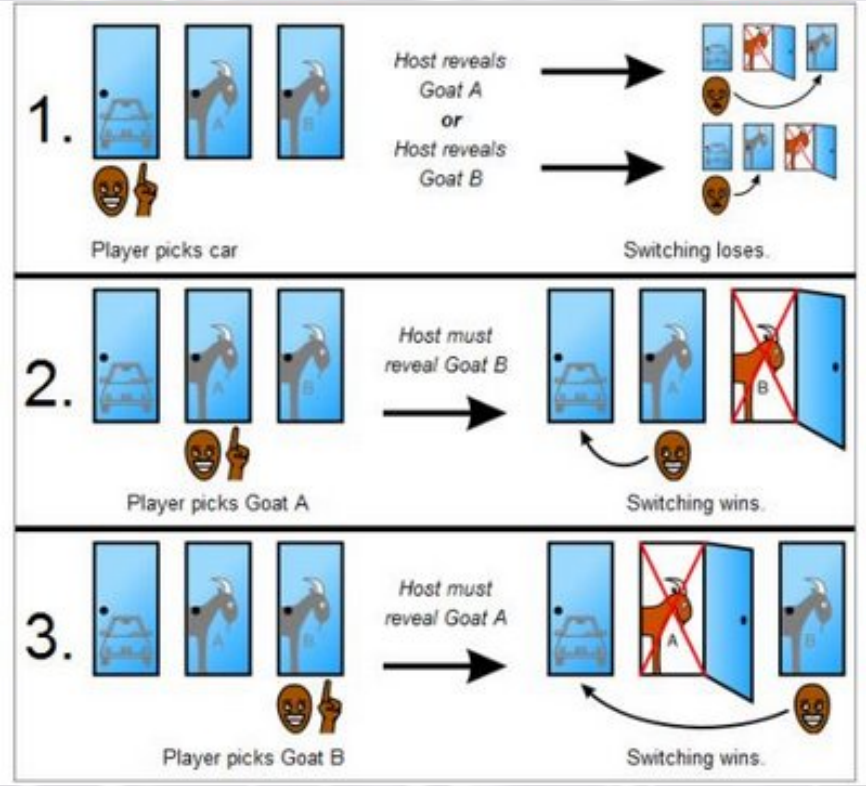


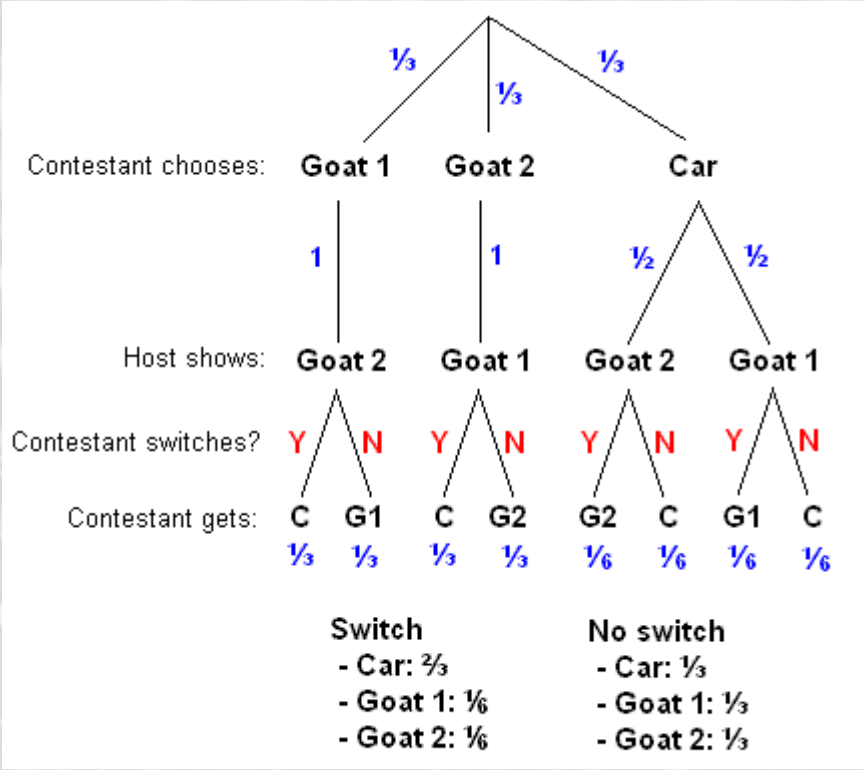
Car has a 1/3 chance of being behind the player's pick and a 2/3 chance of being behind one of the other two doors.



The host opens a door, the odds for the two sets don't change but the odds move to 0 for the open door and 2/3 closed door.

Digression: The Monty Hall Problem





Uncertainty

- Agents need to handle uncertainty, whether due to partial observability, non-determinism, or a combination of the two.

- In Chapter 4, we encountered problem-solving agents designed to handle uncertainty by monitoring a **belief-state** – a representation of the set of all possible world states in which the agent might find itself (e.g. AND-OR graphs).

- The agent generated a **contingency plan** that handles every possible eventuality that its sensors report during execution.

Uncertainty

- Despite its many virtues, however, this approach has many **significant drawbacks**:

(*) With partial information, an agent must consider *every* possible eventuality, no matter how unlikely. This leads to impossibly large and complex belief-state representations.

(*) A correct contingency plan that handles every possible outcome can grow arbitrarily large and must consider arbitrarily unlikely contingencies.

(*) Sometimes there is, in fact, no plan that is guaranteed to achieve a stated goal – yet the agent must act. It must have some way to compare the merits of plans that are not guaranteed.

Uncertainty

Let action A_t = leave for airport $_t$ minutes before flight

Will A_t get me there on time?

1. Problems:

    1. Partial observability (road state, other drivers' plans, etc.)
    2. Noisy sensors (traffic reports)

    Uncertainty in action outcomes (flat tire, etc.)

    Immense complexity of modeling and predicting traffic

4. Hence a purely logical approach either

    risks falsehood: "A_{25} will get me there on time", or

    leads to conclusions that are too weak for decision making:

» "A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

» (A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport …)

Uncertainty

- Consider a trivial example of uncertain reasoning for medical diagnosis.

(*) Toothache => Cavity (this is faulty)

Me amend it:

(*) Tootache => Cavity V Gum Problem V Abscess…

Problem is that we would need to add an almost unlimited list of possible symptoms.

- We could instead attempt to turn the rule into a *causal rule*.

(*) Cavity => Tootache (this is also incorrect; not all cavities cause pain).

Uncertainty

- The only way to fix the rule, it seems, is to make it logically exhaustive! (i.e. augment the left-hand side with all the qualifications required for a cavity to cause a toothache).

- This approach though naturally fails for at least (3) reasons:

  (1) **Laziness:** far too much work is required to compile the entire list.

  (2) **Theoretical Ignorance:** Medical science is theoretically incomplete.

  (3) **Practical Ignorance:** Even if we knew all the rules, we might be uncertain about a particular patient, because not all of the necessary tests have been run.

Uncertainty

- Typically, an agent's knowledge can at best provide only a **degree of belief**.
- Our main tool for dealing with degrees of belief is **probability theory**.
- Probability provides a way of summarizing the uncertainty that comes from our *laziness* and *ignorance*.

Uncertainty and Rational Decisions

- So how best can an agent make rational decisions in the face of uncertainty?

- To make choices, the agent must first have **preferences** between possible **outcomes** of the various plans.

- An outcome is a completely specified state, including such factors as whether the agent arrives on time (e.g. the "airport problem").

- We use **utility theory** to represent reason with preferences. Utility theory asserts that every state has a degree of *usefulness*, or utility, to an agent and that the agent will prefer states with higher utility.

Uncertainty and Rational Decisions

- Preferences, as expressed by utilities, are combined with probabilities in the general theory of rational decisions called **decision theory**:

  Decision Theory = Probability Theory + Utility Theory

- Fundamental idea: *an agent is **rational** iff it chooses the action that yields the highest expected utility, averaged over all possible outcomes of the action*. (The principle of maximum expected utility (**MEU**)).

- Note that this is none other than a computation of **expected value**.

Probability

Probabilistic assertions summarize effects of

- laziness: failure to enumerate exceptions, qualifications, etc.
- ignorance: lack of relevant facts, initial conditions, etc.

- Subjective probability:

Probabilities relate propositions to agent's own state of knowledge

e.g., $P(A_{25} \mid \text{no reported accidents}) = 0.06$

These are not assertions about the world

» Probabilities of propositions change with new evidence:

» e.g., $P(A_{25} \mid \text{no reported accidents, 5 a.m.}) = 0.15$

Making decisions under uncertainty

Suppose I believe the following:

$P(A_{25}$ gets me there on time | …) = 0.04
$P(A_{90}$ gets me there on time | …) = 0.70
$P(A_{120}$ gets me there on time | …) = 0.95
– $P(A_{1440}$ gets me there on time | …) = 0.9999

- Which action to choose?

» Depends on my preferences for missing flight vs. time spent waiting, etc.

– Utility theory is used to represent and infer preferences
– Decision theory = probability theory + utility theory

Syntax

- Basic element: random variable
- Similar to propositional logic: possible worlds defined by assignment of values to random variables.

Boolean random variables
– e.g., *Cavity* (do I have a cavity?)

- Discrete random variables
  e.g., *Weather* is one of <*sunny,rainy,cloudy,snow*>
- Domain values must be exhaustive and mutually exclusive

» Elementary proposition constructed by assignment of a value to a
» random variable: e.g., *Weather* = *sunny*, *Cavity* = *false*
» (abbreviated as ¬*cavity*)

» Complex propositions formed from elementary propositions and standard logical connectives e.g., *Weather* = *sunny* ∨ *Cavity* = *false*

Syntax

- Atomic event: A complete specification of the state of the world about which the agent is uncertain

  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  Cavity = *false* ∧ *Toothache* = *false*
  Cavity = *false* ∧ *Toothache* = *true*
  Cavity = *true* ∧ *Toothache* = *false*
  • *Cavity* = *true* ∧ *Toothache* = *true*

» Atomic events are mutually exclusive and exhaustive

Axioms of probability

- The set of all possible "worlds" is the **sample space (omega)**. The possible worlds are **mutually exclusive** and **exhaustive**.
- A fully specified probability model associates a numerical probability *P(ω)* with each possible world (we assume discrete, countable worlds).

$$0 \le P(\omega) \le 1 \ \textit{for every}\ \omega, \ \textit{and}\ \sum_{\omega \in \Omega} P(w) = 1$$

Axioms of probability

- Probabilistic assertions are usually about **sets** instead of particular possible worlds.
- These sets are commonly referred to as **events**.

- In AI, the sets are described by propositions in a formal language. The probability associated with a proposition is defined to be the sum of probabilities of the worlds in which it holds:

$$For\ any\ \ proposition\ \phi,\ \ P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Axioms of probability

- Probabilistic assertions are usually about **sets** instead of particular possible worlds.
- These sets are commonly referred to as **events**.

- In AI, the sets are described by propositions in a formal language. The probability associated with a proposition is defined to be the sum of probabilities of the worlds in which it holds:

$$For\ any\ \ proposition\ \phi,\ \ P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Axioms of probability

- The basic axioms of probability imply certain relationships among the degrees of belief that can be accorded to logically-related propositions. Example:

$$
\begin{aligned}
P(\neg a) &= \sum_{\omega \in \neg a} P(\omega) \\
&= \sum_{\omega \in \neg a} P(\omega) + \sum_{\omega \in a} P(\omega) - \sum_{\omega \in a} P(\omega) \\
&= \sum_{\omega \in \Omega} P(\omega) - \sum_{\omega \in a} P(\omega) \\
&= 1 - P(a)
\end{aligned}
$$

Axioms of probability

- **Inclusion-Exclusion** (probability of a *disjunction*):

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

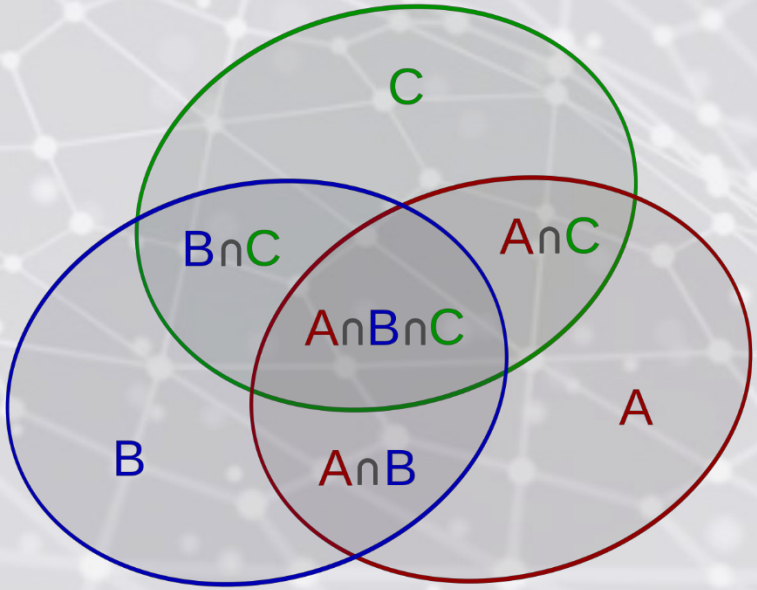
- Now derive the general formula for three or more sets…

Axioms of probability

- **Inclusion-Exclusion** (probability of a *disjunction*):

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

- Now derive the general formula for three or more sets…



Prior probability

- Prior or unconditional probabilities of propositions

  – e.g., P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72 correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

P(*Weather*) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)

Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables

P(*Weather,Cavity*) = a 4 × 2 matrix of values:

» *Weather* =	sunny	rainy	cloudy	snow
» *Cavity* = true	0.144	0.02	0.016	0.02
» *Cavity* = false	0.576	0.08	0.064	0.08

» Every question about a domain can be answered by the joint distribution

Conditional probability

- Conditional or posterior probabilities
  e.g., P(*cavity* | *toothache*) = 0.8

  i.e., given that *toothache* is all I know

(Notation for conditional distributions:
  – **P**(*Cavity* | *Toothache*) = 2-element vector of 2-element vectors)

- If we know more, e.g., *cavity* is also given, then we have
  P(*cavity* | *toothache,cavity*) = 1

» New evidence may be irrelevant, allowing simplification, e.g.,
  – P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8
» This kind of inference, sanctioned by domain knowledge, is crucial

Conditional probability

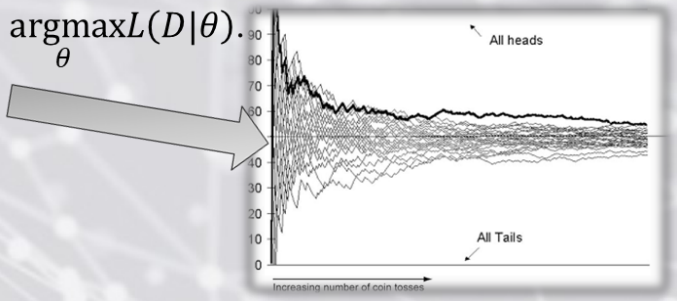
- Definition of conditional probability:
- $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$

- Product rule gives an alternative formulation:
- $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

- A general version holds for whole distributions, e.g.,
- $\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\, \mathbf{P}(Cavity)$
- (View as a set of 4×2 equations, not matrix mult.)
-
- Chain rule is derived by successive application of product rule:

Bayesian and Frequentist Probability

- (2) General paradigms for statistics and statistical inference: *frequentist* vs. *Bayesian*.
- Frequentists: Parameters are fixed; there is a (Platonic) model; parameters remain constant.
- Bayesians: Data are fixed; data are observed from realized sample; we encode prior beliefs; parameters are described probabilistically.
- Frequentists commonly use the *MLE* (**maximum likelihood estimate**) as a cogent *point estimate* of the model parameters of a probability distribution: $\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} L(D|\theta).$
- Using the *Law of Large Numbers (LLN)*, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0,$ one can consequently show that: $\hat{\theta}_{MLE} \xrightarrow{P} \theta.$



Potential issues with frequentist approach: philosophical reliance on long-term 'frequencies', *the problem of induction* (Hume) and the black swan paradox, as well as the presence of limited exact solutions for a small class of settings.

Bayesian and Frequentist Probability

In the Bayesian framework, conversely, probability is regarded as a measure of uncertainty pertaining to the practitioner's knowledge about a particular phenomenon.
The prior belief of the experimenter is not ignored but rather encoded in the process of calculating probability.
As the Bayesian gathers new information from experiments, this information is used, in conjunction with prior beliefs, to update the measure of certainty related to a specific outcome.
These ideas are summarized elegantly in the familiar *Bayes' Theorem*:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Where *H* here connotes '*hypothesis*' and *D* connotes '*data*'; the leftmost probability is referred to as the *posterior* (of the hypothesis), and the numerator factors are called the *likelihood* (of the data) and the *prior* (on the hypothesis), respectively; the denominator expression is referred to as the *marginal likelihood*.

Typically, the point estimate for a parameter used in Bayesian statistics is the *mode* of the *posterior distribution*, known as the **maximum a posterior** (MAP) estimate, which is given as:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)P(\theta)$$

Practice Problems

(1) Derive **Inclusion-Exclusion** from Equations (13.1) and (13.2) in the text.

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

$$0 \le P(\omega) \le 1 \text{ for every } \omega, \text{ and } \sum_{\omega \in \Omega} P(w) = 1 \quad (13.1)$$

$$\text{For any proposition } \phi, \; P(\phi) = \sum_{\omega \in \phi} P(\omega) \quad (13.2)$$

(2) Consider the set of all possible five-card poker hands dealt fairly (i.e. randomly) from a single, standard deck.

(i) How many atomic events are there in the joint probability distribution?

(ii) What is the probability of each atomic event?

(iii) What is the probability of being dealt a royal flush?

(iv) Four of a kind?

(v) Given that my first two cards are aces, what is the probability that my total hand consists of four aces?

Inference by enumeration

- Start with the joint probability 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:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

Inference by enumeration

- Start with the joint probability 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:

$$P(\phi) = \sum_{\omega \in \phi} P(\omega)$$

» P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

- Start with the joint probability distribution:

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

Can also compute conditional probabilities:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

»

$$= 0.4$$

»

Normalization

	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

P(*Cavity* | *toothache*) = α, **P**(*Cavity*,*toothache*)
= α, [**P**(*Cavity*,*toothache*,*catch*) + **P**(*Cavity*,*toothache*,¬ *catch*)]
= α, [<0.108,0.016> + <0.012,0.064>]
= α, <0.12,0.08> = <0.6,0.4>

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

Inference by enumeration

Typically, we are interested in
the posterior joint distribution of the query variables **Y**
given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** - **E**

- Then the required summation of joint entries is done by summing out the hidden variables:

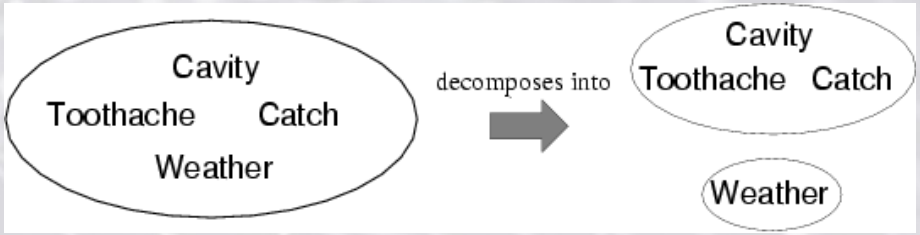
  – $\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}) = \alpha \Sigma_h \mathbf{P}(\mathbf{Y}, \mathbf{E} = \mathbf{e}, \mathbf{H} = \mathbf{h})$

2. The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables

» Obvious problems:
  – Worst-case time complexity $O(d^n)$ where d is the largest arity
  – Space complexity $O(d^n)$ to store the joint distribution
  – How to find the numbers for $O(d^n)$ entries?

Independence

- *A* and *B* are independent iff
  $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$



$\mathbf{P}(Toothache, Catch, Cavity, Weather)$
$= \mathbf{P}(Toothache, Catch, Cavity)\,\mathbf{P}(Weather)$

- 32 entries reduced to 12; for *n* independent biased coins, $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare

» Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  (1) **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)
- The same independence holds if I haven't got a cavity:

  (2) **P**(*catch* | *toothache*,¬ *cavity*) = **P**(*catch* | ¬ *cavity*)

Catch is conditionally independent of *Toothache* given *Cavity*:
- **P**(*Catch* | *Toothache,Cavity*) = **P**(*Catch* | *Cavity*)

Equivalent statements:
- **P**(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)
- **P**(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

Conditional independence

- Write out full joint distribution using chain rule:

$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&\quad = \mathbf{P}(Toothache \mid Catch, Cavity)\ \mathbf{P}(Catch, Cavity) \\
&\quad = \mathbf{P}(Toothache \mid Catch, Cavity)\ \mathbf{P}(Catch \mid Cavity)\ \mathbf{P}(Cavity) \\
&\quad = \mathbf{P}(Toothache \mid Cavity)\ \mathbf{P}(Catch \mid Cavity)\ \mathbf{P}(\mathrm{Cavity})
\end{aligned}$$

- I.e., 2 + 2 + 1 = 5 independent numbers

» In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*.

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

Practice Problems II

(1) Show that the (3) forms of "absolute" independence are equivalent.

$$\mathbf{P}(A \mid B) = \mathbf{P}(A) \quad \text{or } \mathbf{P}(B \mid A) = \mathbf{P}(B) \quad \text{or } \mathbf{P}(\mathrm{A}, \mathrm{B}) = \mathbf{P}(A)\,\mathbf{P}(B)$$

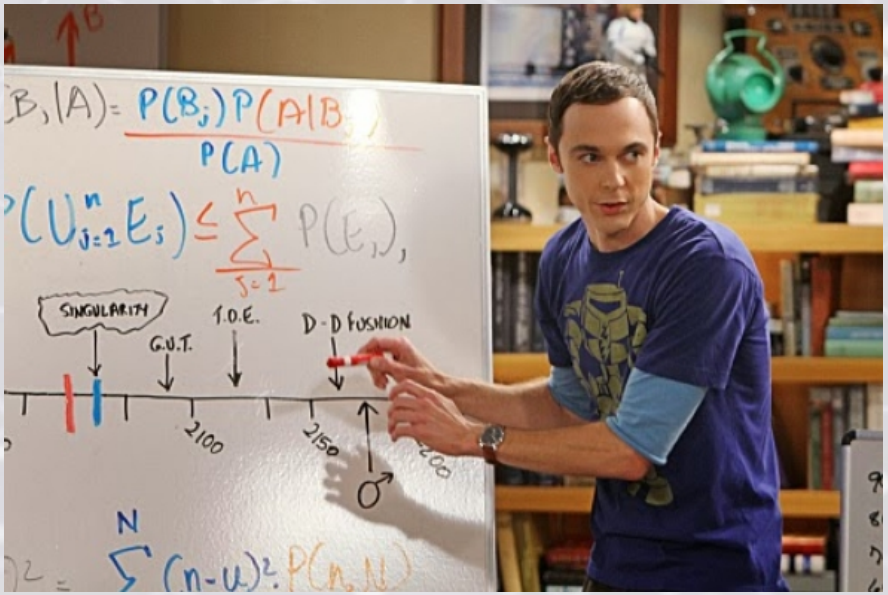
(2) Suppose that X, Y are independent random variables; let Z be a function of X and Y. Must X and Y be conditionally independent, given Z? Explain.

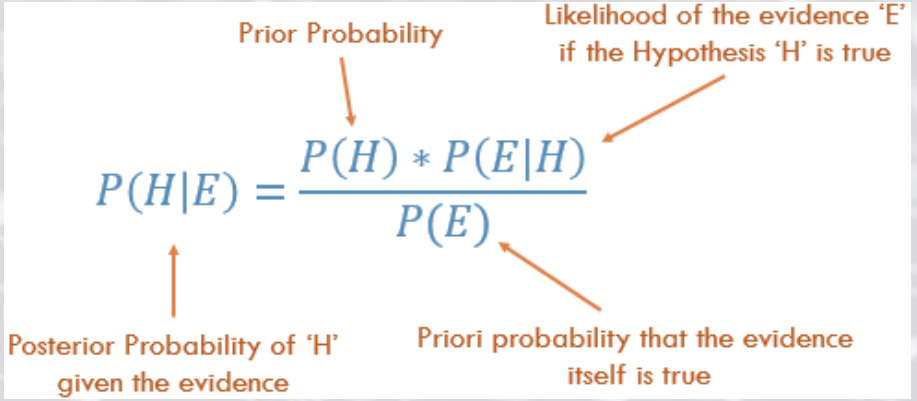
Bayes' Rule

- Product rule $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

  $\Rightarrow$ Bayes' rule: $P(a \mid b) = P(b \mid a)\, P(a) \,/\, P(b)$

» Derive Bayes' Rule…





Bayes' Rule

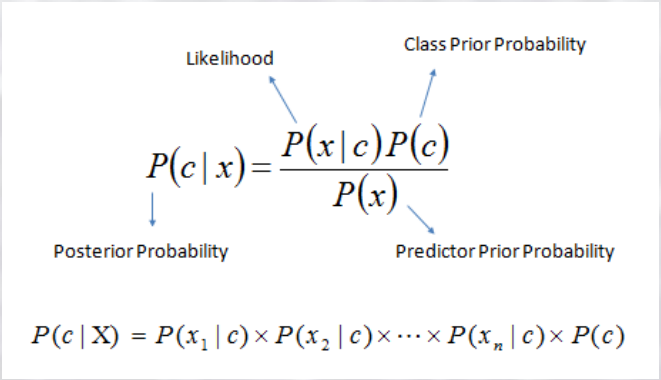
- In distribution form:

$$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\,\mathbf{P}(Y) / \mathbf{P}(X) = \alpha\mathbf{P}(X|Y)\,\mathbf{P}(Y)$$

- Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let *M* be meningitis, *S* be stiff neck:
  - P(m|s) = P(s|m) P(m) / P(s) = 0.8 × 0.0001 / 0.1 = 0.0008
  - Note: posterior probability of meningitis still very small!

Bayes' Rule and conditional independence

$\mathbf{P}(Cavity \mid toothache \wedge catch)$
$= \alpha\mathbf{P}(toothache \wedge catch \mid Cavity)\ \mathbf{P}(Cavity)$
$= \alpha\mathbf{P}(toothache \mid Cavity)\ \mathbf{P}(catch \mid Cavity)\ \mathbf{P}(Cavity)$



- This is an example of a naïve Bayes model:
  - $\mathbf{P}(\text{Cause},\text{Effect}_1, \ldots, \text{Effect}_n) = \mathbf{P}(\text{Cause})\ \pi_i \mathbf{P}(\text{Effect}_i \mid \text{Cause})$

» Total number of parameters is linear in n.

Terminology

- ***Prior probability* of *h*:**
  - $P(h)$: Probability that hypothesis h is true given our prior knowledge
  - If no prior knowledge, all $h \in H$ are equally probable
- ***Posterior probability* of *h*:**
  - $P(h \mid D)$: Probability that hypothesis h is true, given the data D.
- ***Likelihood* of *D*:**
  - $P(D \mid h)$: Probability that we will see data D, given hypothesis h is true.
- ***Marginal likelihood of D***
  -

$$P(D) = \sum_h P(D \mid h) P(h)$$

A Bayesian Approach to the Monty Hall Problem

You are a contestant on a game show.

There are 3 doors, A, B, and C. There is a new car behind one of them and goats behind the other two.

Monty Hall, the host, knows what is behind the doors. He asks you to pick a door, any door. You pick door A.

Monty tells you he will open a door, different from A, that has a goat behind it. He opens door B: behind it there is a goat.

Monty now gives you a choice: Stick with your original choice A or switch to C.

Bayesian probability formulation

Hypothesis space *H*:

h_1 = Car is behind door A

h_2 = Car is behind door B

h_3 = Car is behind door C

Data *D*: After you picked door A, Monty opened B to show a goat

What is $P(h_1 \mid D)$?

What is $P(h_2 \mid D)$?

What is $P(h_3 \mid D)$?

Prior probability:

$P(h_1) = 1/3$ $P(h_2) = 1/3$ $P(h_3) = 1/3$

Likelihood:

$P(D \mid h_1) = 1/2$

$P(D \mid h_2) = 0$

$P(D \mid h_3) = 1$

Marginal likelihood:

$$P(D) = p(D|h_1)p(h_1) + p(D|h_2)p(h_2) + p(D|h_3)p(h_3) = 1/6 + 0 + 1/3 = 1/2$$

By Bayes rule:

$$P(h_1 \mid D) = \frac{P(D \mid h_1)P(h_1)}{P(D)} = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)(2) = \frac{1}{3}$$

$$P(h_2 \mid D) = \frac{P(D \mid h_2)P(h_2)}{P(D)} = (0)\left(\frac{1}{3}\right)(2) = 0$$

$$P(h_3 \mid D) = \frac{P(D \mid h_3)P(h_3)}{P(D)} = (1)\left(\frac{1}{3}\right)(2) = \frac{2}{3}$$

So you should switch!

MAP ("maximum a posteriori") Learning

Bayes rule: $P(h \mid D) = \dfrac{P(D \mid h)P(h)}{P(D)}$

Goal of learning: Find maximum a posteriori hypothesis h_{MAP}:

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\, P(h \mid D)$$

$$= \underset{h \in H}{\text{argmax}} \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\text{argmax}}\, P(D \mid h)P(h)$$

because $P(D)$ is a constant independent of h.

Note: If every $h \in H$ is equally probable, then

$$h_{\text{MAP}} = \underset{h \in H}{\text{argmax}}\ P(D \mid h)$$

h_{MAP} is called the "maximum likelihood hypothesis".

A Medical Example

Toby takes a test for leukemia. The test has two outcomes: positive and negative. It is known that if the patient has leukemia, the test is positive 98% of the time. If the patient does not have leukemia, the test is positive 3% of the time. It is also known that 0.008 of the population has leukemia.

Toby's test is positive.

Which is more likely: Toby has leukemia or Toby does not have leukemia?

- **Hypothesis space:**
  h_1 = T. has leukemia
  h_2 = T. does not have leukemia

- **Prior:** 0.008 of the population has leukemia. Thus
  $P(h_1) = 0.008$
  $P(h_2) = 0.992$

- **Likelihood:**
  $P(+ \mid h_1) = 0.98$, $P(- \mid h_1) = 0.02$
  $P(+ \mid h_2) = 0.03$, $P(- \mid h_2) = 0.97$

- **Posterior knowledge:**
  Blood test is + for this patient.

- In summary

$$P(h_1) = 0.008,\ P(h_2) = 0.992$$

$$P(+ \mid h_1) = 0.98,\ P(- \mid h_1) = 0.02$$

$$P(+ \mid h_2) = 0.03,\ P(- \mid h_2) = 0.97$$

- Thus:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\ P(D \mid h) P(h)$$

$$P(+ \mid leukemia) P(leukemia) = (0.98)(0.008) = 0.0078$$

$$P(+ \mid \neg leukemia) P(\neg leukemia) = (0.03)(0.992) = 0.0298$$

$$h_{MAP} = \neg leukemia$$

Naive Bayes Classifier

Let $f(\mathbf{x})$ be a target function for classification: $f(\mathbf{x}) \in \{+1, -1\}$.

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$

We want to find the most probable class value, h_{MAP}, given the data $\mathbf{x}$:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class | D)$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class | x_1, x_2, ..., x_n)$$

By Bayes Theorem:

$$class_{MAP} = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \frac{P(x_1, x_2, ..., x_n \mid class) P(class)}{P(x_1, x_2, ..., x_n)}$$

$$= \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(x_1, x_2, ..., x_n \mid class) P(class)$$

$P(class)$ can be estimated from the training data. How?

However, in general, not practical to use training data to estimate $P(x_1, x_2, ..., x_n \mid class)$. Why not?

- Naive Bayes classifier: Assume

$$P(x_1, x_2, \ldots, x_n \mid class) = P(x_1 \mid class)P(x_2 \mid class)\boxed{?}P(x_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$:

Naive Bayes classifier:

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class)\prod_i P(x_i \mid class)$$

To train: Estimate the values of these various probabilities over the training set.

Training data:

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Test data:

D15	Sunny	Cool	High	Strong	?

Use training data to compute a probabilistic *model*:

$$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$$
$$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$$
$$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$$

$$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$$
$$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$$
$$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$$

$$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$$
$$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$$

$$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$$
$$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$$

Use training data to compute a probabilistic *model*:

$P(Outlook = Sunny \mid Yes) = 2/9$ $P(Outlook = Sunny \mid No) = 3/5$

$P(Outlook = Overcast \mid Yes) = 4/9$ $P(Outlook = Overcast \mid No) = 0$

$P(Outlook = Rain \mid Yes) = 3/9$ $P(Outlook = Rain \mid No) = 2/5$

$P(Temperature = Hot \mid Yes) = 2/9$ $P(Temperature = Hot \mid No) = 2/5$

$P(Temperature = Mild \mid Yes) = 4/9$ $P(Temperature = Mild \mid No) = 2/5$

$P(Temperature = Cool \mid Yes) = 3/9$ $P(Temperature = Cool \mid No) = 1/5$

$P(Humidity = High \mid Yes) = 3/9$ $P(Humidity = High \mid No) = 4/5$

$P(Humidity = Normal \mid Yes) = 6/9$ $P(Humidity = Normal \mid No) = 1/5$

$P(Wind = Strong \mid Yes) = 3/9$ $P(Wind = Strong \mid No) = 3/5$

$P(Wind = Weak \mid Yes) = 6/9$ $P(Wind = Weak \mid No) = 2/5$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

Use training data to compute a probabilistic *model*:

$P(Outlook = Sunny \mid Yes) = 2/9 \quad P(Outlook = Sunny \mid No) = 3/5$

$P(Outlook = Overcast \mid Yes) = 4/9 \quad P(Outlook = Overcast \mid No) = 0$

$P(Outlook = Rain \mid Yes) = 3/9 \quad P(Outlook = Rain \mid No) = 2/5$

$P(Temperature = Hot \mid Yes) = 2/9 \quad P(Temperature = Hot \mid No) = 2/5$

$P(Temperature = Mild \mid Yes) = 4/9 \quad P(Temperature = Mild \mid No) = 2/5$

$P(Temperature = Cool \mid Yes) = 3/9 \quad P(Temperature = Cool \mid No) = 1/5$

$P(Humidity = High \mid Yes) = 3/9 \quad P(Humidity = High \mid No) = 4/5$

$P(Humidity = Normal \mid Yes) = 6/9 \quad P(Humidity = Normal \mid No) = 1/5$

$P(Wind = Strong \mid Yes) = 3/9 \quad P(Wind = Strong \mid No) = 3/5$

$P(Wind = Weak \mid Yes) = 6/9 \quad P(Wind = Weak \mid No) = 2/5$

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D15	Sunny	Cool	High	Strong	?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class) \prod_i P(x_i \mid class)$$

Estimating probabilities / Smoothing

- **Recap:** In previous example, we had a training set and a new example, (Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong)
- We asked: What classification is given by a naive Bayes classifier?
- Let n_c be the number of training instances with class c.
- Let $n_c^{x_i=a_k}$ be the number of training instances with attribute value $x_i=a_k$ and class c.

Then: $$P(x_i = a_i \mid c) = \frac{n_c^{x_i=a_k}}{n_c}$$

- **Problem with this method:** If n_c is very small, gives a poor estimate.

- E.g., $P(Outlook = Overcast \mid no) = 0$.

- Now suppose we want to classify a new instance:

(Outlook=overcast, Temperature=cool, Humidity=high, Wind=strong)

Then:

$$P(\text{no})\prod_i P(x_i \mid \text{no}) = 0$$

This incorrectly gives us zero probability due to small sample.

One solution: *Laplace smoothing* (also called "add-one" smoothing)

For each class c and attribute x_i with value a_k, add one "virtual" instance.

That is, for each class c, recalculate:

$$P(x_i = a_i \mid c) = \frac{n_c^{x_i = a_k} + 1}{n_c + K}$$

where K is the number of possible values of attribute a.

Training data:

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Laplace smoothing: Add the following virtual instances for *Outlook*:

Outlook=Sunny: **Yes** *Outlook=Overcast*: **Yes** *Outlook=Rain*: **Yes**
Outlook=Sunny: **No** *Outlook=Overcast*: **No** *Outlook=Rain*: **No**

$$P(Outlook = overcast \mid \mathbf{No}) = \frac{0}{5} \rightarrow \frac{n_c^{x_i=a_k} + 1}{n_c + K} = \frac{0+1}{5+3} = \frac{1}{8}$$

$$P(Outlook = overcast \mid \mathbf{Yes}) = \frac{4}{9} \rightarrow \frac{n_c^{x_i=a_k} + 1}{n_c + K} = \frac{4+1}{9+3} = \frac{5}{12}$$

$P(Outlook = Sunny \mid Yes) = 2/9 \rightarrow 3/12 \quad P(Outlook = Sunny \mid No) = 3/5 \rightarrow 4/8$

$P(Outlook = Overcast \mid Yes) = 4/9 \rightarrow 5/12 \quad P(Outlook = Overcast \mid No) = 0/5 \rightarrow 1/8$

$P(Outlook = Rain \mid Yes) = 3/9 \rightarrow 4/12 \quad P(Outlook = Rain \mid No) = 2/5 \rightarrow 3/8$

$P(Humidity = High \mid Yes) = 3/9 \rightarrow 4/11 \quad P(Humidity = High \mid No) = 4/5 \rightarrow 5/7$

$P(Humidity = Normal \mid Yes) = 6/9 \rightarrow 7/11 \quad P(Humidity = Normal \mid No) = 1/5 \rightarrow 2/7$

Etc.

Naive Bayes on continuous-valued attributes

- How to deal with continuous-valued attributes?

Two possible solutions:

- Discretize
- Assume particular probability distribution of classes over values (estimate parameters from training data)

Discretization: Equal-Width Binning

For each attribute x_i, create k equal-width bins in interval from $min(x_i)$ to $max(x_i)$.

The discrete "attribute values" are now the bins.

Questions: What should k be? What if some bins have very few instances?

Problem with balance between *discretization bias* and *variance*.

The more bins, the lower the bias, but the higher the variance, due to small sample size.

Discretization: Equal-Frequency Binning

For each attribute x_i, create k bins so that each bin contains an equal number of values.

Also has problems: What should k be? Hides outliers. Can group together instances that are far apart.

Gaussian Naïve Bayes

Assume that within each class, values of each numeric feature are normally distributed:

$$p(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu_{i,c}$ is the mean of feature *i* given the class c, and $\sigma_{i,c}$ is the standard deviation of feature *i* given the class c

We estimate $\mu_{i,c}$ and $\sigma_{i,c}$ from training data.

Example

x_1	x_2	**Class**
3.0	5.1	**POS**
4.1	6.3	**POS**
7.2	9.8	**POS**
2.0	1.1	NEG
4.1	2.0	NEG
8.1	9.4	NEG

Example

x_1	x_2	**Class**
3.0	5.1	**POS**
4.1	6.3	**POS**
7.2	9.8	**POS**
2.0	1.1	NEG
4.1	2.0	NEG
8.1	9.4	NEG

$P(\mathbf{POS}) = 0.5$

$P(\mathbf{NEG}) = 0.5$

$$\mu_{1,\mathbf{POS}} = \frac{(3.0 + 4.1 + 7.2)}{3} = 4.8$$

$$\sigma_{1,\mathbf{POS}} = \sqrt{\frac{(3.0 - 4.8)^2 + (4.1 - 4.8)^2 + (7.2 - 4.8)^2}{3}} = 1.8$$

$$\mu_{1,\mathbf{NEG}} = \frac{(2.0 + 4.1 + 8.1)}{3} = 4.7$$

$$\sigma_{1,\mathbf{NEG}} = \sqrt{\frac{(2.0 - 4.7)^2 + (4.1 - 4.7)^2 + (8.1 - 4.7)^2}{3}} = 2.5$$

$$\mu_{2,\mathbf{POS}} = \frac{(5.1 + 6.3 + 9.8)}{3} = 7.1$$

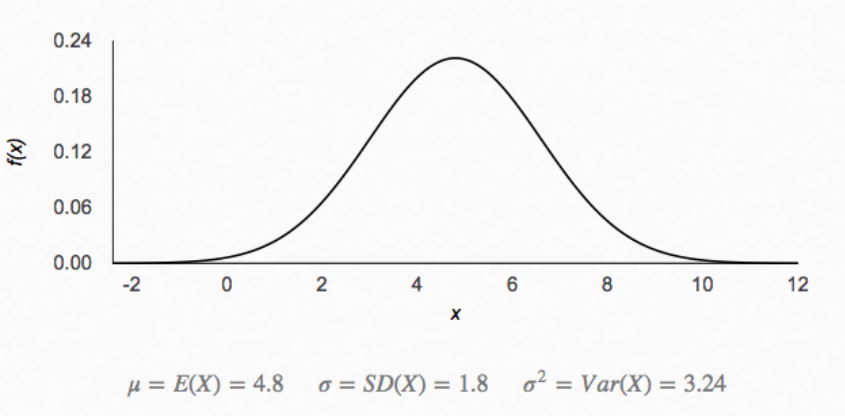
$$\sigma_{2,\mathbf{POS}} = \sqrt{\frac{(5.1 - 7.1)^2 + (6.3 - 7.1)^2 + (9.8 - 7.1)^2}{3}} = 2.0$$

$$\mu_{2,\mathbf{NEG}} = \frac{(1.1 + 2.0 + 9.4)}{3} = 4.2$$

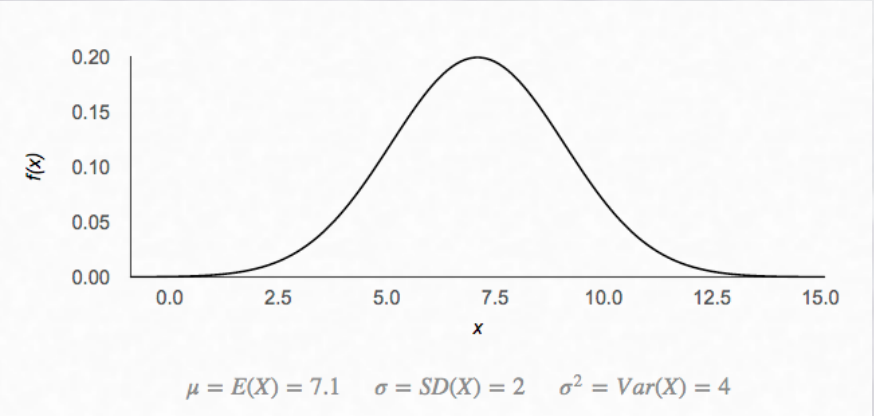
$$\sigma_{2,\mathbf{NEG}} = \sqrt{\frac{(1.1 - 4.2)^2 + (2.0 - 4.2)^2 + (9.4 - 4.2)^2}{3}} = 3.7$$

http://homepage.stat.uiowa.edu/~mbognar/applets/normal.html

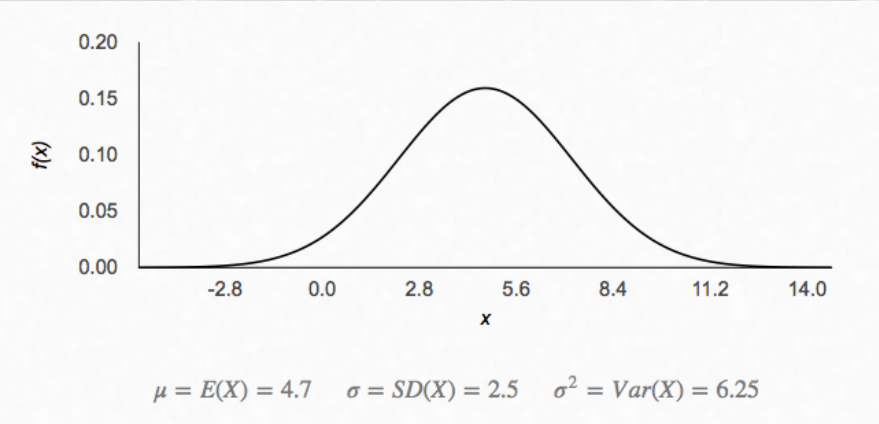
$N_{1,\mathbf{POS}} = N(x; 4.8, 1.8)$



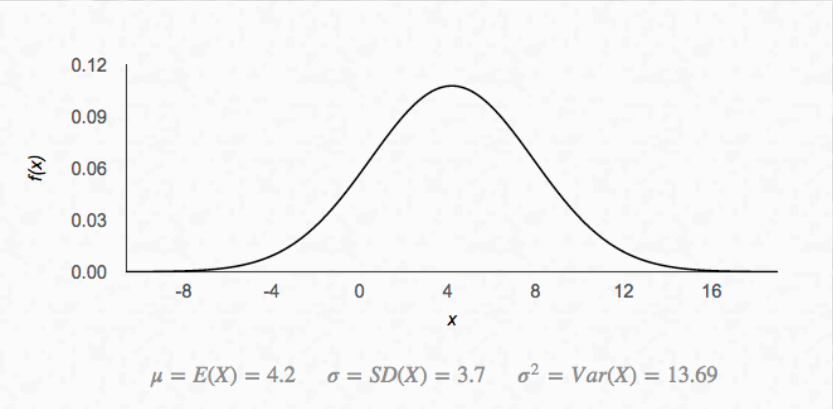
$N_{2,\mathbf{POS}} = N(x; 7.1, 2.0)$



$N_{1,\mathbf{NEG}} = N(x; 4.7, 2.5)$



$N_{2,\mathbf{NEG}} = N(x; 4.2, 3.7)$



Now, suppose you have a new example **x**, with $x_1 = 5.2, x_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

Now, suppose you have a new example **x**, with $x_1 = 5.2$, $x_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} \; P(class) \prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Note: N is the probability density function, but can be used analogously to probability in Naïve Bayes calculations.

Now, suppose you have a new example **x**, with $x_1 = 5.2, x_2 = 6.3$.

What is $class_{NB}$ (**x**) ?

$$class_{NB}(\mathbf{x}) = \underset{class \in \{+1,-1\}}{\operatorname{argmax}} P(class) \prod_i P(x_i \mid class)$$

$$P(x_i \mid c) = N(x_i; \mu_{i,c}, \sigma_{i,c})$$

where

$$N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(x_1 \mid \mathbf{POS}) = \frac{1}{\sqrt{2\pi}(1.8)} e^{-\frac{(5.2-4.8)^2}{2(1.8)^2}} = .22$$

$$P(x_2 \mid \mathbf{POS}) = \frac{1}{\sqrt{2\pi}(2.0)} e^{-\frac{(6.3-7.1)^2}{2(2.0)^2}} = .18$$

$$P(x_1 \mid \mathbf{NEG}) = \frac{1}{\sqrt{2\pi}(2.5)} e^{-\frac{(5.2-4.7)^2}{2(2.5)^2}} = .16$$

$$P(x_2 \mid \mathbf{NEG}) = \frac{1}{\sqrt{2\pi}(3.7)} e^{-\frac{(6.3-4.2)^2}{2(3.7)^2}} = .09$$

Positive:

$$P(\mathbf{POS})P(x_1 \mid \mathbf{POS})P(x_2 \mid \mathbf{POS}) = (.5)(.22)(.18) = .02$$

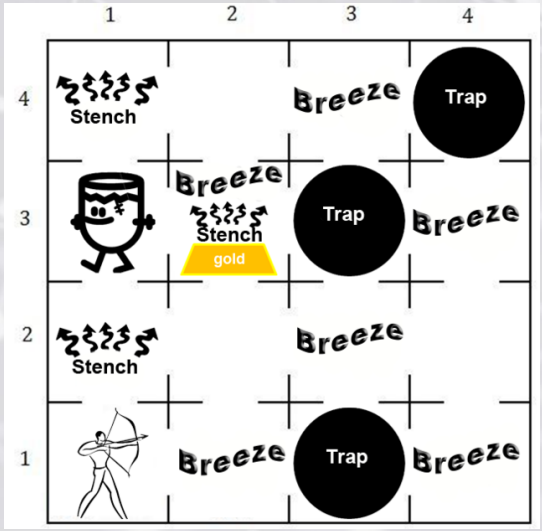
Negative:

$$P(\mathbf{NEG})P(x_1 \mid \mathbf{NEG})P(x_2 \mid \mathbf{NEG}) = (.5)(.16)(.09) = .0072$$

$$class_{NB}(\mathbf{x}) = \mathbf{POS}$$

Wumpus World

- The **Wumpus world problem** deals with an AI robot navigating its way through a 4x4 puzzle to try and find gold. The robot must safely navigate its way around bottomless pits of death and evil Wumpus creatures to locate the gold hidden on the board. After it has successfully found the gold, it must safely navigate its way back to the starting point. The robot must use its light sensors and the signals sent to it at each square to determine which way to properly navigate to reach its goal.
- As the AI robot moves across the puzzle board, it will receive a set of signals at each square it moves onto. There will be one signal given to the robot per each adjacent square around it. There are four different signals given to the robot that indicate either a wumpus, pit, empty square, or gold is nearby. The robot must make a decision as to which square is safe to move into and which one may contain the gold.



Wumpus World

$P_{ij} = true$ iff $[i, j]$ contains a pit

$B_{ij} = true$ iff $[i, j]$ is breezy

Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Wumpus World

Specifying the probability model

The full joint distribution is $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$

Apply product rule: $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$

(Do it this way to get $P(Effect|Cause)$.)

First term: 1 if pits are adjacent to breezes, 0 otherwise

Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Wumpus World

Observations and query

We know the following facts:

$$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$$
$$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$$

Query is $\mathbf{P}(P_{1,3}|known, b)$

Define $Unknown = P_{ij}$s other than $P_{1,3}$ and $Known$

For inference by enumeration, we have

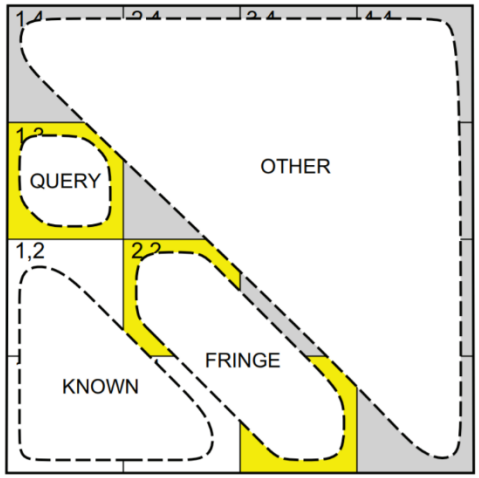
$$\mathbf{P}(P_{1,3}|known, b) = \alpha \Sigma_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

Wumpus World

Using conditional independence

Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares



Define $Unknown = Fringe \cup Other$

$\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$

Manipulate query into a form where we can use this!

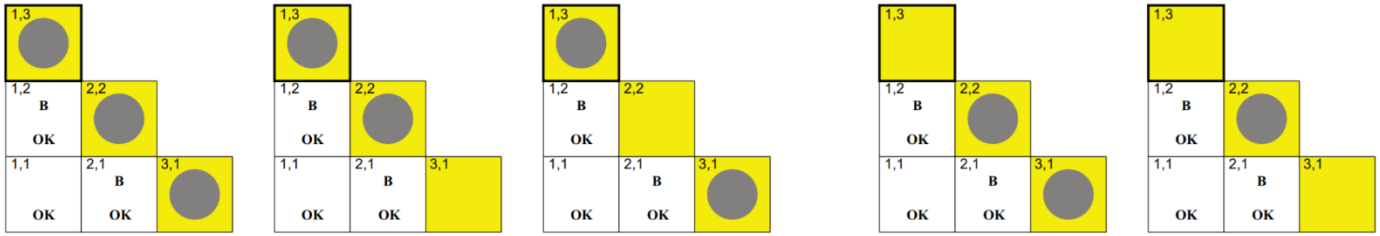
Wumpus World

Using conditional independence contd.

$$
\begin{aligned}
\mathbf{P}(P_{1,3}|known, b) &= \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\
&= \alpha\, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other) \\
&= \alpha' \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)
\end{aligned}
$$

Wumpus World

Using conditional independence contd.



$$\mathbf{P}(P_{1,3}|known, b) = \alpha' \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16) \rangle$$
$$\approx \langle 0.31, 0.69 \rangle$$

$$\mathbf{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14 \rangle$$

Bayes Nets

Another example

A patient comes into a doctor's office with a bad cough and a high fever.

Hypothesis space *H*:

h_1: patient has flu

h_2: patient does not have flu

Data *D*:

coughing = true, *fever* = true

Prior probabilities:

$p(h_1) = .1$

$p(h_2) = .9$

Likelihoods

$p(\boldsymbol{D} \mid h1) = .8$

$p(\boldsymbol{D} \mid h2) = .4$

Prob. of data

$P(\boldsymbol{D}) =$

Posterior probabilities:

$P(h_1|\boldsymbol{D}) =$

$P(h_2|\boldsymbol{D}) =$

- Let’s say we have the following random variables:

cough
fever
flu
smokes

Full joint probability distribution

smokes				
	cough		¬ *cough*	
	Fever	¬ *Fever*	*Fever*	¬ *Fever*
flu	p_1	p_2	p_3	p_4
¬*flu*	p_5	p_6	p_7	p_8

¬ *smokes*				
	cough		¬ *cough*	
	fever	¬ *fever*	*fever*	¬ *fever*
flu	p_9	p_{10}	p_{11}	p_{12}
¬*flu*	p_{13}	p_{14}	p_{15}	p_{16}

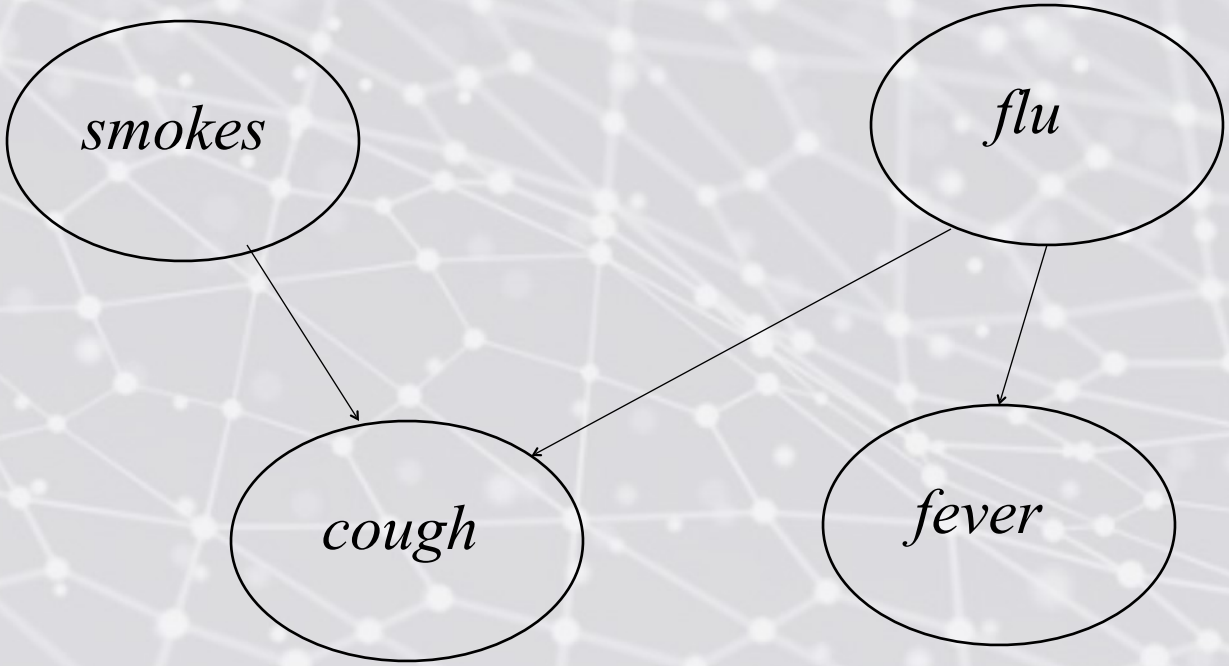
Sum of all boxes is 1.

In principle, the full joint distribution can be used to answer any question about probabilities of these combined parameters.

However, size of full joint distribution scales exponentially with number of parameters so is expensive to store and to compute with.

Bayesian networks

- Idea is to represent dependencies (or causal relations) for all the variables so that space and computation-time requirements are minimized.



“Graphical Models”

Conditional probability tables for each node

smoke	
true	0.2
false	0.8

		cough	
flu	***smoke***	true	false
True	True	0.95	0.05
True	False	0.8	0.2
False	True	0.6	0.4
false	false	0.05	0.95

flu	
true	0.01
false	0.99

	fever	
flu	true	false
true	0.9	0.1
false	0.2	0.8

Semantics of Bayesian networks

- If network is correct, can calculate full joint probability distribution from network.

$$P((X_1 = x_1) \wedge (X_2 = x_2) \ldots \wedge (X_n = x_n))$$

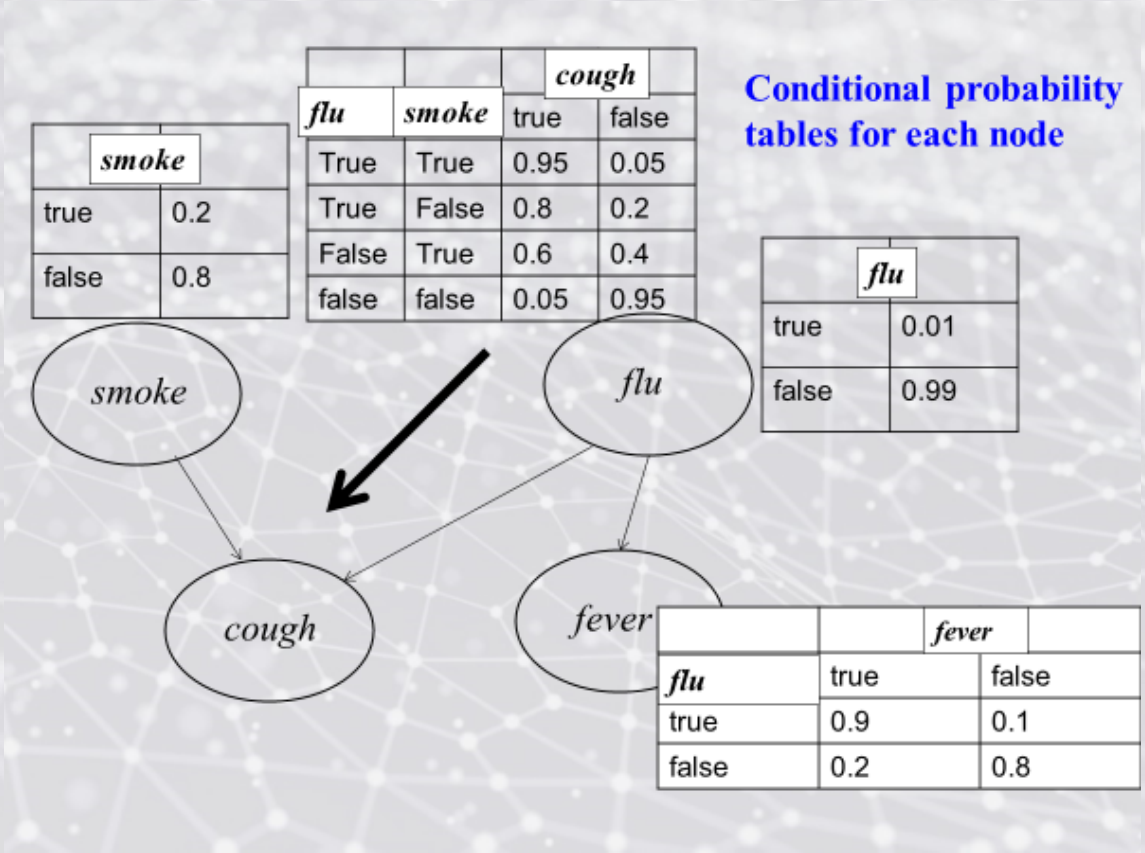
$$= \prod_{i=1}^{n} P(X_i = x_i \mid parents(X_i))$$

where parents(X_i) denotes specific values of parents of X_i.

Example

- Calculate

$$P[(cough = t) \wedge (fever = f) \wedge (flu = t) \wedge (smoke = f)]$$

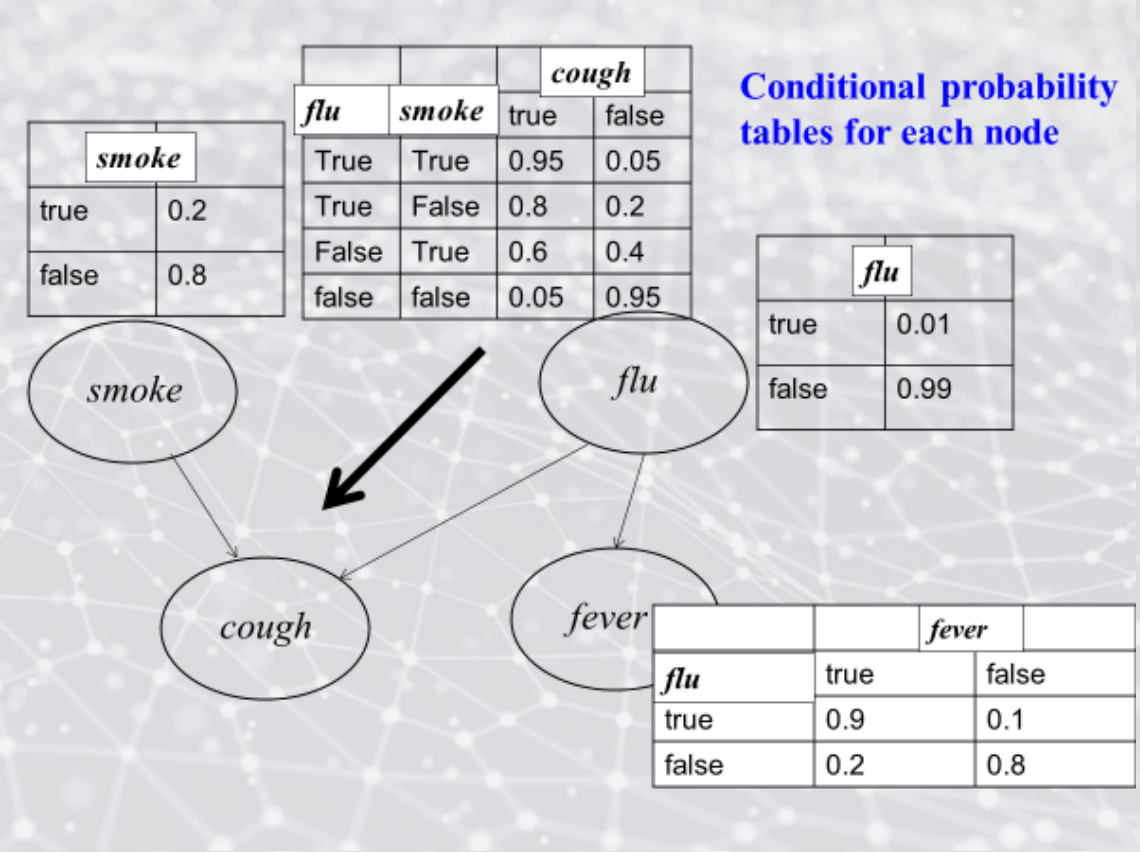


Example

- Calculate

$$P[(cough = t) \wedge (fever = f) \wedge (flu = t) \wedge (smoke = f)]$$

$$= P(flu = t)P(smoke = f)P(cough = t \mid smoke = f, flu = t)P(fever = f \mid flu = t)$$



Different types of inference in Bayesian Networks

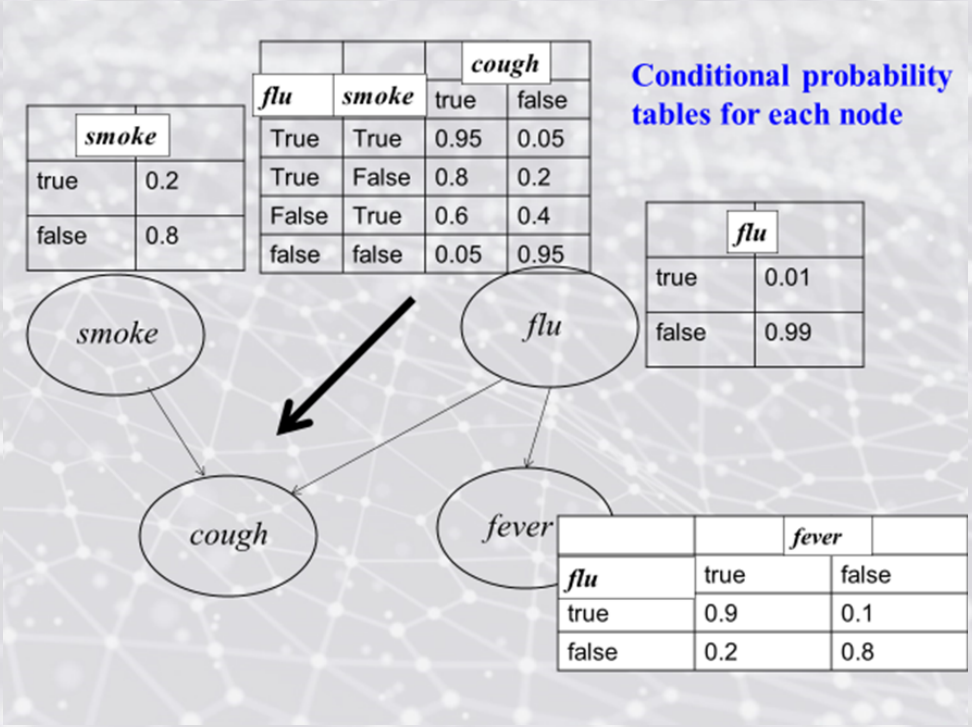
Causal inference

Evidence is cause, inference is probability of effect

Example:

Instantiate evidence *flu* = *true.* What is *P*(*fever* | *flu*)?

$$P(fever \mid flu) = .9$$

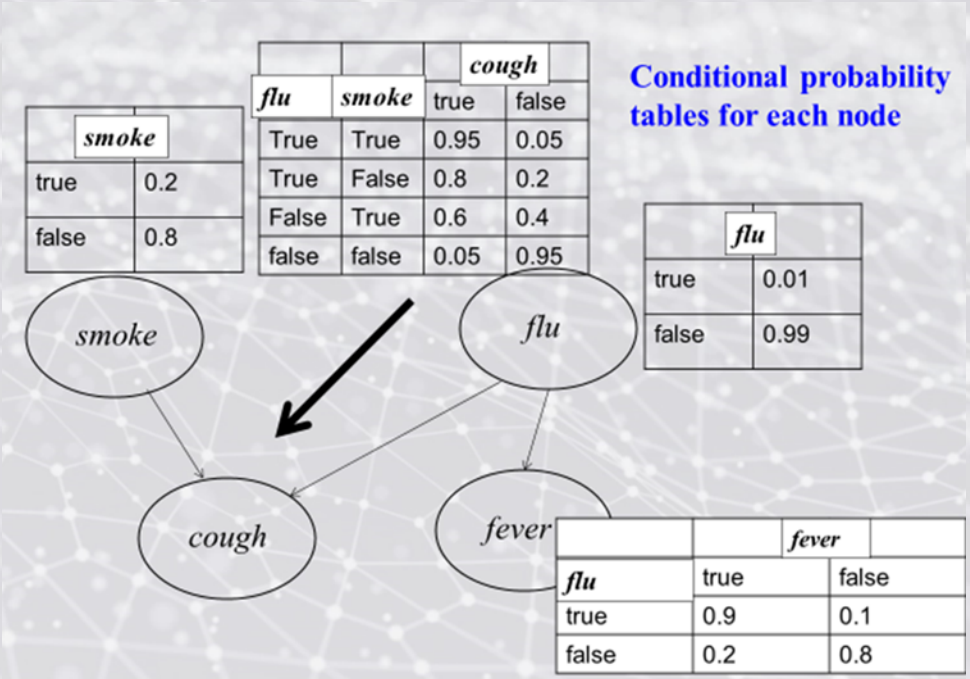


Diagnostic inference

Evidence is effect, inference is probability of cause

Example: Instantiate evidence *fever* = true. What is $P(flu \mid fever)$?

$$P(flu \mid fever) = \frac{P(fever \mid flu)P(flu)}{P(fever)} = \frac{(.9)(.01)}{.207} = .043$$

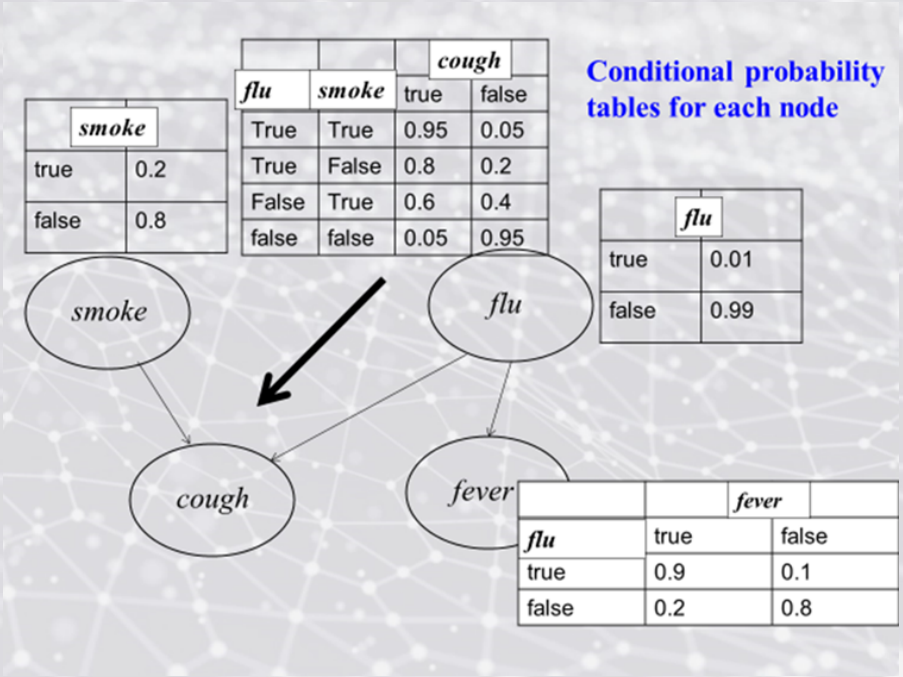


Example: What is P(*flu*|*cough*)?

$$P(flu \mid cough) = \frac{P(cough \mid flu)P(flu)}{P(cough)} =$$

$$= \frac{[P(cough \mid flu, smoke)\,p(smoke) + P(cough \mid flu, \neg smoke)\,p(\neg smoke)]P(flu)}{p(cough)}$$

$$= \frac{[(.95)(.2) + (.8)(.8)](.01)}{.167} = .0497$$



Inter-causal inference

Explain away different possible causes of effect

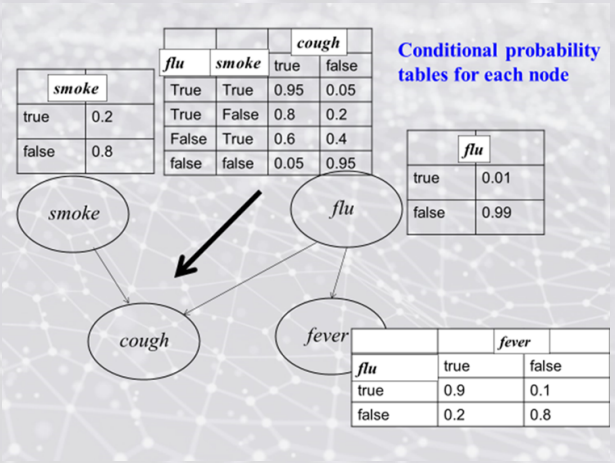
Example: What is P(*flu*|*cough*,*smoke*)?

$$P(flu \mid cough, smoke) = \frac{p(flu \wedge cough \wedge smoke)}{p(cough \wedge smoke)}$$

$$= \frac{p(cough \mid flu, smoke)\,p(flu)\,p(smoke)}{p(cough \mid flu, smoke)\,p(flu)\,p(smoke) + p(cough \mid smoke, \neg flu)\,p(smoke)\,p(\neg flu)}$$

$$= (.95)(.01)(.2)/[(.95)(.01)(.2) + (.6)(.2)(.99)]$$

$$= 0.016$$



Why is P(*flu*|*cough*,*smoke*) < P(*flu*|*cough*)?

Complexity of Bayesian Networks

For n random Boolean variables:

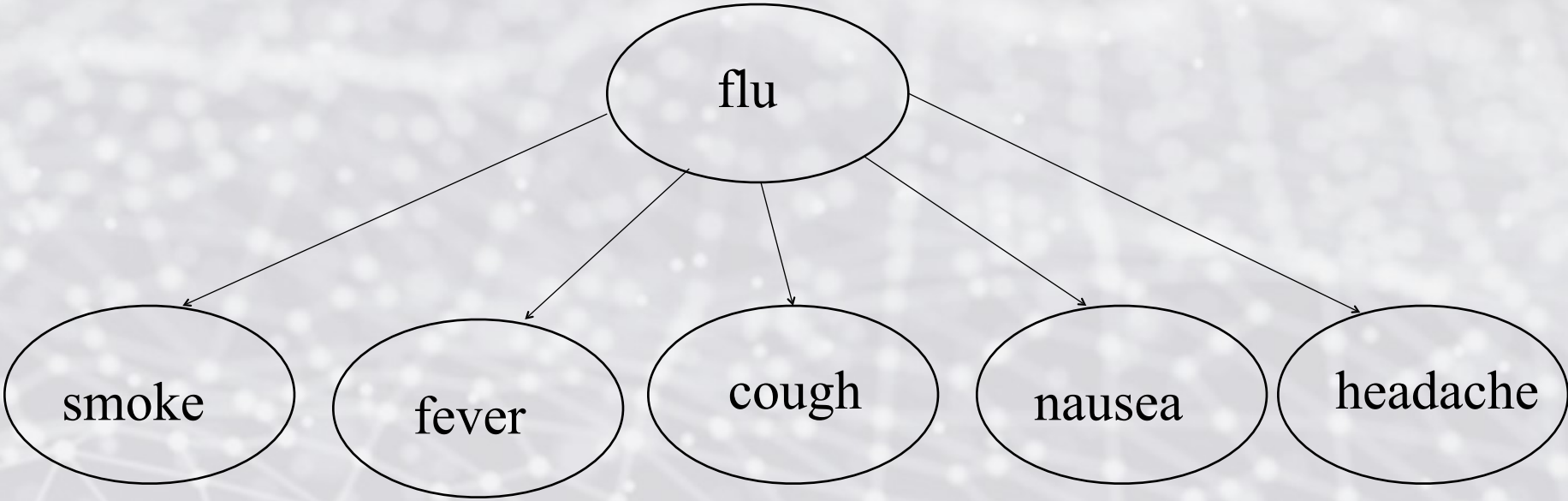
- Full joint probability distribution: 2^n entries
- Bayesian network with at most k parents per node:
  - Each conditional probability table: at most 2^k entries
  - Entire network: $n\ 2^k$ entries

What are the advantages of Bayesian networks?

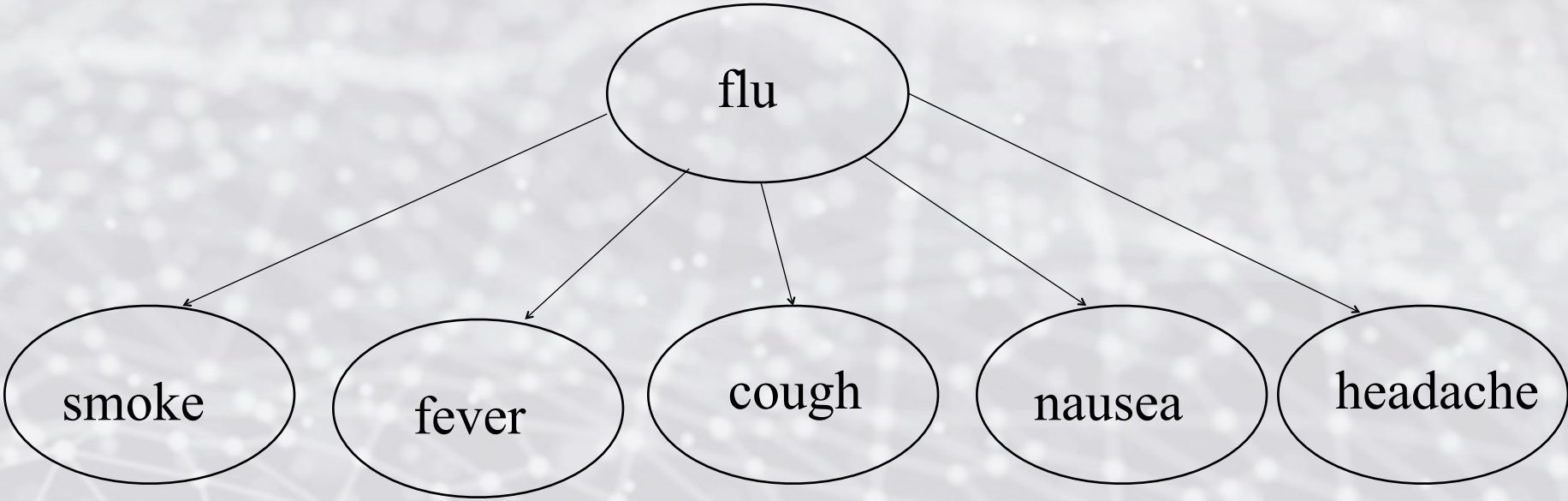
- Intuitive, concise representation of joint probability distribution (i.e., conditional dependencies) of a set of random variables.
- Represents “beliefs and knowledge” about a particular class of situations.
- Efficient (?) (approximate) inference algorithms
- Efficient, effective learning algorithms

Issues in Bayesian Networks

- Building / learning network topology
- Assigning / learning conditional probability tables
- Approximate inference via sampling



$$P(flu \wedge smoke \wedge fever \wedge cough \wedge nausea \wedge headache)$$
$$= P(flu)P(smoke \mid flu)P(fever \mid flu)P(cough \mid flu)$$
$$P(nausea \mid flu)P(headache \mid flu)$$



More generally, for classification :

$$P(C = c_j \mid X_1 = x_1, \ldots, X_1 = x_n)$$
$$= P(C = c_j) \prod_i P(X_i = x_i \mid C = c_j)$$

"Naive Bayes"

Learning network topology

- Many different approaches, including:
  - Heuristic search, with evaluation based on information theory measures
  - Genetic algorithms
  - Using “meta” Bayesian networks!

Learning conditional probabilities

- In general, random variables are not binary, but real-valued
- Conditional probability tables conditional probability distributions
- Estimate parameters of these distributions from data
- If data is missing on one or more variables, use “expectation maximization” algorithm

Approximate inference via sampling

- Recall: We can calculate full joint probability distribution from network.

$$P(X_1,...,X_d) = \prod_{i=1}^{d} P(X_i \mid parents(X_i))$$

where parents(X_i) denotes specific values of parents of X_i.

- We can do diagnostic, causal, and inter-causal inference

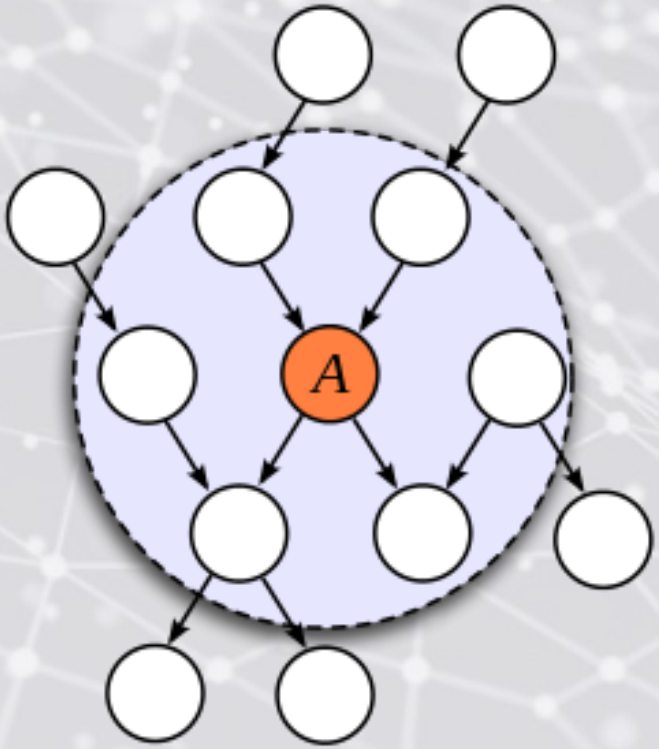
- But if there are a lot of nodes in the network, this can be very slow!

Need efficient algorithms to do approximate calculations!

Markov Chain Monte Carlo Sampling

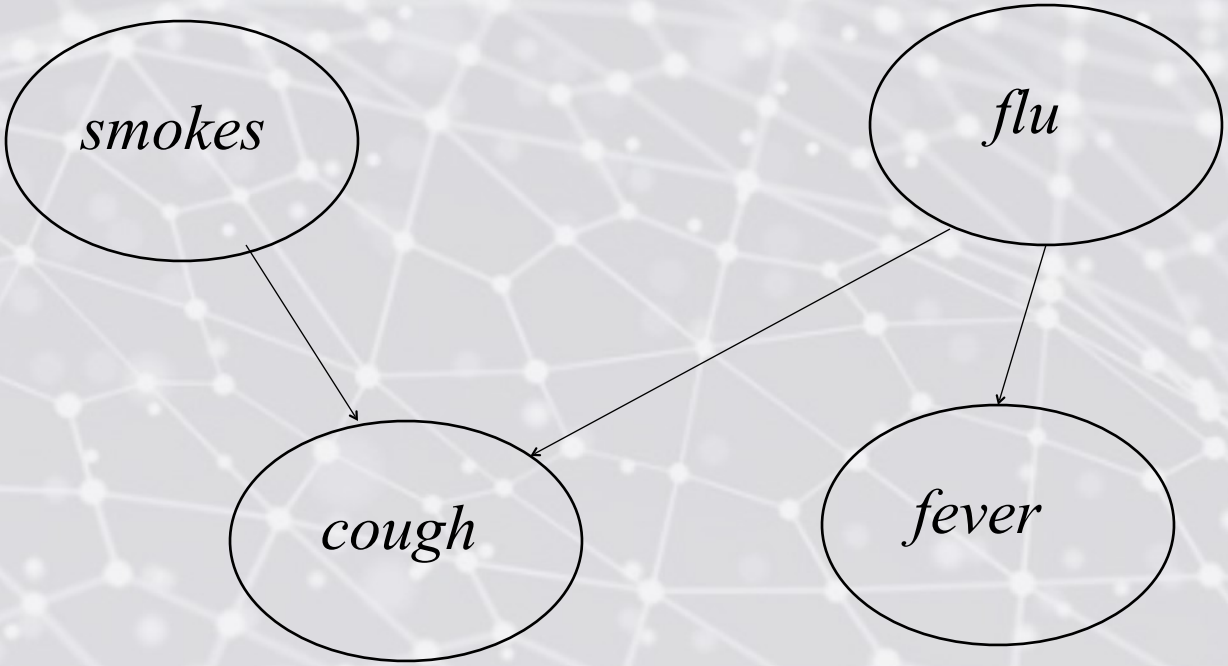
- One of most common methods used in real applications.
- By construction of Bayesian network, a node is conditionally independent of its non-descendants, given its parents.
- A node can be conditionally dependent on its children and on the other parents of its children.
- **Definition:** The ***Markov blanket*** of a variable X_i is X_i*'s* parents, children, and children's other parents.

- **Theorem:** A node X_i is conditionally independent of all other nodes in the network, given its Markov blanket.



Example

- What is the Markov blanket of *cough*? of *flu*?



Markov Chain Monte Carlo (MCMC) Sampling

- Start with random sample from variables: $(x_1, ..., x_n)$. This is the current "state" of the algorithm.

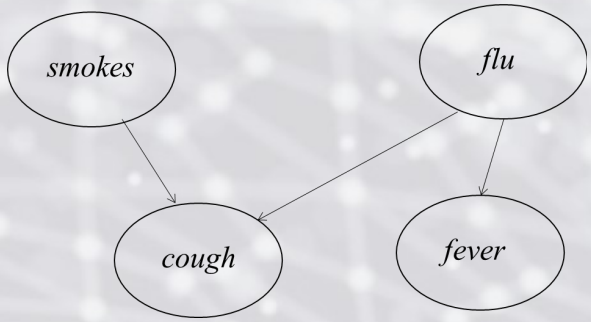
- Next state: Randomly sample value for one non-evidence variable X_i , conditioned on current values in "Markov Blanket" of X_i.

Example

- Query: What is P(*cough*| *smoke*)?

- MCMC:
  - Random sample, with evidence variables fixed:

flu	***smoke***	*fever*	*cough*
true	**true**	false	true



  - Repeat:
    1. Sample *flu* probabilistically, given current values of its Markov blanket: *smoke* = true, *fever* = false*, cough = true*

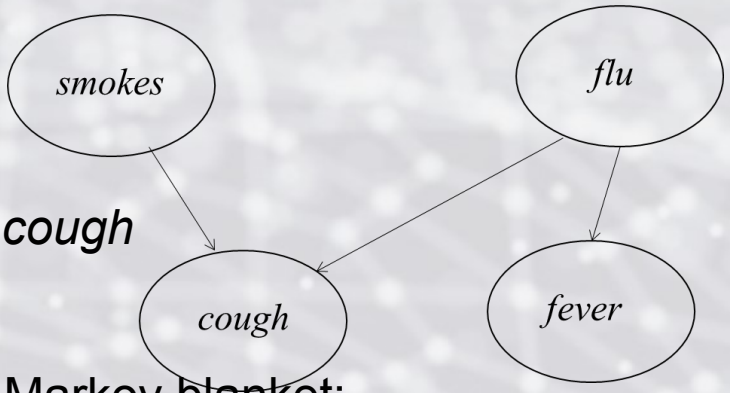
Suppose result is *false*. New state:

flu	***smoke***	*fever*	*cough*
false	**true**	false	true

2. Sample *cough*, given current values of its Markov blanket:
   smoke = true , flu = false

   Suppose result is *true*.
   New state:

flu	***smoke***	*fever*	*cough*
false	**true**	false	true



3. Sample *fever*, given current values of its Markov blanket:
   flu = false

Suppose result is *true*.
   New state:

flu	***smoke***	*fever*	*cough*
false	**true**	true	true

- Each sample contributes to estimate for query
  P(*cough* | *smoke*)

- Suppose we perform 100 such samples, 20 with *cough* = true and 80 with *cough*= false.

- Then answer to the query is
  P(cough | smoke) = .20

- **Theorem:** MCMC settles into behavior in which each state is sampled exactly according to its posterior probability, given the evidence.

Summary

- Uncertainty arises because of both laziness and ignorance. It is **inescapable** in complex, nondeterministic, or partially observable environments.

- Probability is a rigorous formalism for uncertain knowledge. Probabilities summarize the agent's believes relative to the evidence.

- **Decision Theory** combines the agent's beliefs and desires, defining the best action as the one that maximizes expected utility.

- Basic probability statements include **priors probabilities** and **conditional probabilities**. Joint probabilities distributions specifiy a probability of every atomic event.

Summary

- **Absolute independence** between subsets of random variables allows the full joint distribution to be factored into smaller joint distributions, greatly reducing its complexity. Absolute independence seldom occurs in practice.

- **Bayes' Rule** allows unknown probabilities to be computer from known conditional probabilities, usually in the casual direction.

- **Conditional independence** brought about by direct causal relationships in the domain might allow the full joint distribution to be factored into smaller, conditional distributions.

- The **naïve Bayes** model assumes the conditional independence of all effect variables, given a single cause variable, and grows linearly with the number of effects.