CSCE A405/A605 (Adv) Artificial Intelligence

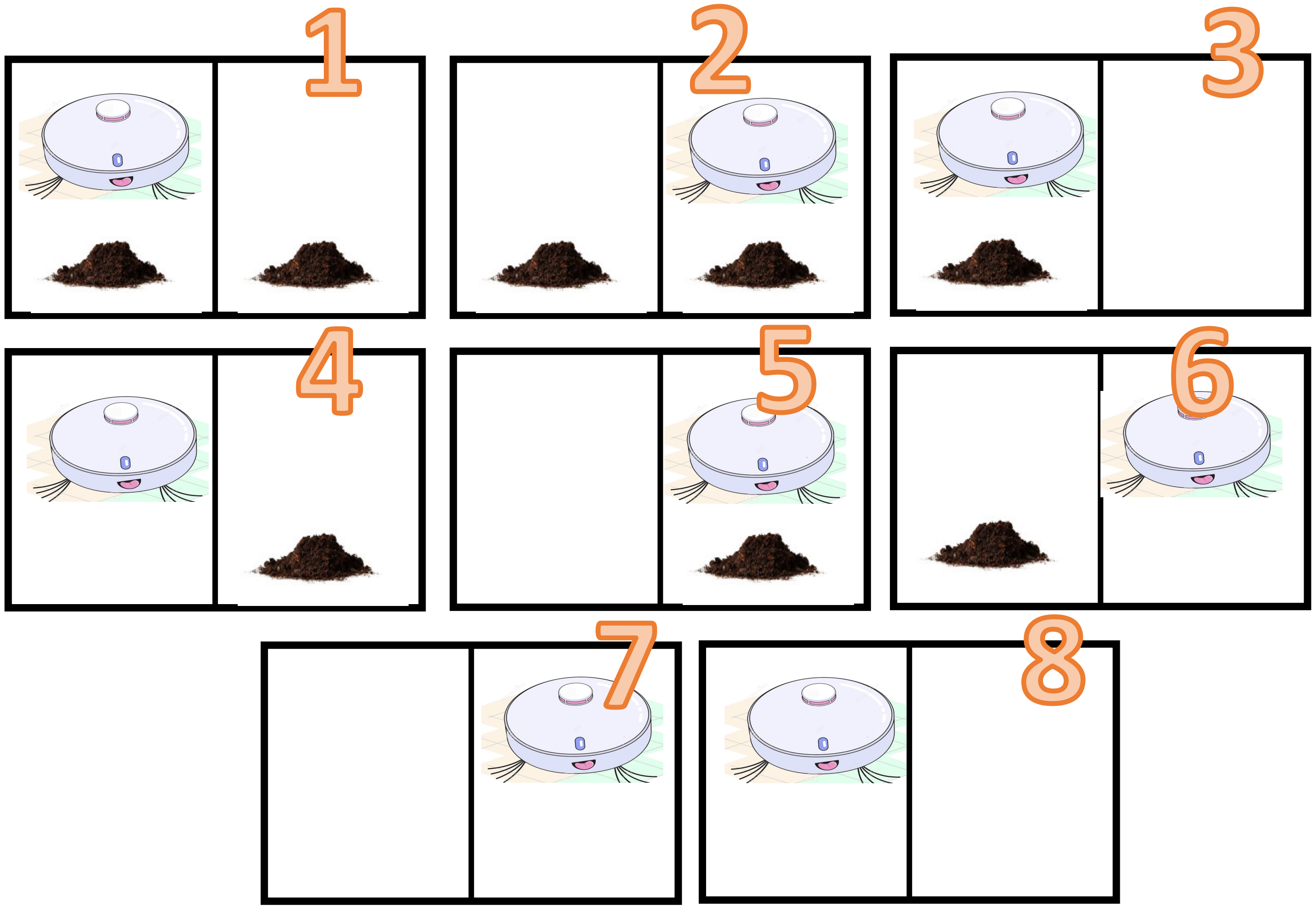
Quantifying Uncertainty

Ref: Artificial Intelligence: A Modern Approach, 4th ed by Stuart Russell and Peter Norvig, chapter 12

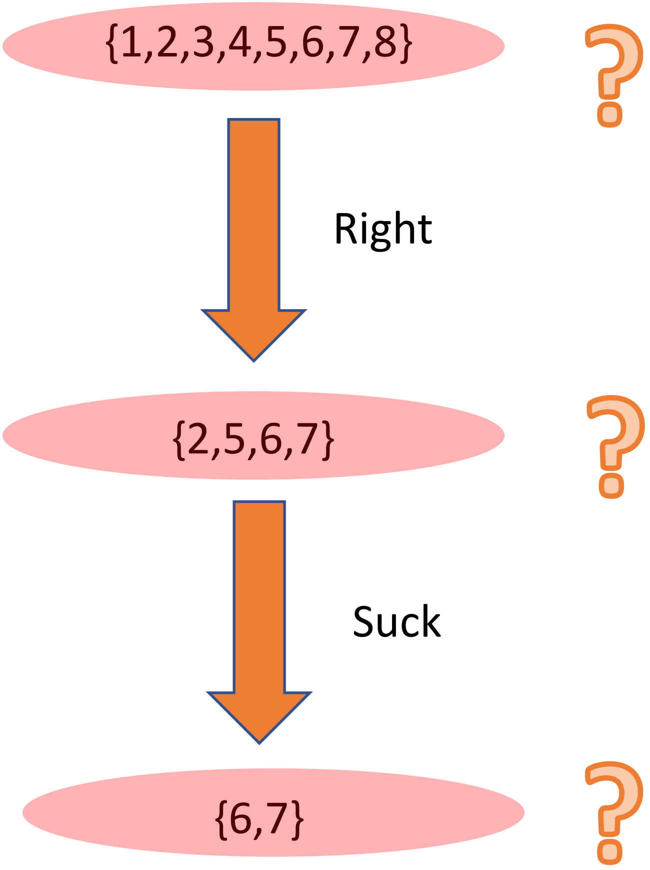
Bayesian Reasoning and Machine Learning, David Barber, chapter 1

Instructor: Masoumeh Heidari (mheidari2@Alaska.edu)

UAA College of Engineering
UNIVERSITY *of* ALASKA ANCHORAGE



Problem-solving agents handle uncertainty by keeping track of a belief state—a representation of the set of all possible world states that it might be in.



{1,2,3,4,5,6,7,8} ?

Right

{2,5,6,7} ?

Suck

{6,7} ?

- This approach works on simple problems, but it has drawbacks:

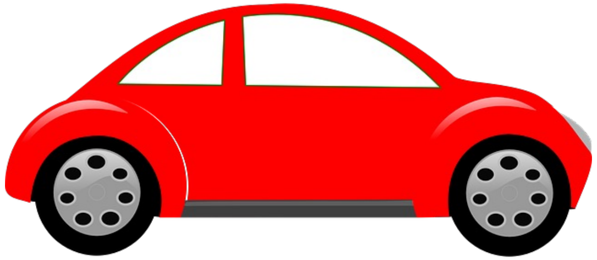
1. The agent must consider every possible explanation for its sensor observations, no matter how unlikely. This leads to a large belief-state full of unlikely possibilities.

2. A correct contingent plan that handles every eventuality can grow arbitrarily large and must consider arbitrarily unlikely contingencies.

3. Sometimes there is no plan that is guaranteed to achieve the goal—yet the agent must act. It must have some way to compare the merits of plans that are not guaranteed.

A_{90}?

A_{180}?



5 miles

Performance Measure:

- Getting to the airport in time for the flight.
- Avoiding a long, unproductive wait at the airport.
- Avoiding speeding tickets along the way.

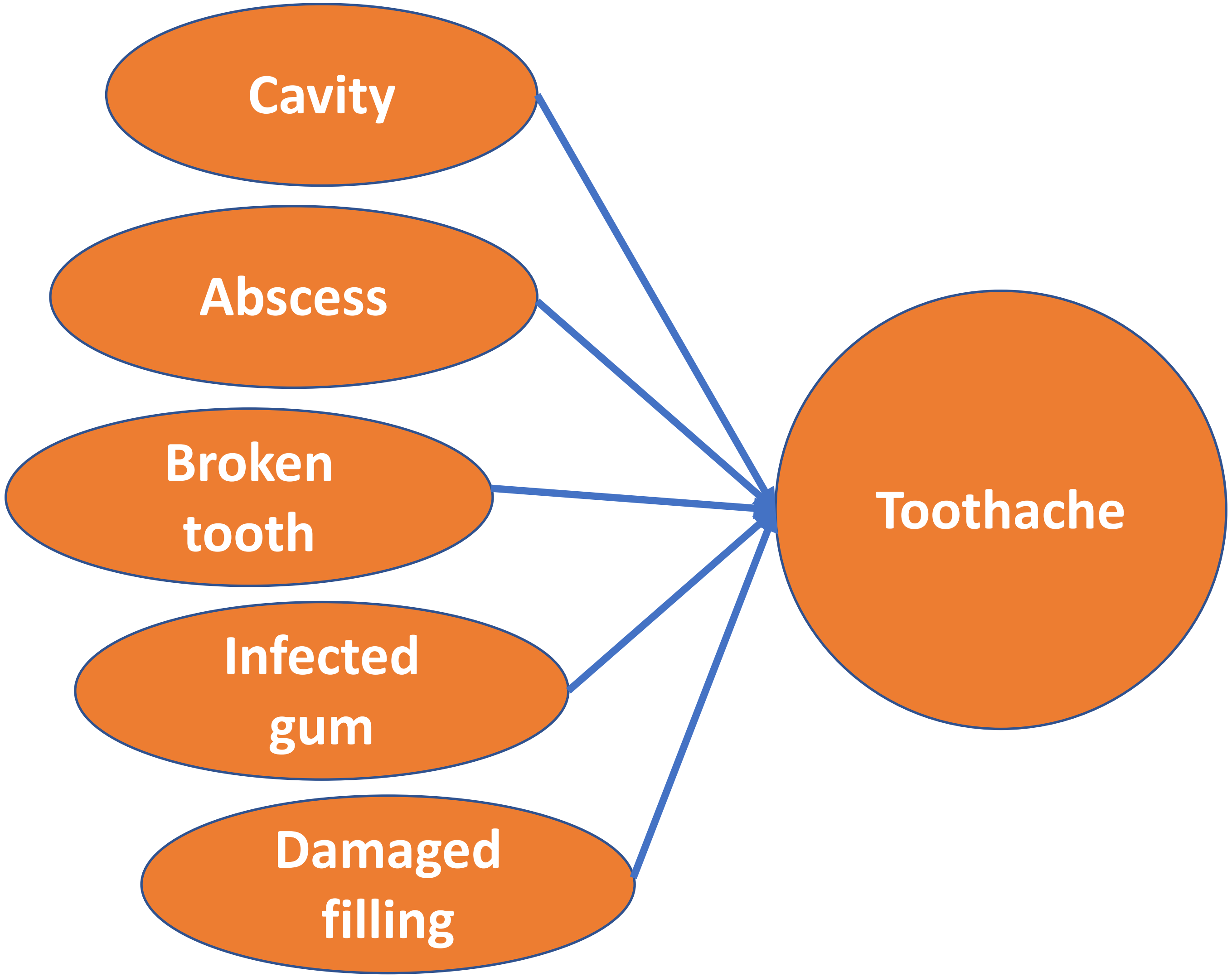
- The agent’s knowledge cannot guarantee any of these outcomes, but it can provide some degree of belief that they will be achieved.

- The right thing to do—the rational decision—therefore depends on both the relative importance of various goals and the likelihood that, and degree to which, they will be achieved.

Summarizing uncertainty

We fail to come up with an exhaustive list due to:

- **Laziness:** It is too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule.
- **Theoretical ignorance:** Medical science has no complete theory for the domain.
- **Practical ignorance:** Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.



Summarizing uncertainty

- The agent’s knowledge can at best provide only a degree of belief.

- Our main tool for dealing with degrees of belief is probability theory.

- A probabilistic agent may have a numerical degree of belief between 0 (certainly false) and 1 (certainly true).

- The theory of probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance.

Summarizing uncertainty

- We might not know for sure what afflicts a particular patient, but we believe that there is, say, an 80% chance—that is, a probability of 0.8—that the patient who has a toothache has a cavity.

- That is, we expect that out of all the situations that are indistinguishable from the current situation as far as our knowledge goes, the patient will have a cavity in 80% of them.

- This belief could be derived from statistical data—80% of the toothache patients seen so far have had cavities—or from some general dental knowledge, or from a combination of evidence sources.

Probability refresher

- **Random variables:** A random variable is some aspect of the world about which we (may) have uncertainty.
  - **C :** Is it cavity?
  - **T:** Is it hot or cold?
  - **R:** Is it raining?

- We denote random variables with capital letters. A set of variables will typically be denoted by a calligraphic symbol ($\mathcal{V}$).

Probability refresher

- The domain of a variable V, dom(V), denotes the states V can take.
  - dom(C) = {T, F}
  - dom(T) = {hot, cold}
  - dom(W) = {sun, rain, snow, fog}
  - dom(D) = [0,∞). How long will it take to drive to work?

- **Probability distribution:** Given a variable,V, its domain dom(V) and a full specication of the probability values for each of the variable states, p(V), we have a distribution for V.

Probability refresher

Probability distribution

T	p(T)
hot	0.5
cold	0.5

W	p(W)
sun	0.35
rain	0.2
snow	0.35
fog	0.1

p(W = rain) = 0.35
p(rain) = 0.35

Probability refresher

- **Events:** are expressions about random variables, such as Two heads in 6 coin tosses.

- Two events are mutually exclusive if they cannot both be true. For example the events “The coin is heads” and “The coin is tails” are mutually exclusive.

Rules of probability for discrete variables:

- The probability p(V = v) of variable V being in state v is represented by a value between 0 and 1.

- p(V = v) = 1 means that we are certain V is in state v. Conversely, p(V = v) = 0 means that we are certain V is not in state v. Values between 0 and 1 represent the degree of certainty of state occupancy.

Rules of probability for discrete variables:

- **Normalization condition:** The summation of the probability over all the states is 1:

$\sum_{v \in dom(V)} p(V = v) = 1$ or $\sum_{V} p(V)$ = 1

- Two variables x and y can interact through

$p(X\ or\ Y\ )$ = p(X) + p(Y) – P(X,Y)

Joint distribution

- A joint distribution over a set of random variables: $V_1, V_2, \ldots V_n$, specifies a real number for each assignment (or *outcome*):

$p(V_1 = v_1, V_2 = v_2, \ldots, V_n = v_n)$ or $p(v_1, v_2, \ldots, v_n)$

- Typically, the events we care about are partial assignments, like p(T=hot)

T	W	P(T,W)
hot	sun	0.3
hot	rain	0.1
hot	snow	0.05
hot	fog	0.05
cold	sun	0.05
cold	rain	0.1
cold	snow	0.3
cold	fog	0.05

Example

- P(+x, +y) ? **0.2**

- P(+x) ? **0.5**

- P(-y OR +x) ? **0.6**

$$P(X,Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal distributions

- Given a joint distribution p(X, Y) the distribution of a single variable is given by $p(X) = \sum_Y p(X, Y)$.

- The process of computing a marginal from a joint distribution is called marginalization. More generally, one has:

$$p(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) = \sum_{X_i} p(X_1, \dots, X_n)$$

- Marginal distributions are sub-tables which eliminate variables.

Marginal distributions

p(T=hot) = p(T=hot, W=sun) + p(T=hot, W=rain) + p(T=hot, W=snow) + p(T=hot, W=fog).

$$p(T) = \sum_{W} T, W$$

T	W	P(T,W)
hot	sun	0.3
hot	rain	0.1
hot	snow	0.05
hot	fog	0.05
cold	sun	0.05
cold	rain	0.1
cold	snow	0.3
cold	fog	0.05

T	p(T)
hot	0.5
cold	0.5

W	p(W)
sun	0.35
rain	0.2
snow	0.35
fog	0.1

$$p(W) = \sum_{T} T, W$$

Marginal distributions

X	Y	P(X,Y)
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$$P(x) = \sum_y P(x, y)$$

$P(X)$

X	P
+x	0.5
-x	0.5

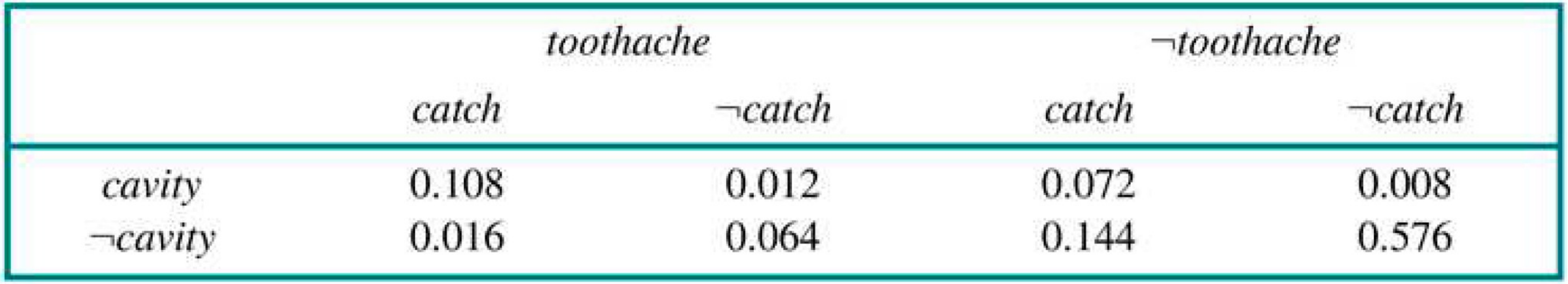
$$P(y) = \sum_x P(x, y)$$

$P(Y)$

Y	P
+y	0.6
-y	0.4

UAA College of Engineering
UNIVERSITY *of* ALASKA ANCHORAGE

Marginal distributions



	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2.$$

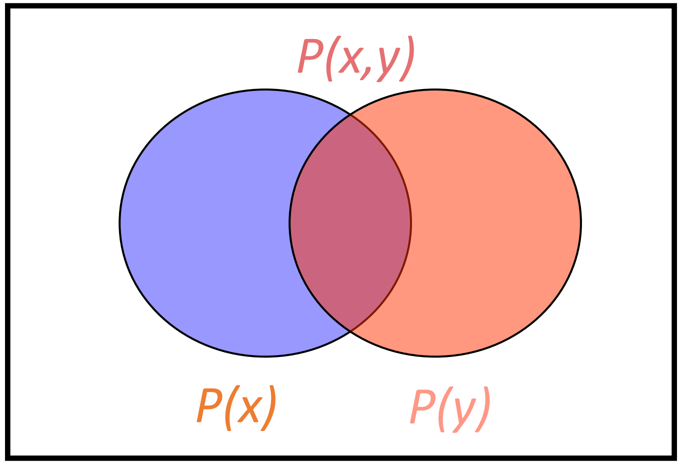
$$\begin{aligned}\mathbf{P}(Cavity) &= \mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch) \\ &+ \mathbf{P}(Cavity, \neg toothache, catch) + \mathbf{P}(Cavity, \neg toothache, \neg catch) \\ &= \langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle + \langle 0.072, 0.144\rangle + \langle 0.008, 0.576\rangle \\ &= \langle 0.2, 0.8\rangle\ .\end{aligned}$$

Conditional probabilities

- The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as:

T	W	P(T,W)
hot	sun	0.3
hot	rain	0.1
hot	snow	0.05
hot	fog	0.05
cold	sun	0.05
cold	rain	0.1
cold	snow	0.3
cold	fog	0.05

$$p(x|y) \equiv \frac{p(x,y)}{p(y)}$$



$$\text{p(T=hot|W=sun)} = \frac{p(T{=}hot\ ,W{=}sun)}{W{=}sun} = \frac{0.3}{0.35}$$

UAA College of Engineering
UNIVERSITY of ALASKA ANCHORAGE

Conditional probabilities

X	Y	P(X,Y)
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- P(+x | +y) ? $\frac{1}{3}$

- P(-x | +y) ? $\frac{2}{3}$

- P(-y | +x) ? $\frac{3}{5}$

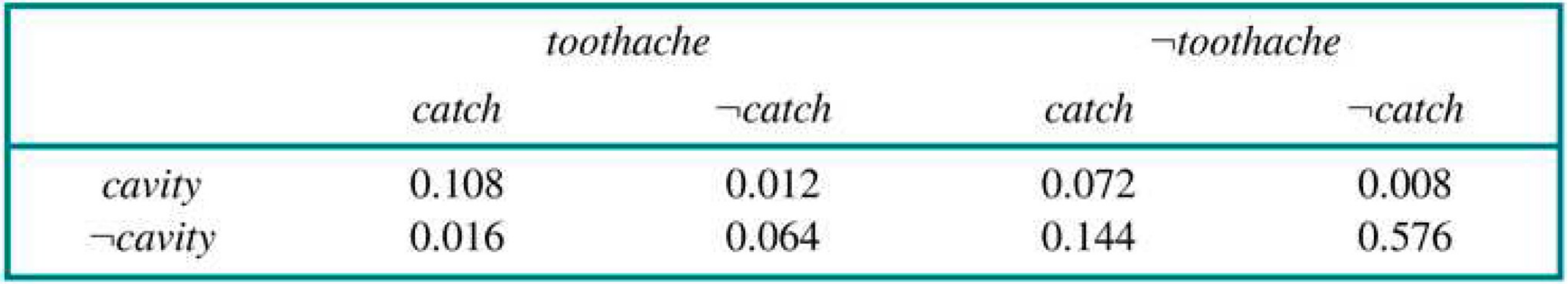
Conditional probabilities

Imagine a circular dart board, split into 20 equal sections, labelled from 1 to 20. Randy, a dart thrower, hits any one of the 20 sections uniformly at random. Hence the probability that a dart thrown by Randy occurs in any one of the 20 regions is p(region i) = 1/20.

A friend of Randy tells him that he hasn't hit the 20 region. What is the probability that Randy has hit the 5 region?

$$p(region\ 5|not\ region\ 20) = \frac{p(region\ 5, not\ region\ 20)}{p(not\ region\ 20)} = \frac{p(region\ 5)}{p(not\ region\ 20)} = \frac{1/20}{19/20} = \frac{1}{19}$$

Example

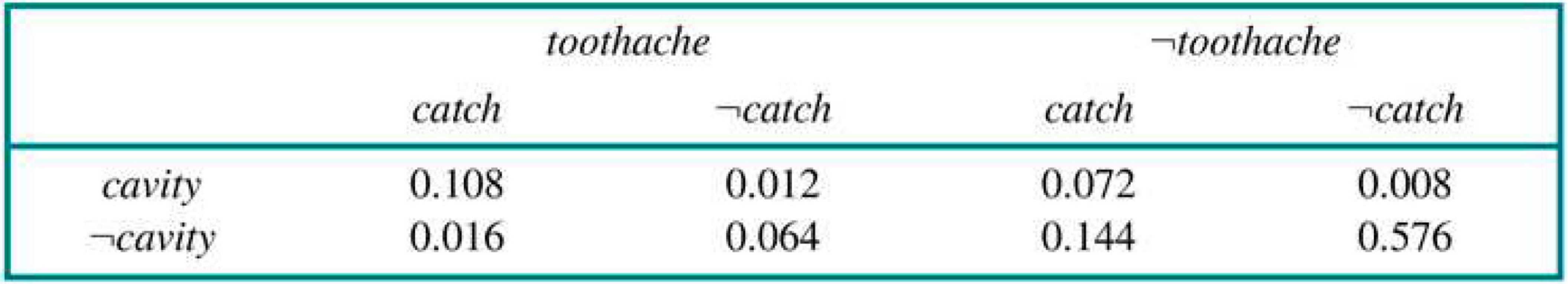


	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

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

$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6.$$

Example

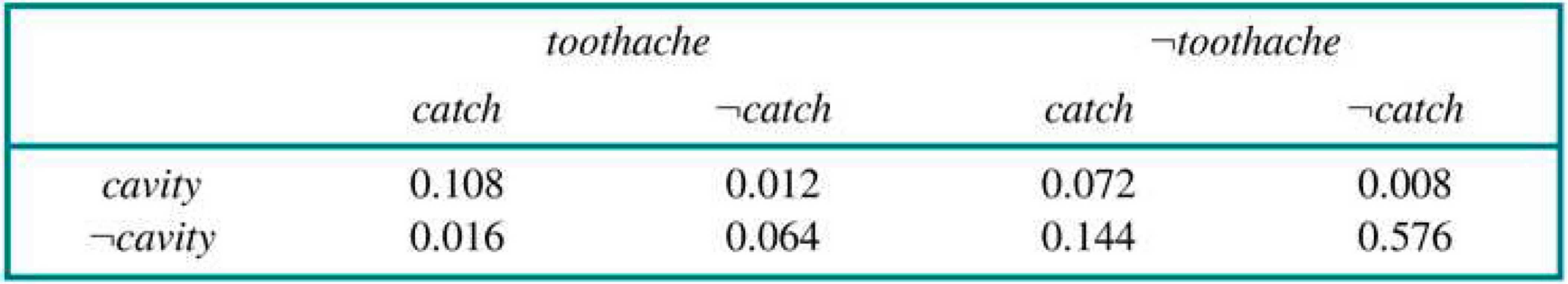


	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$P(\neg cavity \,|\, toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4.$$

Example



	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

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

Conditional distributions

- Conditional distributions are probability distributions over some variables given fixed values of others.

W	P(W|T=hot)
sun	0.6
rain	0.2
snow	0.1
fog	0.1

T	W	P(T,W)
hot	sun	0.3
hot	rain	0.1
hot	snow	0.05
hot	fog	0.05
cold	sun	0.05
cold	rain	0.1
cold	snow	0.3
cold	fog	0.05

Product rule

- The definition of conditional probability, can be written in a different form, called product rule:

$$p(x|y) \equiv \frac{p(x,y)}{p(y)}$$

$$p(x,y) \equiv p(x|y)p(y)$$

The product rule is perhaps easier to remember:
For x and y to be true, we need y to be true, and we also need x to be true given y.

Full joint probability distribution

- A probability model is completely determined by the joint distribution for all of the random variables.

- We use the full joint distribution as the “knowledge base” from which answers to all questions may be derived.

- It does not scale well, however: for a domain described by n Boolean variables, it requires an input table of size $O(2^n)$.

- For these reasons, the full joint distribution in tabular form is seldom a practical tool for building reasoning systems. Instead, it should be viewed as the theoretical foundation on which more effective approaches may be built.

- The remainder of this class introduces some of the basic ideas required in preparation for the development of realistic systems.

Independence

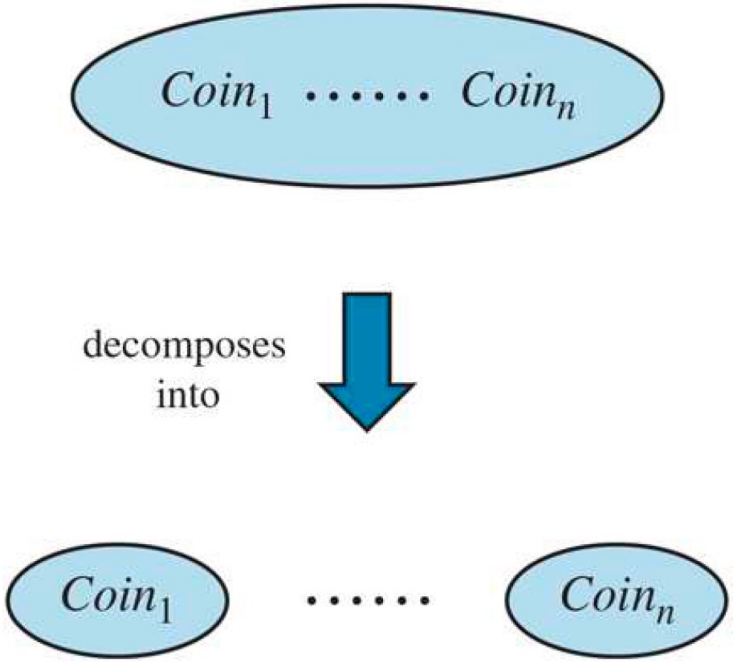
- Let X denote the day of the week in which females are born, and Y denote the day in which males are born.

- dom(X) = dom(Y) = {1,2,3,4,5,6,7}

- We randomly select a woman from the phone book, Alice, and find out that she was born on a Tuesday.

- We also randomly select a male at random, Bob. Before phoning Bob and asking him, what does knowing Alice's birthday add to which day we think Bob is born on?

Independence



$$P(cloud \mid toothache, catch, cavity) = P(cloud).$$

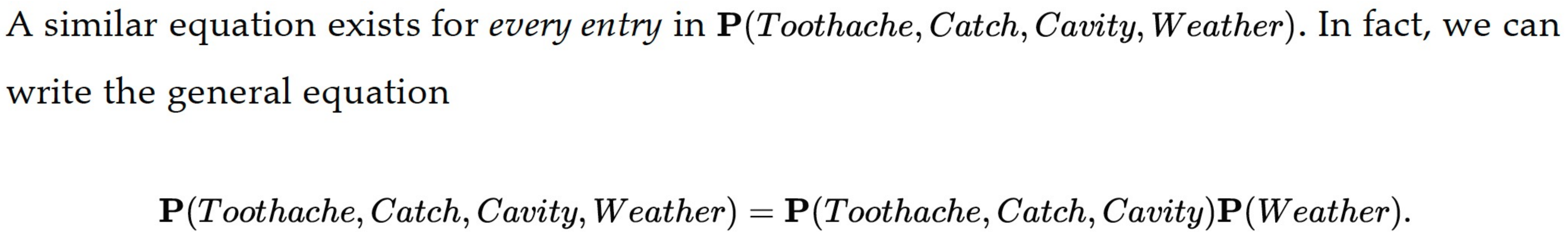
Independence

$$P(toothache, catch, cavity, cloud)$$
$$= P(cloud \mid toothache, catch, cavity) P(toothache, catch, cavity).$$

$$P(toothache, catch, cavity, cloud)$$
$$= \quad P(cloud) \quad P(toothache, catch, cavity).$$

$$P(toothache, catch, cavity, cloud) = P(cloud) P(toothache, catch, cavity).$$

Independence



A similar equation exists for *every entry* in $\mathbf{P}(Toothache, Catch, Cavity, Weather)$. In fact, we can write the general equation

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

Thus, the 32-element table for four variables can be constructed from one 8-element table and one 4-element table.

Independence

- Variables x and y are independent if knowing the state (or value in the continuous case) of one variable gives no extra information about the other variable. Mathematically, this is expressed by

$$p(X,Y) = p(X)p(Y)$$

- Provided that $p(X) \neq 0$ and $p(Y) \neq 0$ independence of X and Y is equivalent to

$$p(X|Y) = p(X) \Leftrightarrow \mathrm{p(Y|X) = p(Y)}$$

Independence

Consider variables X and Y are both binary (their domain consist of two states). We define distribution such that X and Y are always both in a certain joint state:

p(X=a, Y=1) = 1, p(X=a,Y=2) = 0, p(X=b,Y=2) = 0, p(X=b,Y=1)=0

Are X and Y dependent?

p(X=a) = 1, p(X=b)=0, and p(Y=1)=1, p(Y=2)=0
Hence p(X,Y) = p(X)p(Y) for all states of X and Y, and therefore X and Y are independent.

Independence

- Independence assertions are usually based on knowledge of the domain. As the toothache–weather example illustrates, they can dramatically reduce the amount of information necessary to specify the full joint distribution.

- If the complete set of variables can be divided into independent subsets, then the full joint distribution can be factored into separate joint distributions on those subsets.

- When they are available, then, independence assertions can help in reducing the size of the domain representation and the complexity of the inference problem.

Bayes' rule and its use

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

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

$$p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

This equation is known as Bayes' rule (also Bayes' law or Bayes' theorem). This simple equation underlies most modern AI systems for probabilistic inference.

Bayes' rule

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\mathbf{P}(Y)}{\mathbf{P}(X)}.$$

- We will also have occasion to use a more general version conditionalized on some background evidence :

$$\mathbf{P}(Y \mid X, \mathbf{e}) = \frac{\mathbf{P}(X \mid Y, \mathbf{e})\mathbf{P}(Y \mid \mathbf{e})}{\mathbf{P}(X \mid \mathbf{e})}.$$

Applying Bayes' rule

- Often, we perceive as evidence the effect of some unknown cause and we would like to determine that cause. In that case, Bayes' rule becomes:

$$P(cause \mid effect) = \frac{P(effect \mid cause) P(cause)}{P(effect)}.$$

- The conditional probability *p(effect|cause)* quantifies the relationship in the causal direction, whereas *p(cause|effect)* describes the diagnostic direction.

Bayes’ rule

- For example, a doctor knows that the disease meningitis causes a patient to have a stiff neck, say, 70% of the time.

- The doctor also knows some unconditional facts: the prior probability that any patient has meningitis is 1/50000.

- The prior probability that any patient has a stiff neck is 1%.

- P(m|s)?

$$
\begin{aligned}
P(s \mid m) &= 0.7 \\
P(m) &= 1/50000 \\
P(s) &= 0.01
\end{aligned}
$$

Bayes’ rule

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}.$$

$$P(s \mid m) = 0.7$$

$$P(m) = 1/50000$$

$$P(s) = 0.01$$

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014.$$

That is, we expect only 0.14% of patients with a stiff neck to have meningitis. Notice that even though a stiff neck is quite strongly indicated by meningitis (with probability 0.7), the probability of meningitis in patients with stiff necks remains small.

This is because the prior probability of stiff necks (from any cause) is much higher than the prior for meningitis.

Example

- Consider the following fictitious scientific information:

- Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus p(Hamburger Eater |KJ) = 0.9.

- The probability of an individual having KJ is currently rather low, about one in 100,000.

- Assuming eating lots of hamburgers is rather widespread, say p(Hamburger Eater) = 0.5, what is the probability that a hamburger eater will have Kreuzfeld-Jacob disease?

Example 1

$$p(KJ\ |Hamburger\ Eater) = \frac{p(Hamburger\ Eater, KJ\ )}{p(Hamburger\ Eater)} = \frac{p(Hamburger\ Eater|KJ\ )p(KJ\ )}{p(Hamburger\ Eater)}$$

$$= \frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{2}} = 1.8 \times 10^{-5}$$

If the fraction of people eating hamburgers was rather small, p(Hamburger Eater) = 0.001, what is the probability that a regular hamburger eater will have Kreuzfeld-Jacob disease?

This is much higher than in scenario (1) since here we can be more sure that eating hamburgers is related to the illness.

$$\frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{1000}} \approx 1/100$$

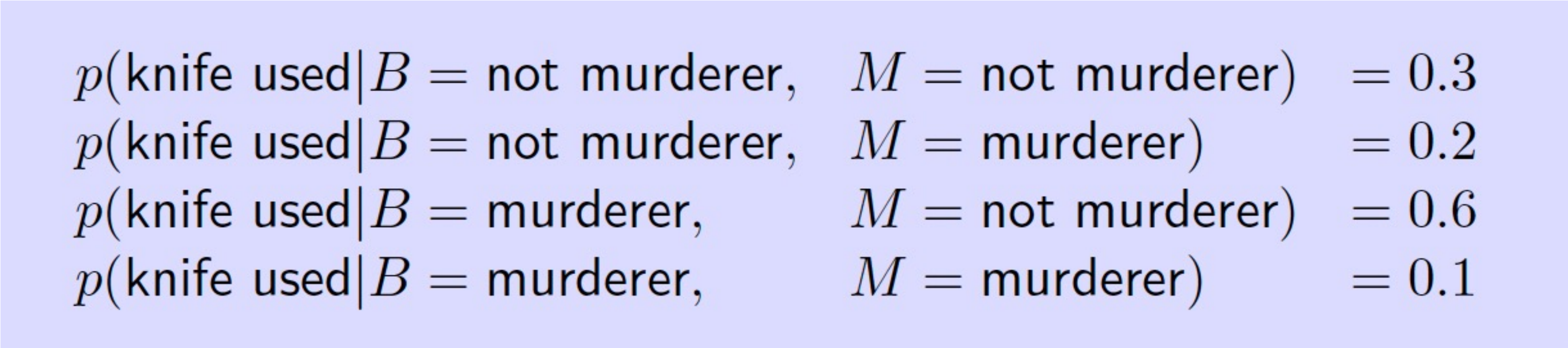
Example 2

- Inspector Clouseau arrives at the scene of a crime. The victim lies dead in the room alongside the possible murder weapon, a knife. The Butler (B) and Maid (M) are the inspector’s main suspects and the inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer.

- These beliefs are independent in the sense that p(B,M) = p(B)p(M). (It is possible that both the Butler and the Maid murdered the victim or neither).

Example 2

- What are the variables?
  - M, B, K
  - dom(M) = dom(B) = {murderer, not murderer}, dom(K) = {knife used, knife not used}
  - P(B= murderer) = 0.6, p(M= murderer)=0.2



$$\begin{aligned}
p(\text{knife used}|B &= \text{not murderer}, & M &= \text{not murderer}) &= 0.3\\
p(\text{knife used}|B &= \text{not murderer}, & M &= \text{murderer}) &= 0.2\\
p(\text{knife used}|B &= \text{murderer}, & M &= \text{not murderer}) &= 0.6\\
p(\text{knife used}|B &= \text{murderer}, & M &= \text{murderer}) &= 0.1
\end{aligned}$$

Assuming that the knife is the murder weapon, what is the probability that the Butler is the murderer?

UAA College of Engineering
UNIVERSITY *of* ALASKA ANCHORAGE

Example 2

$$
\begin{aligned}
p(\text{knife used} | B = \text{not murderer}, \quad & M = \text{not murderer}) && = 0.3 \\
p(\text{knife used} | B = \text{not murderer}, \quad & M = \text{murderer}) && = 0.2 \\
p(\text{knife used} | B = \text{murderer}, \quad & M = \text{not murderer}) && = 0.6 \\
p(\text{knife used} | B = \text{murderer}, \quad & M = \text{murderer}) && = 0.1
\end{aligned}
$$

p(B,M) = p(B)p(M)
p(B= murderer) = 0.6, p(M= murderer)=0.2

p(b) = 0.6, p(m) =0.2

P(b|k) = ? $\frac{p(b,k)}{P(k)}$? ?

P(k|¬b,¬m)	0.3
P(k|¬b, m)	0.2
P(k|b,¬m)	0.6
P(k|b, m)	0.1

$$p(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = \sum_{X_i} p(X_1, \ldots, X_n)$$

P(b,k) = $\sum_M p(b,k) = p(b,k,m) + p(b,k,\neg m)$

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

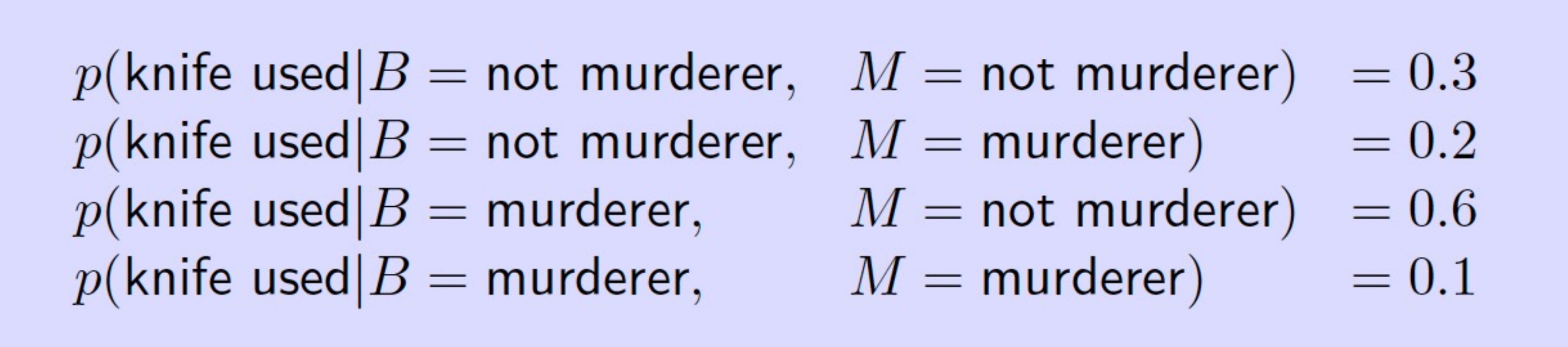
P(b,k) = $p(k|m,b)p(m,b) + p(k|\neg m,b)p(\neg m,b)$

$$p(x,y) = p(x)p(y)$$

P(b,k) = $p(k|m,b)p(m)p(b) + p(k|\neg m,b)p(\neg m)p(b)$

P(b,k) = $0.1 \times 0.2 \times 0.6 + 0.6 \times 0.8 \times 0.6 = 0.3$

Example 2



$$
\begin{aligned}
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{not murderer}) &= 0.3 \\
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{murderer}) &= 0.2 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{not murderer}) &= 0.6 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{murderer}) &= 0.1
\end{aligned}
$$

p(B,M) = p(B)p(M)
p(B= murderer) = 0.6, p(M= murderer)=0.2

p(b) = 0.6, p(m) =0.2

P(b|k) = ? $\frac{p(b,k)}{P(k)}$?

P(k|¬b,¬m)	0.3
P(k|¬b, m)	0.2
P(k|b,¬m)	0.6
P(k|b, m)	0.1

P(k) = $p(k|m,b)p(m)p(b) + p(k|\neg m,b)p(\neg m)p(b) + p(k|m,\neg b)p(m)p(\neg b) + p(k|\neg m,\neg b)p(\neg m)p(\neg b)$

P(k) = $0.1 \times 0.2 \times 0.6 + 0.6 \times 0.8 \times 0.6 + 0.2 \times 0.2 \times 0.4 \; + \; 0.3 \times 0.8 \times 0.4 = 0.412$

P(b|k) = $\frac{p(b,k)}{P(k)} \; = \frac{0.3}{0.412}$ ≈ 0.73

UAA College of Engineering
UNIVERSITY of ALASKA ANCHORAGE

Example 2

$$p(b|k) = \sum_{M} p(b, M|k) = \sum_{M} \frac{p(b, M, k)}{p(k)} = \frac{\sum_M p(k|b, M) p(b, M)}{\sum_{M,B} p(k|B, M) p(B, M)} = \frac{p(b) \sum_M p(k|b, M) p(M)}{\sum_B p(B) \sum_M p(k|B, M) p(M)}$$

$$p(B = \text{murderer}|\text{knife used}) = \frac{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10} \times \frac{2}{10} + \frac{8}{10} \times \frac{3}{10}\right)} = \frac{300}{412} \approx 0.73$$

Example 3

- A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

Example 3

- A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

$$P(t|d) = 0.995$$

Example 3

- A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

$$P(\neg t \mid \neg d) = 0.999$$

Example 3

- A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

$$P(d) = \frac{1}{100000}$$

Example 3

- A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

$P(t|d) = 0.995$

$P(\neg t|\neg d) = 0.999$

$P(d) = \frac{1}{100000}$

$p(d|t) = ?$

Example 3

- A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

$P(t|d) = 0.995$

$P(\neg t|\neg d) = 0.999$

$P(d) = \frac{1}{100000}$

$$p(d|t) = \frac{p(t|d)p(d)}{p(t)} = \frac{p(t|d)p(d)}{\sum_D p(t,D)} = \frac{p(t|d)p(d)}{p(t,d)+p(t,\neg d)}$$

Example 3

P(t|d) = 0.995

P(¬t|¬d) = 0.999

P(d) = $\frac{1}{100000}$

$$p(d/t) = \frac{p(t|d)p(d)}{p(t)} = \frac{p(t|d)p(d)}{\sum_D p(t,D)} = \frac{p(t|d)p(d)}{p(t,d)+p(t,\neg d)}$$

$$= \frac{p(t|d)p(d)}{p(t|d)p(d)+p(t|\neg d)p(\neg d)} = \frac{0.995 \times \frac{1}{100000}}{0.995 \times \frac{1}{100000} + 0.001 \times \frac{99999}{100000}}$$

Example 4

- A casino uses two kinds of dice. One kind of die is fair and is used 99% of the time. The unfair die rolls a six 50% of the time and is used for the rest of the time.

- If we pick up a single die at random, how likely is it that we will roll a six?

Example 4

- A casino uses two kinds of dice. One kind of die is fair and is used 99% of the time. The unfair die rolls a six 50% of the time and is used for the rest of the time.

- If we pick up a single die at random, how likely is it that we will roll a six?

P(six) = $\sum_D p(six, D) = p(six, d_1) + p(six, d_2) = p(six|d_1)p(d_1) + p(six|d_2)p(d_2) \;= \frac{1}{6}\times 0.99 + \frac{1}{2}\times 0.01 = 0.17$

Example 5

- Larry is typically late for school. If Larry is late, we denote this with L = late, otherwise, L = not late. When his mother asks whether or not he was late for school he never admits to being late. The response Larry gives R_L is represented as follows:
  $p(R_L = not\ late|L = not\ late) = 1, p(R_L = late|L = late) = 0$
  $p(R_L = late|L = not\ late) = 0, p(R_L = not\ late|L = late) = 1$

- Given that R_L =not late, what is the probability that Larry was late, i.e. $p(L = late|R_L = not\ late)$?

Example 5

$$p(R_L = not\ late|L = not\ late) = 1, p(R_L = late|L = late) = 0$$
$$p(R_L = late|L = not\ late) = 0, p(R_L = not\ late|L = late) = 1$$

$$p(L = late|R_L = not\ late)?$$

$$= \frac{p(R_L = not\ late|L = late)p(L = late)}{p(L = late, R_L = not\ late) + p(L = not\ late, R_L = not\ late)} = \text{p(late)}$$

Conditional independence

- X and Y are conditionally independent, given a third variable Z, if

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

- Are *Toothache* and *Catch* independent?

- Are *Toothache* and *Catch* independent given *Cavity*?
    - If the probe catches in the tooth, then it is likely that the tooth has a cavity and that the cavity causes a toothache. These variables are independent, however, given the presence or the absence of a cavity.

$$p(toothache, catch|Cavity) = p(toothache|Cavity)p(catch|Cavity)$$

Conditional independence

- $p(X, Y|Z) = p(X|Z)p(Y|Z)$

- $p(X|Y, Z) = p(X|Z)$ and $p(Y|X, Z) = p(Y|Z)$

- Similar to the absolute independence, conditional independence allows a decomposition of the full joint distribution into much smaller pieces.

$p(Toothache, Catch, Cavity) = p(Toothache, Catch|Cavity)p(Cavity)$ Product rule

= $p(Toothache|Cavity)p(Catch|Cavity)p(Cavity)$

Example 6

- In the Larry's example, assume Larry's sister Sue always tells the truth to her mother as to whether or not Larry was late for School. Given that R_S = late and R_L = not late, what is the probability that Larry was late?

$$p(R_s = not\ late|L = not\ late) = 1, p(R_s = late|L = late) = 1$$
$$p(R_s = late|L = not\ late) = 0, p(R_s = not\ late|L = late) = 0$$

$p(L = late|R_L = not\ late, R_S = late)$ =

$$\frac{p(R_S = late|L = late)p(R_L = not\ late|L = late)p(L=late)}{p(R_L=not\ late, R_S=late)}$$

Example 6

$$\frac{1 \times 1 \times p(L{=}late)}{p(R_L{=}not\ late, R_S{=}late)} =$$

$$\frac{p(L{=}late)}{p(R_s = late|L = late)p(R_L = not\ late|L = late)p(L{=}late) + p(R_s = late|L = not\ late)p(R_L = not\ late|L = not\ late)p(L{=}not\ late)} =$$

$$\frac{p(L = late)}{p(L = late) + 0} = 1$$

Since Larry's mother knows that Sue always tells the truth, no matter what Larry says, she knows he was late.

Naïve Bayes model

- Sometimes a single cause directly influences a number of effects, all of which are conditionally independent, given the cause.
- The full joint distribution can be written as:

$$p(Cause, Effect_1, Effect_2, Effect_n) =$$

$$p(Cause|Effect_1)p(Cause|Effect_2)\ldots p(Cause|Effect_n) =$$

$$p(Cause)\prod_i p(Effect_i|Cause)$$

Naïve Bayes model

- It is called “naïve” because it is often used (as a simplifying assumption) in cases where the “effect” variables are not strictly independent given the cause variable.

- In practice, naive Bayes systems often work very well, even when the conditional independence assumption is not strictly true.

Naïve Bayes model

- We can use Naïve bayes model to obtain the probability of the cause given some observed effects.

- Assuming observed effects as E=e, while the remaining effect variables Y are unobserved, the standard method for inference from joint distribution can be applied:

$$p(Cause|\boldsymbol{e}) = \alpha \sum_{y} p(Cause, \boldsymbol{e}, \boldsymbol{y})$$

Naïve Bayes model

$$p(Cause, Effect_1, Effect_2, Effect_n) = p(Cause)\prod_i p(Effect_i|Cause)$$

$p(Cause|\boldsymbol{e}) = \ \alpha \sum_y p(Cause, \boldsymbol{e}, \boldsymbol{y})$=

$\alpha \ \sum_{\boldsymbol{y}} p(Cause)p(\boldsymbol{y}|Cause)(\prod_j p(e_j|Cause))$=

$\alpha \, p(Cause)\left(\prod_j p\left(e_j\middle|Cause\right)\right)\sum_y p(y|Cause)$=

$\alpha \, p(Cause)\prod_j p\left(e_j\middle|Cause\right)$

- For each possible cause, multiply the prior probability of the cause by the product of the conditional probabilities of the observed effects given the cause; then normalize the result.

Text classification with naïve Bayes

- Consider two example sentences, taken from newspaper articles:

    1. Stocks rallied on Monday, with major indexes gaining 1% as optimism persisted over the first quarter earnings season.
    2. Heavy rain continued to pound much of the east coast on Monday, with flood warnings issued in New York City and other locations.

- Can you classify each sentence into a *category*?
    - news, sports, business, weather, or entertainment

Text classification with naïve Bayes

- The naive Bayes model consists of the prior probabilities p(*Category*) and the conditional probabilities p($HasWord_i$|*Category*).
- For each category *c*, p(*Category = c*) is estimated as the fraction of all previously seen documents that are of category *c*.
  - For example, if 9% of articles are about weather, we set p(*Category=weather*) = 0.09.
- Similarly, p($HasWord_i$|*Category*) is estimated as the fraction of documents of each category that contain word *i* ;
  - If 37% of articles about business contain word 6 “stocks” so p($HasWord_6$*=true*|*Category = business*) = 0.37.
- To categorize a new document, we check which key words appear in the document and then apply the Naïve Bayes model to obtain the posterior probability distribution over categories.

Text classification with Naïve Bayes

- The naive Bayes model assumes that words occur independently in documents, with frequencies determined by the document category.
- This independence assumption is clearly violated in practice. For example, the phrase “first quarter” occurs more frequently in business (or sports) articles than would be suggested by multiplying the probabilities of “first” and “quarter.”
- The violation of independence usually means that the final posterior probabilities will be much closer to 1 or 0 than they should be; in other words, the model is overconfident in its predictions.
- Even with these errors, the ranking of the possible categories is often quite accurate.

Text classification with Naïve Bayes

- Naïve Bayes models are widely used for language determination, document retrieval, spam filtering, and other classification tasks.

- For tasks such as medical diagnosis, where the actual values of the posterior probabilities really matter—for example, in deciding whether to perform an appendectomy—one would usually prefer to use the more sophisticated models.

Example 1

- 30 % of all emails are spam.
- 15 % of spam emails contain the word “travel”.
- 1% of nonspam emails contain the word “travel”.
- What is the probability that an email is spam, given that it contains the word “travel”?

$$p(s|t) = \frac{p(t|S)p(s)}{p(t|S)p(s)+p(t|\neg S)p(\neg s)} = \frac{0.15 \times 0.3}{0.15 \times 0.3 + 0.01 \times 0.7} = \frac{0.045}{0.052} \approx 0.86$$

$$\mathrm{p}(\neg s|t) = \alpha\ \mathrm{p}(\mathrm{t}|\neg s)\mathrm{p}(\neg s) = \alpha \times 0.01 \times 0.7 = \alpha \times 0.007$$

$$\mathrm{p}(s|t) = \alpha\ \mathrm{p}(\mathrm{t}|s)\mathrm{p}(s) = \alpha \times 0.15 \times 0.3 = \alpha \times 0.045$$

Example 2

$p(Cause|\boldsymbol{e}) = \alpha \sum_y p(Cause, \boldsymbol{e}, \boldsymbol{y})$=

$\alpha\, p(Cause) \prod_j p\left(e_j \middle| Cause\right)$

- 30 % of all emails are spam.
- 15 % of spam emails contain the word “travel”, and 50 % of them contain the word “enjoy”.
- 1% of nonspam emails contain the word “travel”, and 1 % of them contain “enjoy”.
- What is the probability that an email is spam, given that it contains the word “travel” and “enjoy”?

P(s|t,e) = α p(t,e|s)p(s). Naïve Bayes: p(s|t,e) = αp(s)p(t|s)p(e|s) =

$\alpha \times 0.3 \times 0.15 \times 0.5 = 0.0225\, \alpha$

Example 3

Document #	Words	Class
1	Everything in Alaska is beautiful!	Alaska
2	Anchorage is a city in Alaska.	Alaska
3	AI is the most important course.	AI
4	I had fun in Alaska.	Alaska
5	AI is fun!	AI
6	They play football.	sport
7	Everyone loves AI because it is fun!	?

C	P(C)
Alaska	$\frac{3}{6}$
AI	$\frac{2}{6}$
sport	$\frac{1}{6}$

Keywords
Alaska
Anchorage
AI
Fun
Play

C	P(Alaska|C)
Alaska	1
AI	0
sport	0

C	P(Anchorage|C)
Alaska	$\frac{1}{3}$
AI	0
sport	0

C	P(AI|C)
Alaska	0
AI	1
sport	0

C	P(Fun|C)
Alaska	$\frac{1}{3}$
AI	$\frac{1}{2}$
sport	0

C	P(Play|C)
Alaska	0
AI	0
sport	1

$$p(Alaska|AI\ ,fun) = \ \alpha\, p(AI|Alaska)p(fun|Alaska)p(Alaska) = 0$$

$$p(AI|AI\ ,fun) = \ \alpha\, p(AI|AI)p(fun|AI)p(AI) = \alpha \times 1 \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}\alpha$$

$$p(sport|AI\ ,fun) = \ \alpha\, p(AI|sport)p(fun|sport)p(sport) = 0$$

Recap

- Uncertainty arises because of both laziness and ignorance. It is inescapable in complex, nondeterministic, or partially observable environments.
- The full joint probability distribution specifies the probability of each complete assignment of values to random variables. It is usually too large to create or use in its explicit form, but when it is available it can be used to answer queries simply by adding up entries for the possible worlds corresponding to the query propositions.
- Absolute independence between subsets of random variables allows the full joint distribution to be factored into smaller joint distributions, greatly reducing its complexity.
- Bayes’ rule allows unknown probabilities to be computed from known conditional probabilities, usually in the causal direction. Applying Bayes’ rule with many pieces of evidence runs into the same scaling problems as does the full joint distribution.

Recap

- Conditional independence brought about by direct causal relationships in the domain allows the full joint distribution to be factored into smaller, conditional distributions.
- The naive Bayes model assumes the conditional independence of all effect variables, given a single cause variable; its size grows linearly with the number of effects.