COMP 341 Intro to AI
Representing and Quantifying Uncertainty

Asst. Prof. Barıș Akgün

Koç University

COMP341 So Far

- Agents
- Search
  - Uninformed
  - Informed
  - Adversarial
- Local Search
- CSPs
- Adversarial Search

Uncertainty

- Real world is uncertain
- Where is this uncertainty coming from?
  - Partial Observability (e.g. fog of war, opponents' hand in poker, traffic)
  - Noisy sensors (e.g. GPS, cameras in low light, traffic reports, sonar)
  - Uncertain action outcomes (e.g. flat tire, wheel slip, object too heavy to lift)
  - Unexpected Events (e.g. sudden car accident, earthquake, meteor hitting)
  - Inherent Stochasticity (e.g. quantum physics) - this affects sensors and actuators as well
  - Complexity of Modelling (e.g. market behavior, predicting traffic) – related to unexpected events and other agents

- We need to represent and quantify uncertainty to be able to solve problems!

Uncertainty

- General situation:
    - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
    - **Unobserved variables**: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
    - **Model**: Agent knows something about how the known variables relate to the unknown variables

- Probabilistic reasoning gives us a framework for managing <u>our beliefs</u> and <u>knowledge</u>

Uncertainty

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?

Some Problems:

1. partial observability (road state, other drivers' plans, etc.)
2. noisy sensors (traffic reports)
3. uncertainty in action outcomes (flat tire, etc.)
4. immense complexity of modeling and predicting traffic

If just TRUE/FALSE

1. risks falsehood: “A_{25} will get me there on time”, or
2. 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.” (also look up qualification problem)

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

Probability

- Cannot list all possible conditions for a given statement
- Cannot deduce the truth value for all the statements for sure
- Instead of absolute statements, use **probability** to summarize uncertainty
- Probabilities relate to the degree that an agent believes a statement to be true

$$P(A_{25} | no\ reported\ accidents) = 0.06$$

- The probability changes with new information (evidence)

$$P(A_{25} | no\ reported\ accidents, 5AM) = 0.15$$

- For this class, we treat probability statements as not assertions about the world but as assertions about the knowledge state of the agent

Decision Making Under Uncertainty

- Which action would you chose given 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

- Depends on preferences for missing flight vs. time spent waiting, etc. and willingness to take risk
- How to represent these? Utilities!
  - Utility theory is used to represent and infer preferences
  - Decision theory = probability theory + utility theory

Not going to go into detail

Will cover basics

Example: Utility is exp(-t/500)

Very Brief History of Probability

- Calculation of probabilities 15th and 16th century, inspired by games of chance
- Formal foundations: mid 17th century correspondence between Pascal and Fermat
  - The problem that started it all: Would you bet money to get a roll of double sixes in 24 die rolls?
  - Look up Chevalier de Mere's Problem
- A few years later, first book by Huygens
- John Gaunt analyzes data on death and age, William Petty suggests using similar methods for government decisions (in 17th century!)
- 18th century: Life insurance turns out to be very profitable
- 1761 Thomas Bayes – Bayes' Theorem
- 1812 Laplace takes probability from games of chance to scientific problems
- 1933 Kolmogorov axioms (based on measure theory)
- 1946 Cox's theorem (a Bayesian axiomatization)

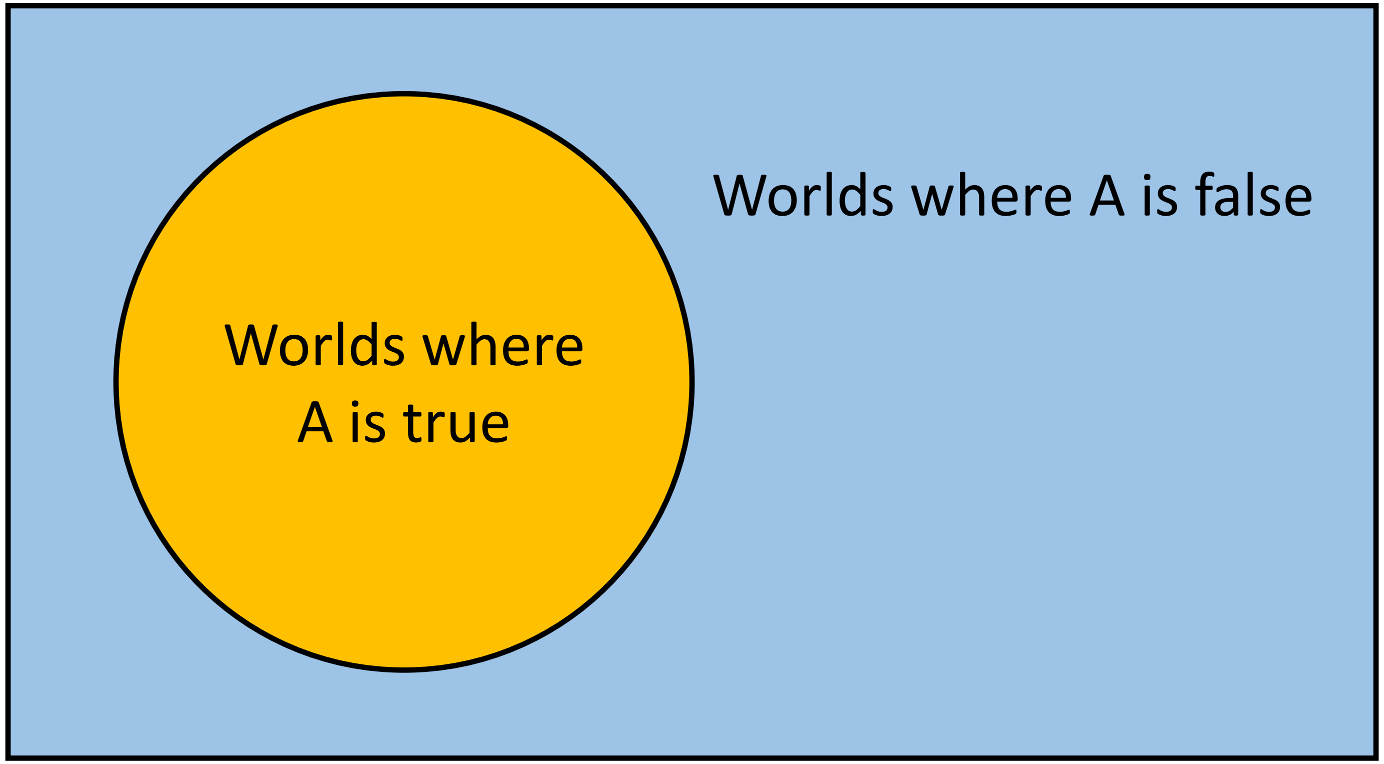
Probability Basics

Random Variables

- A random variable is some aspect of the world about which we are uncertain of
  - Cavity: Do I have a tooth cavity?
  - Weather: How is the weather today?
  - A: How long will it take me to drive to the airport
  - D: Dice roll
- Random variables have domains (remember CSP variables!)
  - Cavity: {true, false}
  - Weather: {sunny, rain, cloudy, snow}
  - A: $[0, \infty)$
  - D: {1,2,3,4,5,6}
- The domain of a random variable is also called the sample space

A Simple Notion of Probability

P(A): Fraction of all possible worlds where **A** is **true**



Space of all possible worlds

Notation

- Let the set Ω be the sample space (e.g. 6 possible rolls of a dice)
- Let $\omega \in \Omega$ be a sample point/possible world/atomic event (e.g. a roll of 3)
- A probability space/probability model is a sample space with an assignment $P(D = \omega)$ for every $\omega \in \Omega$
  - Shorthand $P(D = \omega) = P(\omega)$ if all elements are unique
- An event A is any subset of Ω , $P(A) = \sum_{\omega \in A} P(\omega)$ (e.g. rolling 3 or 6)
- The event space, F, is the power set of Ω
- Random variables are or start with capital letters (e.g. Weather)
- The values are or start with lower case letters (e.g. cloudy)

Probability Axioms

1. Probability of an event is a non-negative real number
$$P(E) \in \mathbb{R}, P(E) \geq 0, \qquad \forall E \in F$$
2. Probability of the entire sample space is 1
$$P(\Omega) = 1$$
3. Probability of observing mutually exclusive events (aka disjoint sets) is additive
$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i),$$

where events E_i are mutually exclusive (i.e. they are disjoint sets)

Immediate consequences:

- $\sum_{\omega \in \Omega} P(\omega) = 1$ (from 2 and 3)
- $1 \geq P(E) \geq 0$ (from 1,2 and 3)
- If $A \subseteq B, P(A) \leq P(B)$ (from 3)

Exercise

- Show that $P(\emptyset) = 0$

Axiom 2: $P(\Omega) = 1$

Axiom 3: $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$

Any set is disjoint with the empty set, including other empty sets

$$\Omega \cup \emptyset \cup \emptyset \cup \emptyset \ldots = \Omega$$

$$P(\Omega \cup \emptyset \cup \emptyset \cup \emptyset \ldots) = P(\Omega) = 1$$

$$P(\Omega) + P(\emptyset) + P(\emptyset) + P(\emptyset) + \cdots = 1$$

$$1 + P(\emptyset) + P(\emptyset) + P(\emptyset) + \cdots = 1$$

$$P(\emptyset) = 0$$

Exercise

- Show that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

A and $B \backslash (A \cap B)$ are mutually exclusive where “\” is the set subtraction: $B \backslash A = \{a \in B \mid a \notin A\}$.

Then (axiom 3):

$$P(A, B \backslash (A \cap B)) = P(A) + P(B \backslash (A \cap B))$$

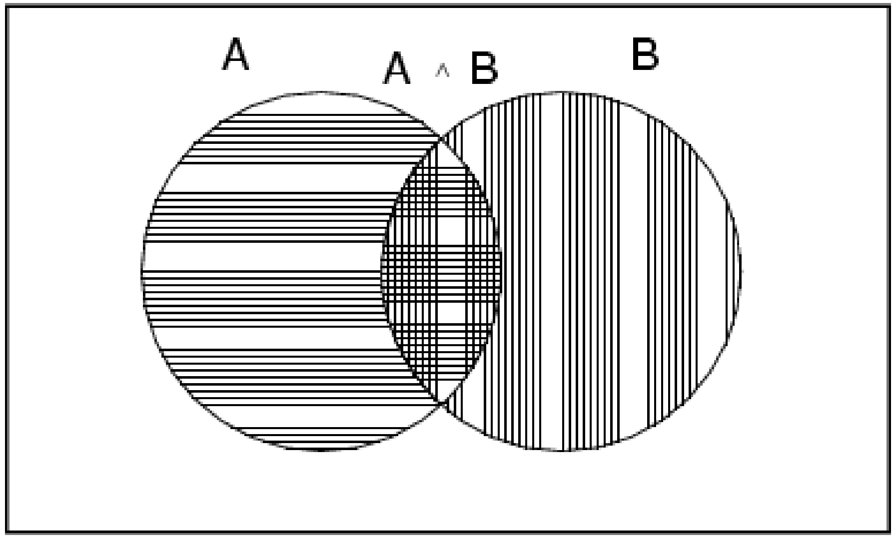
We also have:

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

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

Plug in:

$$P(A, B \backslash (A \cap B)) = P(A) + P(B) - P(A \cap B)$$



Exercise

- Show that $\mathrm{P}(\mathrm{A}^{\mathrm{c}}) = \mathrm{P}(\mathrm{F} \backslash \mathrm{A}) = 1 - \mathrm{P}(\mathrm{A})$

It is easy to see that the event and its complement are mutually exclusive

It is also easy to see that $\mathrm{P}(\mathrm{F}) = \mathrm{P}(\Omega) = 1$

$$\mathrm{P}(\mathrm{F}) = \mathrm{P}(\mathrm{A}^{\mathrm{c}} \cup A) = \mathrm{P}(\mathrm{A}^{\mathrm{c}}) + \mathrm{P}(\mathrm{A})$$

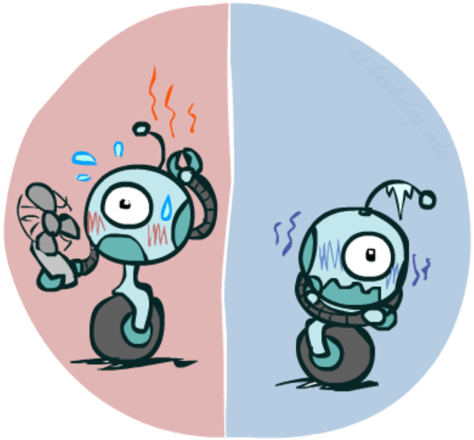
$$1 = \mathrm{P}(\mathrm{A}^{\mathrm{c}}) + \mathrm{P}(\mathrm{A})$$

$$\mathrm{P}(\mathrm{A}^{\mathrm{c}}) = 1 - \mathrm{P}(\mathrm{A})$$

Probability Distributions

- Associate a probability with each value

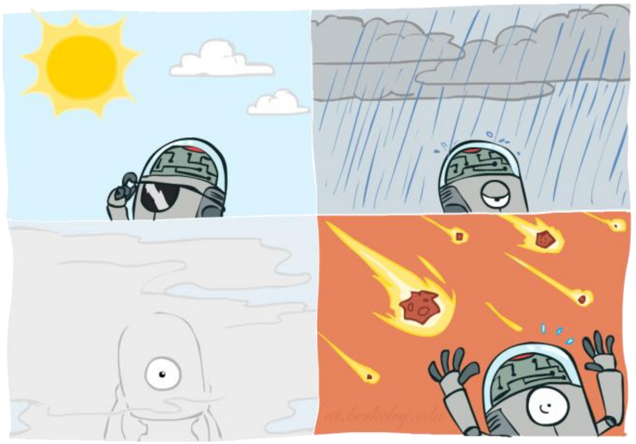
  - Temperature:



$$P(T)$$

T	P
hot	0.5
cold	0.5

  - Weather:



$$P(W)$$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Prior Probability

- Prior or unconditional probabilities reflect agent’s belief prior to arrival of any (new) evidence

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

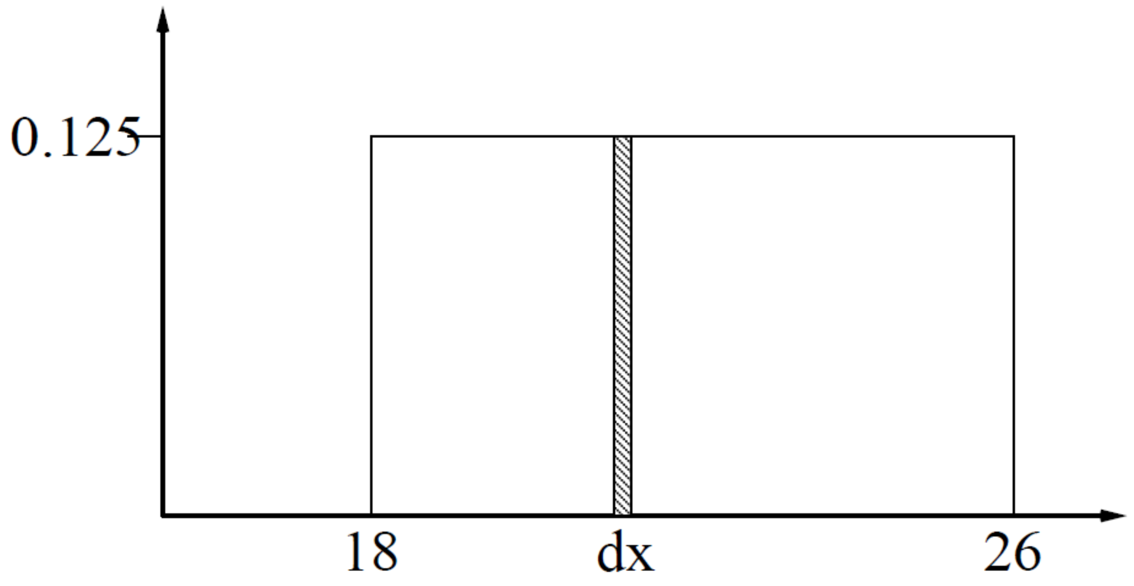
- Probability distributions, in the form of a TABLE, gives values for all possible assignments
- They must sum up to 1
- Note that distributions can be continuous as well!

Continuous Variables

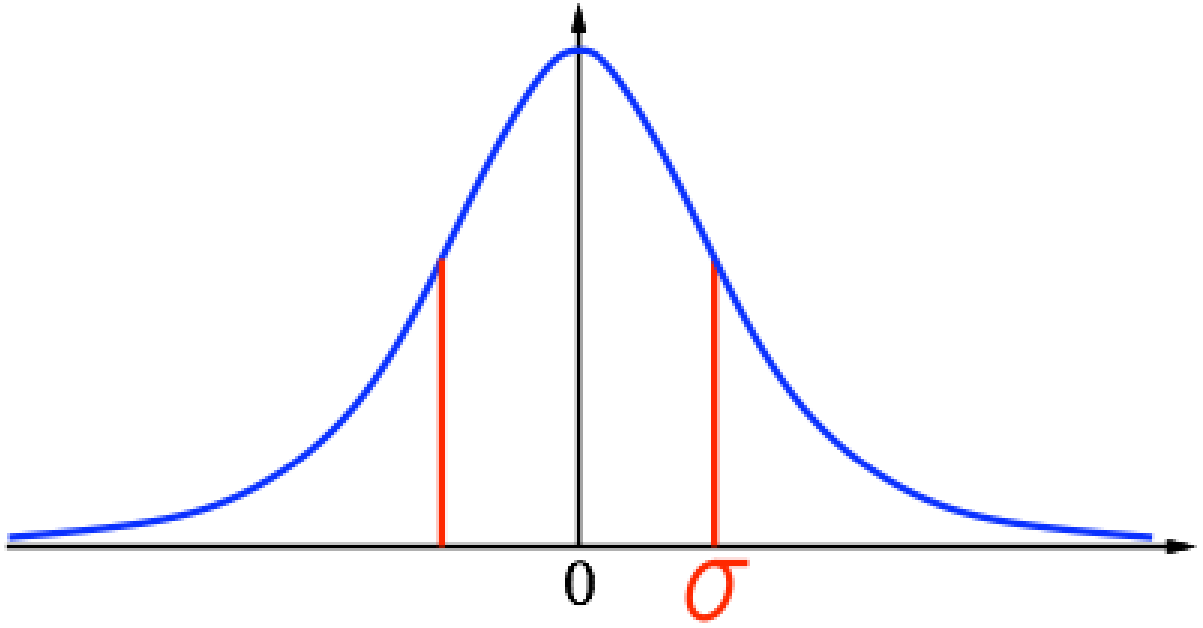
- Express distribution as a parameterized function of value.
- Let f be a probability density function that integrates to 1.

In the remainder of the slides, we are going to deal with discrete variables!

E.g. $f(x) = U[18,26](x)$: Uniform density between 18 and 26



E.g. $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$: Gaussian distribution density function



Joint Probability Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

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

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey: $P(x_1, x_2, \ldots x_n) \geq 0$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Every question about a domain can be answered by the joint distribution because every event is a sum of sample points
- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out!

Exercise: Joint Probabilities and Events

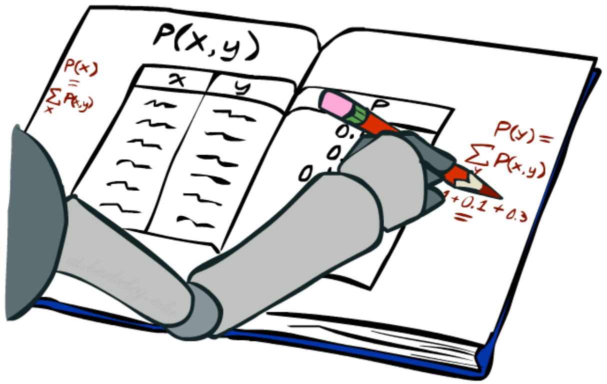
- P(+x, +y) ?
  0.2
- P(+x) ?
  0.2+0.3
- P(-y OR +x) ?
  P(-y ⋃ +x) = P(-y)+P(+x) – P(-y ⋂+x)
  (0.3+0.1)+(0.2+0.3)-0.3=0.6
- P(-y XOR +x) ? (exclusive OR)

$$P(X, Y)$$

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

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding



$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

$$P(s) = \sum_t P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- Conditional or posterior probabilities

  e.g., P(*cavity* | *toothache*) = 0.8

  i.e., given that *toothache* is all I know NOT “if toothache then 80% chance of cavity”

- 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.,

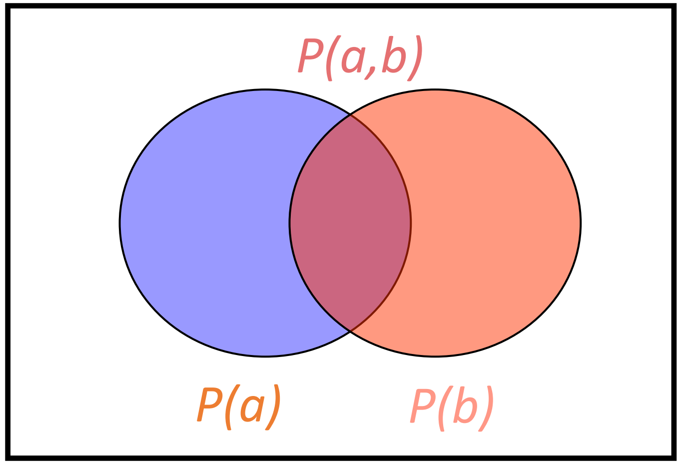
  P(*cavity* | *toothache, sunny*) = P(*cavity* | *toothache*) = 0.8

- This kind of inference, sanctioned by domain knowledge, is crucial

Conditional Probabilities

- Definition of conditional probability: (P(b) ≠ 0)

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



$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$
$$= 0.2 + 0.3 = 0.5$$

Exercise: Conditional Probabilities

$$P(X, Y)$$

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

- P(+x | +y) ?

  0.2/(0.2+0.4)=1/3

- P(-x | +y) ?

  0.4/(0.2+0.4)=2/3

- P(-y | +x) ?

  0.3/(0.2+0.3)=0.6

Conditional Distributions

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

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

Joint Distribution

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Conditional Distribution given T

$$P(W|T)$$

$$P(W|T = hot)$$

W	P
sun	0.8
rain	0.2

$$P(W|T = cold)$$

W	P
sun	0.4
rain	0.6

Conditional Distributions

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\begin{aligned} P(W=s|T=c) &= \frac{P(W=s,T=c)}{P(T=c)} \\ &= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\ &= \frac{0.2}{0.2+0.3} = 0.4 \end{aligned}$$

$$\longrightarrow$$

$$\begin{aligned} P(W=r|T=c) &= \frac{P(W=r,T=c)}{P(T=c)} \\ &= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\ &= \frac{0.3}{0.2+0.3} = 0.6 \end{aligned}$$

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)}$$
$$= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)}$$
$$= \frac{0.2}{0.2+0.3} = 0.4$$

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence

$P(c,W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one)

$P(W|T=c)$

W	P
sun	0.4
rain	0.6

$$P(W=r|T=c) = \frac{P(W=r,T=c)}{P(T=c)}$$
$$= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)}$$
$$= \frac{0.3}{0.2+0.3} = 0.6$$

Normalization Trick

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence

$$P(c,W)$$

T	W	$\bar{P}$
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one)

$$P(W|T=c)$$

W	P
sun	0.4
rain	0.6

$$P(x_1|x_2) = \frac{P(x_1,x_2)}{P(x_2)} = \frac{P(x_1,x_2)}{\sum_{x_1} P(x_1,x_2)}$$

Why does this work? Sum of selection is P(evidence)! (P(T=c), here)

Exercise: Normalization Trick

- P(X | Y=-y) ?

$$P(X, Y)$$

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

SELECT the joint probabilities matching the evidence

X	$\overline{\mathrm{P}}$
+x	0.3
-x	0.1

NORMALIZE the selection (make it sum to one)

X	P
+x	0.75
-x	0.25

Exercise

X	Y	Z	P
+x	+y	+z	0.12
+x	+y	-z	0.18
+x	-y	+z	0.04
+x	-y	-z	0.16
-x	+y	+z	0.18
-x	+y	-z	0.12
-x	-y	+z	0.07
-x	-y	-z	0.13

- P(X | +y, -z)

X	Y	Z	$\overline{P}$	P
+x	+y	-z	0.18	0.6
-x	+y	-z	0.12	0.4

- P(Y, Z | +x)

X	Y	Z	$\overline{P}$	P
+x	+y	+z	0.12	0.24
+x	+y	-z	0.18	0.36
+x	-y	+z	0.04	0.08
+x	-y	-z	0.16	0.32

Inference by Enumeration

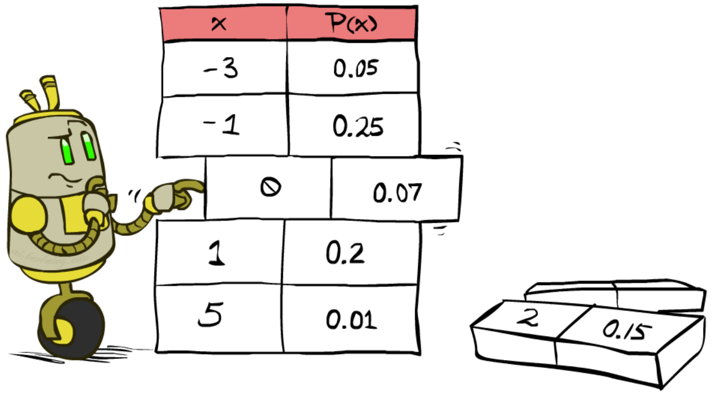
- General case:
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query* variable: Q
  - Hidden variables: $H_1 \dots H_r$

$$\left.\begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array}\right\} \begin{array}{c} X_1, X_2, \dots X_n \\ \textit{All variables} \end{array}$$

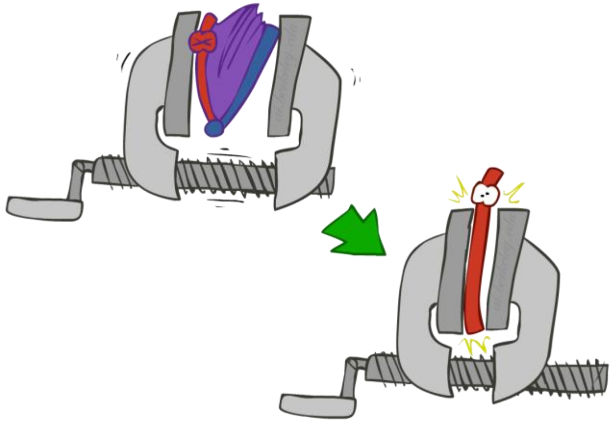
- We want:

$$P(Q|e_1 \dots e_k)$$

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence (marginalize)



$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} \underbrace{P(Q, h_1 \dots h_r, e_1 \dots e_k)}_{X_1, X_2, \dots X_n}$$

- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

* *Works fine with multiple query variables, too*

Inference by Enumeration

- P(W)?
  - Step 1: All of the entries
  - Step 2: Marginalize out S and T (all hidden no evidence)

W	$\overline{\text{P}}$	P
sun	0.3+0.1+0.1+0.15	0.65
rain	0.05+0.05+0.05+0.2	0.35

  - Step 3: Normalize (actually no need for this case)

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter)?
  - Step 1: Rows with S=winter
  - Step 2: Marginalize out T (hidden)
  - Step 3: Normalize

W	$\overline{\text{P}}$	P
sun	0.1+0.15	0.5
rain	0.05+0.2	0.5

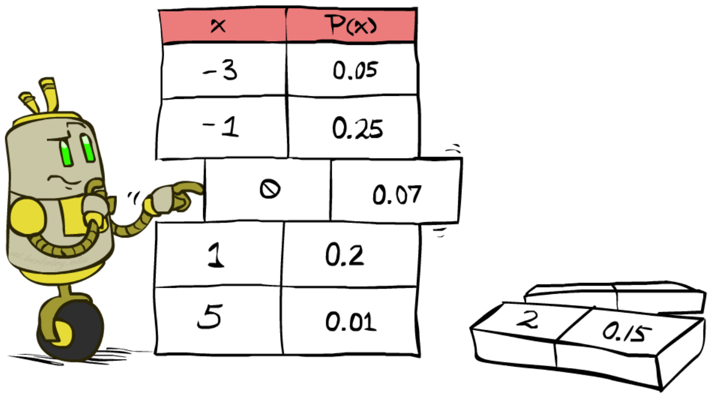
S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

$$P(W = sun \mid S = winter) =$$
$$P(W = sun, T = hot \mid S = winter) + P(W = sun, T = cold \mid S = winter)$$
$$= \alpha(0.1 + 0.15)$$
$$P(W = rain \mid S = winter) =$$
$$P(W = rain, T = hot \mid S = winter) + P(W = rain, T = cold \mid S = winter)$$
$$= \alpha(0.05 + 0.2)$$

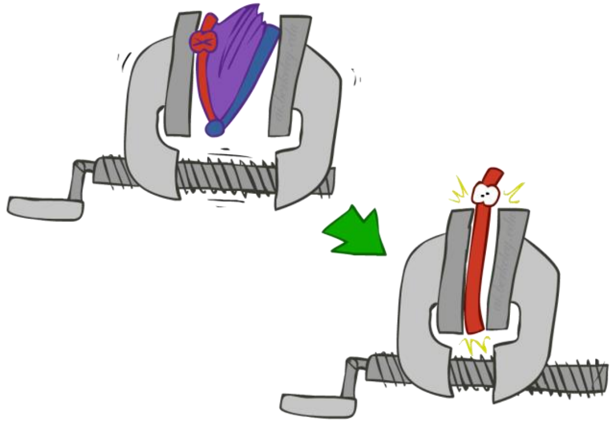
- P(W | winter, hot)?
  - Step 1: Rows with S=winter, T = hot
  - Step 2: No hidden variables
  - Step 3: Normalize

W	$\overline{\text{P}}$	P
sun	0.1	2/3
rain	0.05	1/3

Inference by Enumeration



$\longrightarrow$



$\longrightarrow \quad \times \frac{1}{Z}$

- Obvious problems:
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution

The Product Rule

- Sometimes we have the conditional distributions but we want the joint

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad \Longleftrightarrow \quad P(y)P(x|y) = P(x,y)$$

The Product Rule

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

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

The Product Rule

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

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

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

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

- Chain rule is the product rule applied multiple times, turning a joint probability into conditional probabilities

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

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

- Dividing, we get:

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

- Why is this helpful?
  - Let's us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{given}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

Exercise: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.6
dry	rain	0.4

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

- What is P(W | dry) ?

D	W	$\overline{\text{P}}$	P
dry	sun	0.9*0.8=0.72	0.9
dry	rain	0.4*0.2=0.08	0.1

Probability Summary

- Probability is a rigorous formalism for uncertain knowledge
- *Joint probability distribution* specifies probability of every <u>atomic event</u>

$$P(X, Y)$$

Probability Rules

- Conditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$
- Product rule $$P(x,y) = P(x|y)P(y)$$
- Chain rule $$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1}) \end{aligned}$$
- Bayes rule $$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

Probability Summary

- Probability is a rigorous formalism for uncertain knowledge
- *Joint probability distribution* specifies probability of every atomic event

$$P(X,Y)$$

- Queries can be answered by summing over atomic events

$$P(q,e) = \sum_h P(q,e,H)\,, Z = \sum_q P(Q,e)\,, P(q|e) = P(q,e)/Z$$

- For nontrivial domains, we must find a way to reduce the joint size
- *How?*

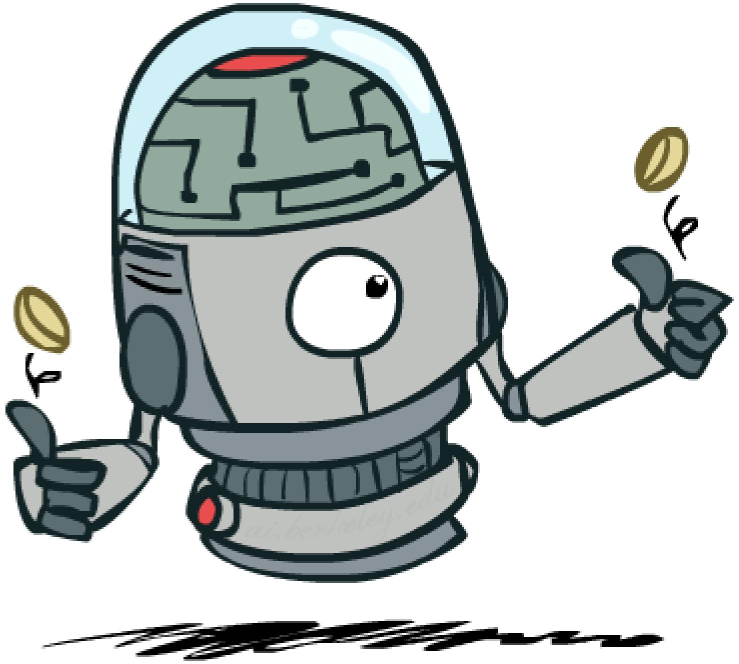
Independence

- Two variables are *independent* in a joint distribution if:

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

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

$$X \perp\!\!\!\perp Y$$



  - Says the joint distribution *factors* into a product of two simple ones
  - Absolute independence is very powerful but not so common

- Can use independence as a *modeling assumption*
  - Independence can be a simplifying assumption
  - *Empirical* joint distributions: at best “close” to independent
  - What could we assume for {Weather, Traffic, Cavity}?

Example: Independence?

$$P(T)$$

T	P
hot	0.5
cold	0.5

$$P_1(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P_2(T,W) = P(T)P(W)$$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$$P(W)$$

W	P
sun	0.6
rain	0.4

Example: Independence

- N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

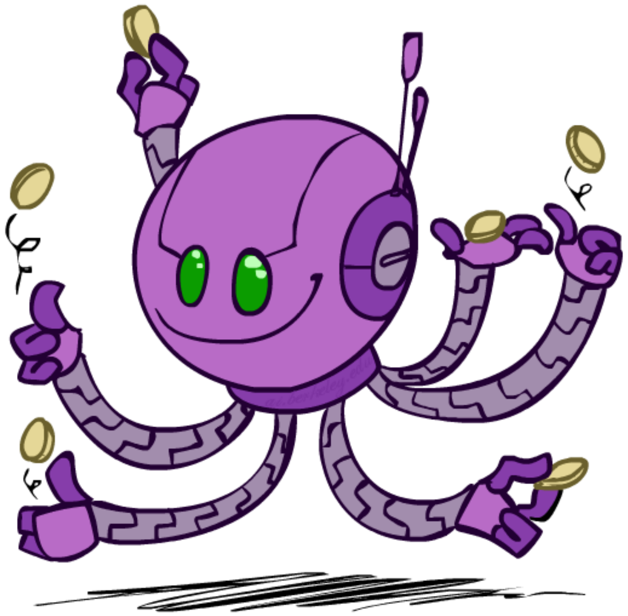
$P(X_2)$

H	0.5
T	0.5

$\cdots$

$P(X_n)$

H	0.5
T	0.5

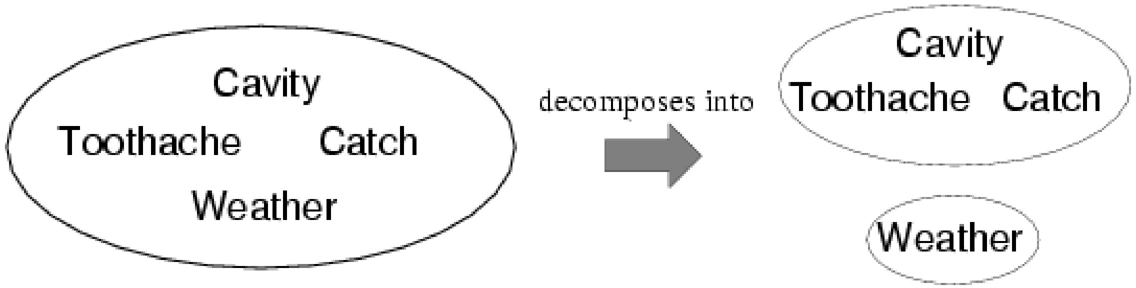


$P(X_1, X_2, \ldots X_n)$

2^n

From $O(2^n) \rightarrow O(n)$

Independence



P(Toothache, Catch, Cavity, Weather)= **P**(Toothache, Catch, Cavity) **P**(Weather)

- 32 entries reduced to 12
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional Independence

- $\mathbf{P}$(Toothache, Cavity, Catch) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches it doesn't depend on whether I have a toothache:
  (1) $\mathbf{P}(catch \mid toothache, cavity) = \mathbf{P}(catch \mid cavity)$
- The same independence holds if I do not have a cavity:
  (2) $\mathbf{P}(catch \mid toothache, \neg cavity) = \mathbf{P}(catch \mid \neg cavity)$
- Catch is conditionally independent of Toothache given Cavity:
  $\mathbf{P}(Catch \mid Toothache, Cavity) = \mathbf{P}(Catch \mid Cavity)$
- Equivalent statements:
  $\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$
  $\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\, \mathbf{P}(Catch \mid Cavity)$
  One can be derived from the other easily

Conditional Independence

- Write out full joint distribution using chain rule:

  P(*Toothache, Catch, Cavity*)

  = **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch, Cavity*)

  = **P**(*Toothache* | *Catch, Cavity*) **P**(*Catch* | *Cavity*) **P**(*Cavity*)

  = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*) **P**(Cavity) (conditional independence)

  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

- Unconditional (absolute) independence is very rare
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y | Z$$

if and only if:

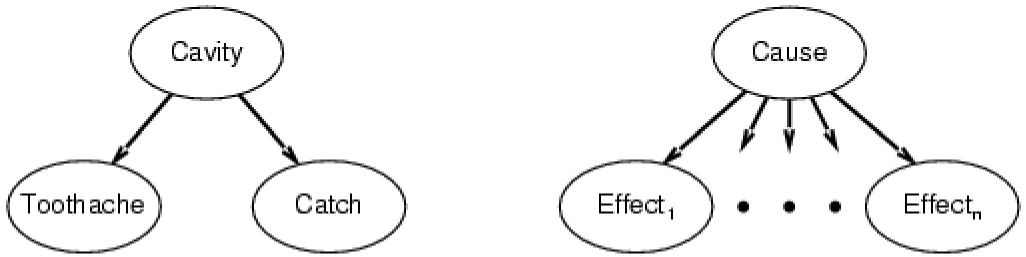
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

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

A Special Case of Conditional Independence

- Let's assume "effects" are conditionally independent given the cause



P(Cavity | toothache ∧ catch)

= α**P**(toothache ∧ catch | Cavity) **P**(Cavity)

= α**P**(toothache | Cavity) **P**(catch | Cavity) **P**(Cavity)

$$P(x|y) = \frac{P(y|x)}{\underbrace{P(y)}_{1/\alpha}} P(x)$$

- Following is an example of a naïve Bayes model:

  P(Cause, $Effect_1$, … , $Effect_n$) = **P**(Cause) $\prod_i$ **P**(Effecti | Cause)

- Total number of parameters is linear in n
- Note that not all models are like this

Probability Recap

- Conditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$
- Product rule $$P(x,y) = P(x|y)P(y)$$
- Chain rule $$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1}) \end{aligned}$$
- X, Y independent if and only if: $\forall x,y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y | Z$

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

Probability Summary

- Probability is a rigorous formalism for uncertain knowledge
- *Joint probability distribution* specifies probability of every atomic event
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size
- *Independence* and *conditional* independence provide the tools