

# **CMSC 671**

## **Fall 2010**

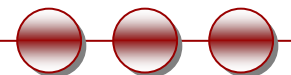
**Thu 10/28/10**

### **Quantifying Uncertainty**

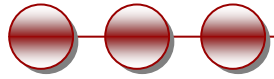
### **Bayes' Rule**

### **Chapter 13.1-13.5**

**Prof. Laura Zavala, [laura.zavala@umbc.edu](mailto:laura.zavala@umbc.edu), ITE 373, 410-455-8775**



# Sources of uncertainty



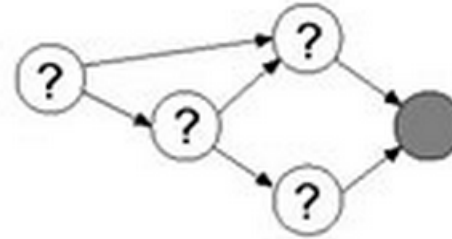
- Uncertain **inputs**
  - Missing data
  - Noisy data
- Uncertain **knowledge**
  - Multiple causes lead to multiple effects
  - Incomplete enumeration of conditions or effects
  - Incomplete knowledge of causality in the domain
  - Probabilistic/stochastic effects
- Uncertain **outputs**
  - Abduction and induction are inherently uncertain
  - Default reasoning, even in deductive fashion, is uncertain
  - Incomplete deductive inference may be uncertain
- ▶ Probabilistic reasoning only gives probabilistic results (summarizes uncertainty from various sources)



# Uncertainty and Artificial Intelligence

- Probabilistic Databases

- traditional DB technology cannot answer queries about items that were never loaded into the dataset
- UAI models are like probabilistic databases



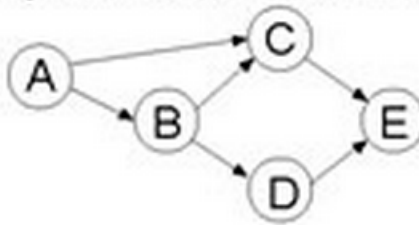
- Automatic System Building

- old expert systems needed hand coding of knowledge and of output semantics
- learning automatically constructs rules and supports all types of queries

# Uncertainty and Artificial Intelligence

## Decision making with uncertainty

- Probabilistic methods can be used to:
  - make decisions given partial information about the world
  - account for noisy sensors or actuators
  - explain phenomena not part of our models
  - describe inherently stochastic behaviour in the world



- Example: you live in California with your spouse and two kids. You listen to the radio on your drive home, and when you arrive you find your burglar alarm ringing.  
Do you think your house was broken into?

# Decision making with uncertainty

- **Rational** behavior:
  - For each possible action, identify the possible outcomes
  - Compute the **probability** of each outcome
  - Compute the **utility** of each outcome
  - Compute the probability-weighted (**expected**) **utility** over possible outcomes for each action
  - Select the action with the highest expected utility (principle of **Maximum Expected Utility**)

# Decision making with uncertainty

- **Rational** behavior:
  - For each possible action, identify the possible outcomes
  - Compute the **probability** of each outcome
  - Compute the **utility** of each outcome
  - Compute the probability-weighted (**expected**) **utility** over possible outcomes for each action
  - Select the action with the highest expected utility (principle of **Maximum Expected Utility**)

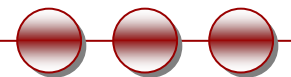
# Bayesian Reasoning

- Probability theory
- Probabilistic approach to inference
- Bayesian inference
  - Use probability theory and information about independence
  - Reason diagnostically (from evidence (effects) to conclusions (causes)) or causally (from causes to effects)
- Bayesian networks
  - Compact representation of probability distribution over a set of propositional random variables
  - Take advantage of independence relationships

# Other Uncertainty Representations

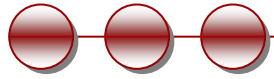


- Default reasoning
  - Nonmonotonic logic: Allow the retraction of default beliefs if they prove to be false
- Rule-based methods
  - Certainty factors (Mycin): propagate simple models of belief through causal or diagnostic rules
- Evidential reasoning
  - Dempster-Shafer theory:  $\text{Bel}(P)$  is a measure of the evidence for  $P$ ;  $\text{Bel}(\neg P)$  is a measure of the evidence against  $P$ ; together they define a belief interval (lower and upper bounds on confidence)
- Fuzzy reasoning
  - Fuzzy sets: How well does an object satisfy a vague property?
  - Fuzzy logic: “How true” is a logical statement?

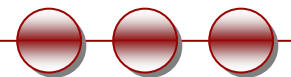




# Abduction



- **Abduction** is a reasoning process that tries to form plausible explanations for abnormal observations
  - Abduction is distinctly different from deduction and induction
  - Abduction is inherently uncertain
- Uncertainty is an important issue in abductive reasoning

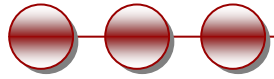


# Abduction examples

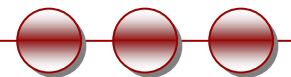
---

- Medical diagnosis
  - Facts: symptoms, lab test results, and other observed findings (called manifestations)
  - KB: causal associations between diseases and manifestations
  - Reasoning: one or more diseases whose presence would causally explain the occurrence of the given manifestations
- Many other reasoning processes (e.g., word sense disambiguation in natural language process, image understanding, criminal investigation) can also be seen as abductive reasoning

# Abduction



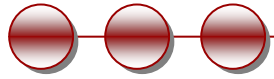
- “Conclusions” are **hypotheses**, not theorems (may be false *even if* rules and facts are true)
  - E.g., misdiagnosis in medicine
- There may be multiple plausible hypotheses
  - Given rules  $A \Rightarrow B$  and  $C \Rightarrow B$ , and fact B, both A and C are plausible hypotheses
  - Hypotheses can be ranked by their plausibility (if it can be determined)



# Abductive Reasoning

- Reasoning is **non-monotonic**
  - The plausibility of hypotheses can increase/decrease as new facts are collected
  - In contrast, deductive inference is **monotonic**: it never change a sentence's truth value, once known
  - In abductive (and inductive) reasoning, some hypotheses may be discarded, and new ones formed, when new observations are made

# Comparing Abduction, Deduction, and Induction



**Deduction:** major premise: All balls in the box are black  
 minor premise: These balls are from the box  
 conclusion: These balls are black

$A \Rightarrow B$ $A$ ----- $B$
--

**Abduction:** rule: All balls in the box are black  
 observation: These balls are black  
 explanation: These balls are from the box

$A \Rightarrow B$ $B$ ----- <b>Possibly A</b>
--

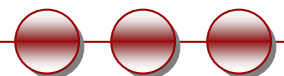
**Induction:** case: These balls are from the box  
 observation: These balls are black  
 hypothesized rule: All ball in the box are black

<b>Whenever</b> <b>A then B</b> ----- <b>Possibly</b> $A \Rightarrow B$
---

**Deduction** reasons from causes to effects

**Abduction** reasons from effects to causes

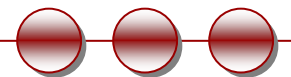
**Induction** reasons from specific cases to general rules



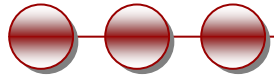
# Uncertainty Tradeoffs



- **Bayesian networks:** Nice theoretical properties combined with efficient reasoning make BNs very popular; limited expressiveness, knowledge engineering challenges may limit uses. Also, they require initial knowledge of many probabilities.
- **Nonmonotonic logic:** Represent commonsense reasoning, but can be computationally very expensive
- **Certainty factors:** Not semantically well founded
- **Dempster-Shafer theory:** Has nice formal properties, but can be computationally expensive, and intervals tend to grow towards  $[0,1]$  (not a very useful conclusion)
- **Fuzzy reasoning:** Semantics are unclear (fuzzy!), but has proved very useful for commercial applications

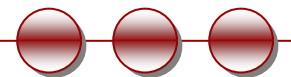


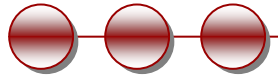
# Ontology and epistemology



- **Ontological commitment** – what the language assumes about the nature of reality
- **Epistemological commitment** – the possible states of knowledge

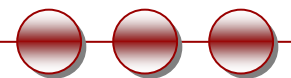
Language	Ontological Commitment (What exists in the world)	Epistemological Commitment (What an agent believes about facts)
Propositional logic	facts	true/false/unknown
First-order logic	facts, objects, relations	true/false/unknown
Temporal logic	facts, objects, relations, times	true/false/unknown
Probability theory	facts	degree of belief 0...1
Fuzzy logic	degree of truth	degree of belief 0...1





# Quantifying uncertainty

## Chapter 13



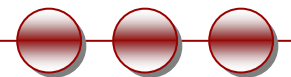
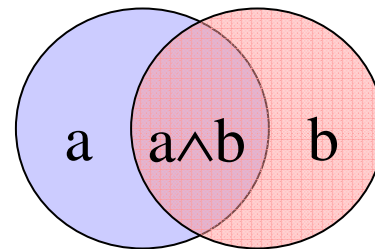
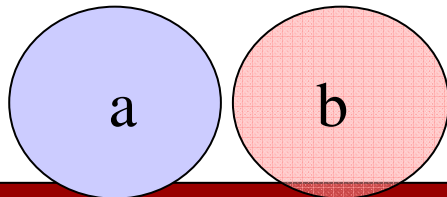


# Why probabilities anyway?

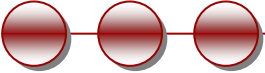
- Kolmogorov showed that three simple axioms lead to the rules of probability theory
  - De Finetti, Cox, and Carnap have also provided compelling arguments for these axioms
- 1. All probabilities are between 0 and 1:
  - $0 \leq P(a) \leq 1$
- 2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0:
  - $P(\text{true}) = 1$  ;  $P(\text{false}) = 0$
- 3. The probability of a disjunction is given by:
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

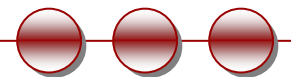
If  $a$  and  $b$  are disjoint, then

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



# Probabilities

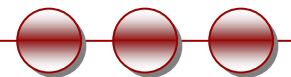
- 
- *The probability that the patient has a cavity, given that she has toothache, is 0.8*
  - *Does the patient have a cavity or not?*
    - *The patient either has a cavity or doesn't*
    - *The patient does not have 0.8 cavity*
  - Probabilities statements are made with respect to a knowledge state not with respect to the real world
  - Similar case for probabilistic predictions
    - *This pneumonia patient has a 93% chance of complete recovery*
      - *The patient either recovers or doesn't*



# Example: disjoint events

- Consider a deck of 52 cards.
- The event A that I will draw a spade and the event B that I will draw a king are clearly not disjoint events.
- Their intersection specifies the event that I will draw the king of spades,  $A \cap B = \{\text{king of spades}\}$ .
- Thus, the probability that I will draw either a king or a spade is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$



# More probabilities

- Probability distribution

- $P(\text{Weather} = \text{sunny}) = 0.6$

$$P(\text{sunny}) = 0.6$$

- $P(\text{Weather} = \text{rain}) = 0.1$

$$P(\text{rain}) = 0.1$$

- $P(\text{Weather} = \text{cloudy}) = 0.29$

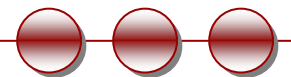
$$P(\text{cloudy}) = 0.29$$

- $P(\text{Weather} = \text{snow}) = 0.01$

$$P(\text{snow}) = 0.01$$

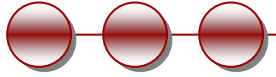
- PDFs (probability density functions) for continuous variables

- $P(\text{NoonTemp} = x) = \text{Uniform}_{[18C, 26C]}(x)$

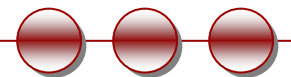


# Joint Probabilities

---



- What does  $P(\text{Alarm}, \text{Burglary})$  mean?



# Joint Probabilities

- In addition to distributions on single variables, we need notation for distributions on multiple variables
- Joint Probability Distribution
  - $P(\textit{Alarm}, \textit{Burglary})$  denotes the probabilities of all the combinations of the values of *Alarm* and *Burglary*
  - $P(\textit{alarm=false}, \textit{Burglary})$

# Full Joint Probability Distribution

- $P(\text{Cavity, Toothache, Weather})$ 
  - 2 x 2 x 4 table (16 entries)
- Every proposition probability is a sum over possible worlds
- A full joint distribution suffices for calculating the probability of any proposition
- Notation:
  - $P(X = x)$  is the probability that random variable  $X$  takes on value  $x$
  - $P(X)$  is the *distribution* of probabilities for all possible values of  $X$

# Probability theory

- **Random variables**

- Domain

- **Atomic event**: complete specification of state

- **Prior probability**: degree of belief without any other evidence

- **Joint probability distribution**: matrix of combined probabilities of a set of variables

- Alarm, Burglary, Earthquake

- Boolean (like these), discrete, continuous

- Alarm=True  $\wedge$  Burglary=True  $\wedge$  Earthquake=False  
alarm  $\wedge$  burglary  $\wedge$   $\neg$ earthquake

- P(burglary) = .1

- P(Alarm, Burglary) =

	alarm	$\neg$ alarm
burglary	.09	.01
$\neg$ burglary	.1	.8



# Conditional Probabilities

- Unconditional or prior probabilities
  - Degree of belief in the absence of any other information
    - $P(\text{burglary}) = 0.1$
    - $P(\text{cavity}) = 0.2$
    - $P(\text{doubles}) = 6/36 = 1/6$
    - $P(\text{double-five}) = 1/36$
- **Conditional or Posterior Probability**
  - Most of the time, we have *some* information, usually called evidence that has already been revealed
  - The probability of some event A, given the occurrence of some other event B
    - $P(\text{cavity} \mid \text{toothache})$
    - $P(\text{burglary} \mid \text{alarm})$
    - $P(\text{doubles} \mid \text{Die1} = 5)$

# Computing Conditional Probabilities

(1)

$$P(a | b) = P(a \wedge b) / P(b)$$

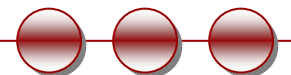
- **Conditional or Posterior Probability**

- Most of the time, we have *some* information, usually called evidence that has already been revealed
- The probability of some event A, given the occurrence of some other event B

- $P(\text{doubles} | \text{Die1} = 5) =$   
 $= P(\text{doubles} \wedge \text{Die1}=5) / P(\text{Die1}=5)$   
 $= (1/36) / (1/6) = 1/6$

- **Product rule**

- $P(a \wedge b) = P(a | b) P(b)$



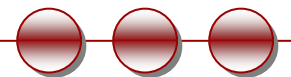
# Computing Conditional Probabilities

(2)

- **Conditional probability:**  
probability of effect given causes
- **Computing conditional probs:**
  - $P(a | b) = P(a \wedge b) / P(b)$
  - $P(b)$ : **normalizing** constant
- **Product rule:**
  - $P(a \wedge b) = P(a | b) P(b)$
- **Marginalizing:**
  - $P(B) = \sum_a P(B, a)$
  - $P(B) = \sum_a P(B | a) P(a)$   
(**conditioning**)
- $P(\text{burglary} | \text{alarm}) =$   
 $P(\text{alarm} | \text{burglary}) =$
- $P(\text{burglary} | \text{alarm}) =$   
 $P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm}) =$
- $P(\text{burglary} \wedge \text{alarm}) =$   
 $P(\text{burglary} | \text{alarm}) P(\text{alarm}) =$
- $P(\text{alarm}) =$   
 $P(\text{alarm} \wedge \text{burglary}) +$   
 $P(\text{alarm} \wedge \neg \text{burglary}) =$

$P(\text{Alarm}, \text{Burglary}) =$

	alarm	$\neg$ alarm
burglary	.09	.01
$\neg$ burglary	.1	.8



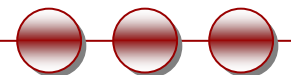
# Computing Conditional Probabilities

(3)

- **Conditional probability:**  
probability of effect given causes
- **Computing conditional probs:**
  - $P(a | b) = P(a \wedge b) / P(b)$
  - $P(b)$ : **normalizing** constant
- **Product rule:**
  - $P(a \wedge b) = P(a | b) P(b)$
- **Marginalizing or Summing out:**
  - $P(B) = \sum_a P(B, a)$
  - $P(B) = \sum_a P(B | a) P(a)$   
(**conditioning**)
- $P(\text{burglary} | \text{alarm}) = .47$   
 $P(\text{alarm} | \text{burglary}) = .9$
- $P(\text{burglary} | \text{alarm}) =$   
 $P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm})$   
 $= .09 / .19 = .47$
- $P(\text{burglary} \wedge \text{alarm}) =$   
 $P(\text{burglary} | \text{alarm}) P(\text{alarm}) =$   
 $.47 * .19 = .09$
- $P(\text{alarm}) =$   
 $P(\text{alarm} \wedge \text{burglary}) +$   
 $P(\text{alarm} \wedge \neg \text{burglary}) =$   
 $.09 + .1 = .19$

$P(\text{Alarm}, \text{Burglary}) =$

	alarm	$\neg$ alarm
burglary	.09	.01
$\neg$ burglary	.1	.8



# Inference from the Joint

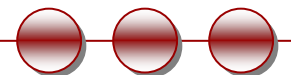
- Simple method for probabilistic inference
- Computation of posterior probabilities for queries given the observed evidence and the full joint distribution
- A full joint distribution suffices for calculating the probability of any proposition

$P(\text{Alarm}, \text{Burglary})$

	alarm	$\neg$ alarm
burglary	.09	.01
$\neg$ burglary	.1	.8

Observed: *alarm*

Query:  $P(\text{burglary} \mid \text{alarm})$



# Example: Inference from the joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

$P(\text{burglary} \mid \text{alarm}) =$

$P(\neg\text{burglary} \mid \text{alarm}) =$

# Example: Inference from the joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

$$\begin{aligned} P(\text{burglary} \mid \text{alarm}) &= .09 / .19 \\ &= .474 \end{aligned}$$

$$\begin{aligned} P(\neg\text{burglary} \mid \text{alarm}) &= .1 / .19 \\ &= .526 \end{aligned}$$

# Example: Inference from the joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

$$\begin{aligned} P(\text{burglary} \mid \text{alarm}) &= .09 / .19 \\ &= .474 \end{aligned}$$

$$\begin{aligned} P(\neg\text{burglary} \mid \text{alarm}) &= .1 / .19 \\ &= .526 \end{aligned}$$



# Example: Inference from the joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

$$P(\text{burglary} \mid \text{alarm}) = .474$$

$$P(\neg\text{burglary} \mid \text{alarm}) = .526$$

$$P(\text{Burglary} \mid \text{alarm}) = \alpha P(\text{Burglary}, \text{alarm})$$

$$= \alpha [P(\text{Burglary}, \text{alarm}, \text{earthquake}) + P(\text{Burglary}, \text{alarm}, \neg\text{earthquake})]$$

$$= \alpha [ (.01, .01) + (.08, .09) ]$$

$$= \alpha [ (.09, .1) ]$$

$$= \langle .474, .526 \rangle$$

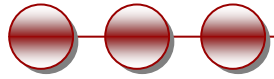
$$\alpha = 1/((.09+.1) = 5.26 \text{ (since } P(\text{burglary} \mid \text{alarm}) + P(\neg\text{burglary} \mid \text{alarm}) = 1)$$

# Exercise: Inference from the joint

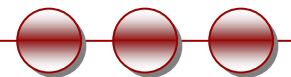
$p(\text{smart} \wedge \text{study} \wedge \text{prep})$	smart		$\neg$ smart	
	study	$\neg$ study	study	$\neg$ study
prepared	.432	.16	.084	.008
$\neg$ prepared	.048	.16	.036	.072

- Queries:
  - What is the prior probability of *smart*?
  - What is the prior probability of *study*?
  - What is the conditional probability of *prepared*, given *study* and *smart*?
- Save these answers for next time! 😊

# Inference Using Full Joint Distributions



- A full joint distribution suffices for calculating the probability of any proposition
- Full joint distribution is not practical for building reasoning systems
  - Suppose we have a joint *distribution*  $P(X_1, X_2, \dots, X_n)$  of  $n$  random variables with domain sizes  $d$
  - What is the size of the probability table?
  - Impossible to write out completely for all but the smallest distributions
- What can be done?



# Independence

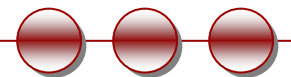
- When two sets of propositions do not affect each others' probabilities, we call them **independent**, and can easily compute their joint and conditional probability:

Independent (A, B)  $\rightarrow$   $\mathbf{P(A \wedge B) = P(A) P(B)}$ ,  $\mathbf{P(A | B) = P(A)}$

# Back in slide 23 ...

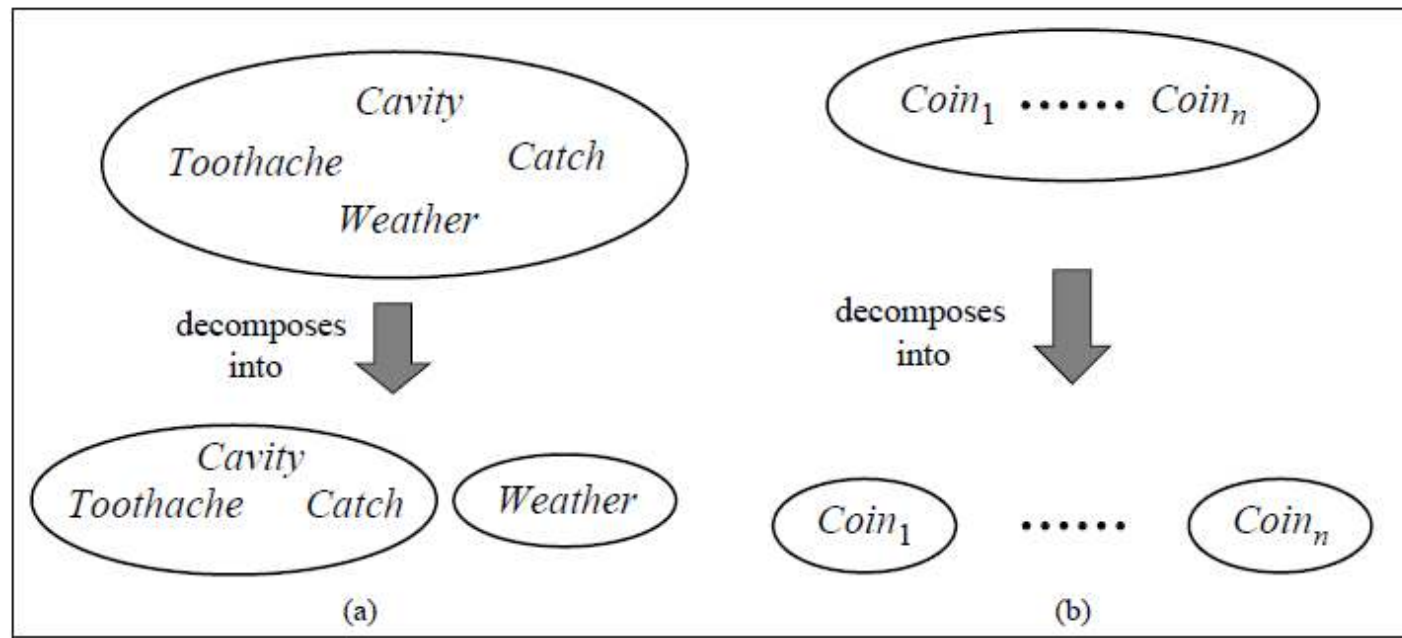
- P(Cavity, Toothache, Weather)
  - 2 x 2 x 4 table (16 entries)
- Dental problems do not affect the weather
- The weather does not influence dental variables
- $P(\text{sunny} \mid \text{tootache, cavity}) = P(\text{sunny})$
- $P(\text{tootache, cavity, sunny}) = P(\text{sunny}) P(\text{tootache, cavity})$

Independent (A, B)  $\rightarrow P(A \wedge B) = P(A) P(B)$ ,  $P(A \mid B) = P(A)$



# Independence Example

- Independence assertions are usually based on knowledge of the domain
- They can dramatically reduce the amount of information necessary to specify the full joint distribution



# Independence Example 2

- {moon-phase, light-level} might be independent of {burglary, alarm, earthquake}
  - Then again, it might not: Burglars might be more likely to burglarize houses when there's a new moon (and hence little light)
  - But if we know the light level, the moon phase doesn't affect whether we are burglarized
  - Once we're burglarized, light level doesn't affect whether the alarm goes off
- We need a more complex notion of independence, and methods for reasoning about these kinds of relationships
  - Absolute Independence vs Conditional Independence

# Exercise: Independence

$p(\text{smart} \wedge \text{study} \wedge \text{prep})$	smart		$\neg$ smart	
	study	$\neg$ study	study	$\neg$ study
prepared	.432	.16	.084	.008
$\neg$ prepared	.048	.16	.036	.072

- Queries:
  - Is *smart* independent of *study*?
  - Is *prepared* independent of *study*?



# Exercise: Conditional independence

$p(\text{smart} \wedge \text{study} \wedge \text{prep})$	smart		$\neg$ smart	
	study	$\neg$ study	study	$\neg$ study
prepared	.432	.16	.084	.008
$\neg$ prepared	.048	.16	.036	.072

- Queries:
  - Is *smart* conditionally independent of *prepared*, given *study*?
  - Is *study* conditionally independent of *prepared*, given *smart*?

# Conditional independence

- Absolute independence:
  - A and B are **independent** if  $P(A \wedge B) = P(A) P(B)$ ; equivalently,  $P(A) = P(A | B)$  and  $P(B) = P(B | A)$
- A and B are **conditionally independent** given C if
  - $P(A \wedge B | C) = P(A | C) P(B | C)$
- This lets us decompose the joint distribution:
  - $P(A \wedge B \wedge C) = P(A | C) P(B | C) P(C)$
- Moon-Phase and Burglary are **conditionally independent given** Light-Level
- Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint probability distribution

# Bayes's rule

- Bayes's rule is derived from the product rule (slide 26):
  - $P(a \wedge b) = P(a | b) P(b)$        $P(a \wedge b) = P(b | a) P(a)$
- We also had a formula for computing probabilities (slide 26):
  - $P(a | b) = P(a \wedge b) / P(b)$

- Using the product rule, we derive:

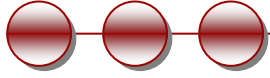
$$P(b | a) = \frac{P(a | b)P(b)}{P(a)}$$

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

**Baye's Rule**

- Often useful for diagnosis:
  - If X are (observed) effects and Y are (hidden) causes,
  - We may have a model for how causes lead to effects ( $P(X | Y)$ )
  - We may also have prior beliefs (based on experience) about the frequency of occurrence of effects ( $P(Y)$ )
  - Which allows us to reason from effects to causes ( $P(Y | X)$ ) (diagnosis)

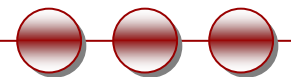
# Bayes's rule


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

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

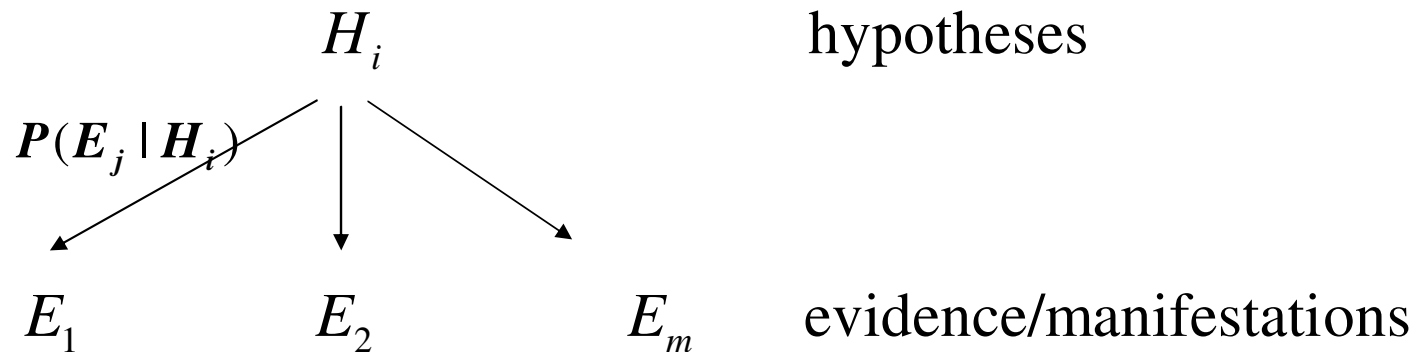
$$P(\text{cause} | \text{effect}) = \frac{P(\text{effect} | \text{cause})P(\text{cause})}{P(\text{effect})}$$

$$P(\text{disease} | \text{symptoms}) = \frac{P(\text{symptoms} | \text{disease})P(\text{disease})}{P(\text{symptoms})}$$



# Bayesian inference

- In the setting of diagnostic/evidential reasoning



– Known prior probability of hypothesis

$$P(H_i)$$

conditional probability

$$P(E_j | H_i)$$

– Want to compute the *posterior probability*

$$P(H_i | E_j)$$

- Bayes' theorem (formula 1):

$$P(H_i | E_j) = \frac{P(E_j | H_i)P(H_i)}{P(E_j)}$$

# Simple Bayesian diagnostic reasoning

- Knowledge base:
  - Evidence / manifestations:  $E_1, \dots, E_m$
  - Hypotheses / disorders:  $H_1, \dots, H_n$ 
    - $E_j$  and  $H_i$  are **binary**; hypotheses are **mutually exclusive** (non-overlapping) and **exhaustive** (cover all possible cases)
  - Conditional probabilities:  $P(E_j | H_i), i = 1, \dots, n; j = 1, \dots, m$
- Cases (evidence for a particular instance):  $E_1, \dots, E_l$
- Goal: Find the hypothesis  $H_i$  with the highest posterior (i.e. the MAP hypothesis)
  - $\text{Max}_i P(H_i | E_1, \dots, E_l)$

# Bayesian diagnostic reasoning II

- Bayes' rule says that
  - $P(H_i | E_1, \dots, E_l) = P(E_1, \dots, E_l | H_i) P(H_i) / P(E_1, \dots, E_l)$
- Assume each piece of evidence  $E_i$  is conditionally independent of the others, *given* a hypothesis  $H_i$ , then:
  - $P(E_1, \dots, E_l | H_i) = \prod_{j=1}^l P(E_j | H_i)$
- The full joint distribution can be written as:
  - $P(H_i | E_1, \dots, E_l) = \alpha P(H_i) \prod_{j=1}^l P(E_j | H_i)$
- Such a probability distribution is known as **Naïve bayes** model

# Naïve Bayes

$$P(\text{Cause} \mid \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i \mid \text{Cause})$$

- Often used in cases where the “effect” variables are **not** actually conditionally independent given the cause variable, and so the name naïve (a simplifying assumption).
- Works surprisingly well in practice (even if the *conditional independence assumption* does not hold).
  - Naïve Bayes classifiers have shown comparable performance to neural networks and decision tree classifiers.
    - Classifying natural language text documents



# Naïve Bayes Model Example: Spam Filtering

- **Bayesian decision theory:** to minimize the probability of error, we should classify a message as spam if  $P(\text{spam} \mid \text{message}) > P(\neg\text{spam} \mid \text{message})$ 
  - *Maximum a posteriori (MAP)* decision

- We have

$$P(\text{spam} \mid \text{message}) = \frac{P(\text{message} \mid \text{spam})P(\text{spam})}{P(\text{message})} \quad \text{and}$$

$$P(\neg\text{spam} \mid \text{message}) = \frac{P(\text{message} \mid \neg\text{spam})P(\neg\text{spam})}{P(\text{message})}$$

- Notice that  $P(\text{message})$  is just a constant normalizing factor and doesn't affect the decision
- Therefore, all we need is to find  $P(\text{message} \mid \text{spam}) P(\text{spam})$  and  $P(\text{message} \mid \neg\text{spam}) P(\neg\text{spam})$

# Naïve Bayes Model Example: Spam Filtering

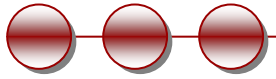
- We need to find  $P(\text{message} \mid \text{spam}) P(\text{spam})$  and  $P(\text{message} \mid \neg\text{spam}) P(\neg\text{spam})$
- The message is a sequence of words  $(w_1, \dots, w_n)$
- **Bag of words** representation
  - The order of the words in the message is not important
  - Each word is conditionally independent of the others given message class (spam or not spam)

$$P(\text{message} \mid \text{spam}) = P(w_1, \dots, w_n \mid \text{spam}) = \prod_{i=1}^n P(w_i \mid \text{spam})$$

- Our filter will classify the message as spam if

$$P(\text{spam}) \prod_{i=1}^n P(w_i \mid \text{spam}) > P(\neg\text{spam}) \prod_{i=1}^n P(w_i \mid \neg\text{spam})$$

# Naïve Bayes Model Example: Spam Filtering



$$P(\textit{spam} \mid w_1, \dots, w_n) = P(\textit{spam}) \prod_{i=1}^n P(w_i \mid \textit{spam})$$

Diagram illustrating the components of the Naïve Bayes model equation:

- $P(\textit{spam} \mid w_1, \dots, w_n)$  is labeled as the **posterior**.
- $P(\textit{spam})$  is labeled as the **prior**.
- $\prod_{i=1}^n P(w_i \mid \textit{spam})$  is labeled as the **likelihood**.



# Limitations of simple Bayesian inference

- Cannot easily handle multi-fault situation, nor cases where intermediate (hidden) causes exist:
  - Disease D causes syndrome S, which causes correlated manifestations  $M_1$  and  $M_2$
- Consider a composite hypothesis  $H_1 \wedge H_2$ , where  $H_1$  and  $H_2$  are independent. What is the relative posterior?
  - $$\begin{aligned} P(H_1 \wedge H_2 | E_1, \dots, E_n) &= \alpha P(E_1, \dots, E_n | H_1 \wedge H_2) P(H_1 \wedge H_2) \\ &= \alpha P(E_1, \dots, E_n | H_1 \wedge H_2) P(H_1) P(H_2) \\ &= \alpha \prod_{j=1}^n P(E_j | H_1 \wedge H_2) P(H_1) P(H_2) \end{aligned}$$
- How do we compute  $P(E_j | H_1 \wedge H_2)$  ??

# Limitations of simple Bayesian inference II

- Assume  $H_1$  and  $H_2$  are independent, given  $E_1, \dots, E_l$ ?
  - $P(H_1 \wedge H_2 \mid E_1, \dots, E_l) = P(H_1 \mid E_1, \dots, E_l) P(H_2 \mid E_1, \dots, E_l)$
- This is a very unreasonable assumption
  - Earthquake and Burglar are independent, but *not* given Alarm:
    - $P(\text{burglar} \mid \text{alarm, earthquake}) \ll P(\text{burglar} \mid \text{alarm})$
- Another limitation is that simple application of Bayes's rule doesn't allow us to handle causal chaining:
  - A: this year's weather; B: cotton production; C: next year's cotton price
  - A influences C indirectly:  $A \rightarrow B \rightarrow C$
  - $P(C \mid B, A) = P(C \mid B)$
- Need a richer representation to model interacting hypotheses, conditional independence, and causal chaining
- Next time: conditional independence and Bayesian networks