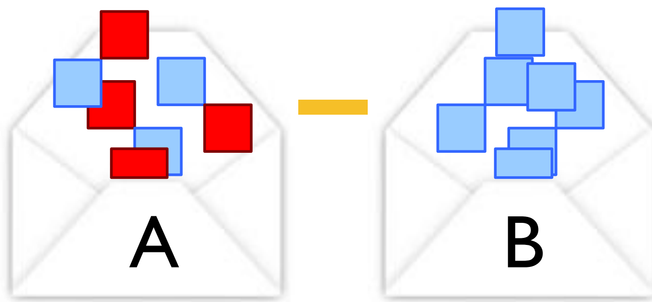
Probabilistic Reasoning

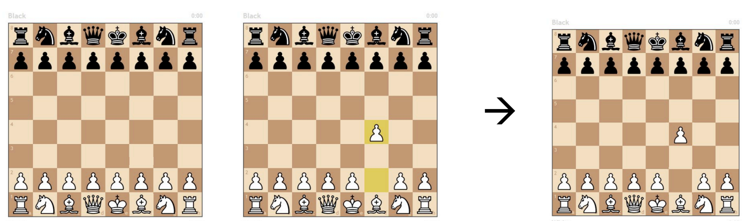


Some material also adapted from www.csc.calpoly.edu/~fkurfess/Courses/CSC-481/W02/Slides/Uncertainty.ppt

1

Probabilistic Reasoning

- So far, mostly, we've done deterministic problems.



- This is a stepping stone to stochastic problem-solving.
- We'll use many of the same techniques and core ideas!
  - Like minimax → expectiminimax

Images: www.instructables.com/id/How-to-Win-a-Chess-Game-in-2-Moves/

4

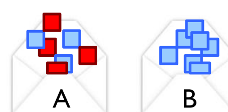
Probabilistic Reasoning

- We don’t (can’t!) know everything about most problems.
- Most problems are not:
  - Deterministic
  - Fully observable
- Or, we can’t calculate everything.
  - Continuous problem spaces
- Probability lets us understand, quantify, and work with this uncertainty.

5

Probability

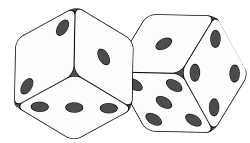
- **World:** The complete set of possible states
- **Random variables**: Problem aspects that take a value
  - “The number of blue squares we have pulled,” B
  - “The combined value of two dice we rolled,” C



- **Event:** Something that happens
- **Sample Space:** All the things (outcomes) that could happen in some set of circumstances
  - Pull 2 squares from envelope A: what is the sample space?
  - How about envelope B?
- **World, redux:** *A complete assignment of values to variables*

6

Basic Probability



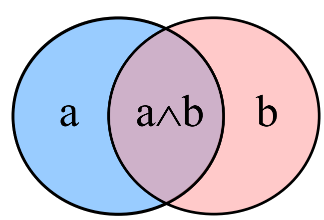
- Each P is a non-negative value in [0,1]
  - P({1,1}) = 1/36
- Total probability of the sample space is 1
  - P({1,1}) + P({1,2}) + P({1,3}) + … + P({6,6}) = 1
- For mutually exclusive events, the probability for at least one of them is the sum of their individual probabilities
  - P(sunny) ∨ P(cloudy) = P(sunny) + P(cloudy)
- Experimental probability: Based on frequency of past events
- Subjective probability: Based on expert assessment

commons.wikimedia.org/wiki/File:2-Dice-Icon.svg

7

Why Probabilities Anyway?

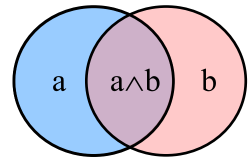
- 3 simple axioms → all rules of probability theory*
- All probabilities are between 0 and 1.
  - $0 \leq P(a) \leq 1$
- Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0.
  - $P(true) = 1$
  - $P(false) = 0$
- The probability of a disjunction is:
  - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



**Kolmogorov – en.wikipedia.org/wiki/Andrey_Kolmogorov*
De Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

8

Compound Probabilities



- Describe independent events
  - Do not affect each other in any way
- Joint probability of two independent events A and B

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

- Union probability of two independent events A and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$
$$= P(A) + P(B) - (P(A) * P(B))$$

9

Probability Theory

- **Random variables:**
  - Domain: possible values
- **Atomic event:**
  - Complete specification of a state
- **Prior probability:**
  - Degree of belief without any *new* evidence
- **Joint probability:**
  - Matrix of **combined probabilities of a set of variables**, $P(A,B)$

- Alarm (A), Burglary (B), Earthquake (E)
  - Boolean, discrete, continuous
- $A=true \wedge B=true \wedge E=false$:
  - alarm ∧ burglary ∧ ¬earthquake
- $P(B) = 0.1$
- $P(A, B) =$

	alarm	¬ alarm
burglary	0.09	0.01
¬ burglary	0.1	0.8

10

Probability Distributions

- A distribution is the probabilities of **all possible values** of a random variable
- Ex: weather can be sunny, rainy, cloudy, or snowy
  - P(Weather = sun) = 0.6
  - P(Weather = rain) = 0.1
  - P(Weather = cloud) = 0.29
  - P(Weather = snow) = 0.01
  - P(Weather) = <0.6, 0.1, 0.29, 0.01> ← shortcut
- P(Weather): **probability distribution on Weather**

11

Probability Theory: Definitions

- Conditional probability: Probability of some effect given that we know cause(s)
  - Example: P(*alarm* | *burglary*)
    - (Technically, we only know b is correlated, not causal)
- Computing it:
  - $P(a \mid b) = \dfrac{P(a \wedge b)}{P(b)}$
- P(*b*): **normalizing constant** (later we'll call this alpha α or rho ϱ)

12

Probability Theory: Definitions

	alarm	¬ alarm
burglary	0.09	0.01
¬ burglary	0.1	0.8

- **Product rule:**
  - $P(a \wedge b) = P(a \mid b)\, P(b)$
- **Marginalizing** (summing out):
  - Finding distribution over *one* or a *subset* of variables
  - Marginal probability of B summed over all alarm states:
  - $P(B) = \Sigma_a P(B, a)$
    - $P(B)$ = sum of $P(B, a)$ for all possible values of A
- **Conditioning** over a subset of variables:
  - $P(B) = \Sigma_a P(B \mid a)\, P(a)$

13

Let's Try It

	alarm	¬ alarm
burglary	0.09	0.01
¬ burglary	0.1	0.8

- **Cond'l probability**
  - P(effect, cause[s])
  - P(a | b) = P(a ∧ b) / P(b)
  - P(b): normalizing constant (1/α)
- **Product rule:**
  - P(a ∧ b) = P(a | b) P(b)
- **Marginalizing:**
  - P(B) = ΣaP(B, a)
  - P(B) = ΣaP(B | a) P(a) (conditioning)

- $P(A \mid B) =$?
- $P(B \mid A) =$?
- $P(B \wedge A) =$?
- $P(A) =$?

14

Marginalizing

- Marginalization: how to **safely ignore variables**.
- Two-variable example (A and B).
- If we know $P(A{=}a{,}B{=}b)$ for *all* values of a and b:
- $P(B{=}b){=}\sum_a P(A{=}a{,}B{=}b)$.
- Here we "marginalized out" the variable A.
- Takes variable(s) in *a* out of consideration

15

Marginalizing

- **Marginalizing** (summing out):
  - Finding distribution over *one* or a *subset* of variables
  - Marginal probability of B summed over all alarm states:
  - $\mathrm{P}(B) = \Sigma_a \mathrm{P}(B,\ a)$
- Takes variable(s) in *a* out of consideration

16

Marginalizing Example

- You are a video games company, and you want to know the probability of a new user winning, $P(W)$. 1000 people already played. Your game has only two characters to choose.
- You know:
  - People who chose character A: $P(A) = 600/1000$
  - People who chose character B: $P(B) = 400/1000$
  - People who chose A and won, $P(W \mid A)$: 75/100
  - People who chose B and won, $P(W \mid B)$: 69/100
- What is P(W)?
- $P(W) = P(W \mid A)P(A) + P(W \mid B)\,P(B)$
  $= 75/100 \times 600/1000 + 69/100 \times 400/1000 = 0.75 \times 0.6 + 0.69 \times 0.4 = 0.726$

www.quora.com/What-is-marginalization-in-probability

17

Exercise: Inference from the Joint

- Queries:
  - What is the prior probability (knowing nothi
  - What is the prior probability of *study*?
  - What is the conditional probability of *prepared*, n *study* and *smart*?

Where do these come from?

P (smart ∧ study ∧ prep)	*smart*		*¬smart*	
	study	*¬study*	*study*	*¬study*
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

18

Exercise: Inference from the joint

- 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*?

***P*(*smart* ∧ *study* ∧ *prep*)**	***smart***		**¬*smart***	
	study	**¬*study***	***study***	**¬*study***
prepared	.432	.16	.084	.008
¬*prepared*	.048	.16	.036	.072

P(*smart*) = .432 + .16 + .048 + .16 = **0.8**

19

Exercise: Inference from the joint

- 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*?

***P*(*smart* ∧ *study* ∧ *prep*)**	***smart***		**¬*smart***	
	study	**¬*study***	***study***	**¬*study***
prepared	.432	.16	.084	.008
¬*prepared*	.048	.16	.036	.072

P(*study*) = .432 + .048 + .084 + .036 = **0.6**

20

Exercise: Inference from the joint

- 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*?

***P*(*smart* ∧ *study* ∧ *prep*)**	***smart***		***¬smart***	
	study	***¬study***	***study***	***¬study***
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

P(*prep*|*smart*,*study*) = P(*prep*, *smart*, *study*)/P(*smart*, *study*)

= .432 / (.432 + .048)

= **0.9**

21

Independence: $\perp\!\!\!\perp$

- **Independent:** Two sets of propositions that do not affect each others' probabilities
- Easy to calculate **joint** and **conditional** probability of independence:
  - $(A, B) \Leftrightarrow P(A \wedge B) = P(A)\,P(B)$ or $P(A \mid B) = P(A)$
- Examples:
  - A = alarm — M = moon phase
  - B = burglary — L = light level
  - E = earthquake

$A \perp\!\!\!\perp B \perp\!\!\!\perp E = f$
$M \perp\!\!\!\perp L = f$
$A \perp\!\!\!\perp M = t$

23

Independence Example

- *{moon-phase, light-level}* ⫫ *{burglary, alarm, earthquake}*
  - But maybe burglaries increase in low light
  - But, if we know the light level, *moon-phase* ⫫ *burglary*
  - Once we're burglarized, light level doesn't affect whether the alarm goes off; *{light-level}* ⫫ *{alarm}*
- We need:
  1. A more complex notion of independence
  2. Methods for reasoning about these kinds of (common) relationships

24

Exercise: Independence

- Is *smart* independent of *study*?
  - P(*smart* | *study*) = P(*smart*)
- Is *prepared* independent of *study*?
  - P(*prep* | *study*) = P(*prep*)

P(smart ∧ study ∧ prep)	***smart***		***¬smart***	
	study	***¬study***	***study***	***¬study***
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

25

Exercise: Independence

- Is *smart* independent of *study*?
  - P(*smart* | *study*) = P(*smart*)
- Is *prepared* independent of *study*?
  - P(*prep* | *study*) = P(*prep*)

Smart	Study		
t	*t*	0.432 + 0.48	0.480
t	*f*	0.16 + 0.16	0.32
f	*t*	0.084 + 0.008	0.092
f	*f*	0.036 + 0.72	0.756

***P* (smart ∧ study ∧ prep)**	***smart***		***¬smart***	
	study	***¬study***	***study***	***¬study***
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

26

Exercise: Independence

- P(*smart* | *study*) = P(*smart*)
- P(*smart* | *study*) = P(*smart*, *study*) / P(*study*)
- 0.8 = (.432 + .048) / .6
- 0.8 = 0.8

***P* (smart ∧ study ∧ prep)**	***smart***		***¬smart***	
	study	***¬study***	***study***	***¬study***
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Smart	Study		
t	*t*	0.432 + 0.48	0.480
t	*f*	0.16 + 0.16	0.32
f	*t*	0.084 + 0.008	0.092
f	*f*	0.036 + 0.72	0.756

27

Conditional Probabilities

- Describes **dependent** events
  - Affect each other in some way
- Typical in the real world
- If we know some event has occurred, what does that tell us about the likelihood of another event?

28

Conditional Independence

- *moon-phase* and *burglary* are **conditionally independent given** *light-level*
  - That is, $M \perp\!\!\!\perp B$ if we already know L
- Conditional independence is:
  - Weaker than absolute independence
  - Useful in decomposing full joint probability distributions

29

Conditional Independence

- Absolute independence: $A \perp\!\!\!\perp B$, if:
  - $P(A \wedge B) = P(A)\,P(B)$
  - Equivalently, $P(A) = P(A \mid B)$ and $P(B) = P(B \mid A)$
- A and B are conditionally independent given C if:
  - $P(A \wedge B \mid C) = P(A \mid C)\,P(B \mid C)$
- This lets us decompose the joint distribution:
  - $P(A \wedge B \wedge C) = P(A \mid C)\,P(B \mid C)\,P(C)$
- What does this mean?

30

Exercise: Conditional Independence

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

***P*(*smart* ∧ *study* ∧ *prep*)**	***smart***		**¬*smart***	
	study	**¬*study***	***study***	**¬*study***
prepared	.432	.16	.084	.008
¬*prepared*	.048	.16	.036	.072

31

Probability

- Worlds, random variables, events, sample space
- **Joint probabilities** of multiple connected variables
- **Conditional probabilities** of a variable, given another variable(s)
- **Marginalizing out** unwanted variables
- **Inference** from the joint probability

The big idea: figuring out the probability of variable(s) taking certain value(s)

32

Bayes' Rule

- Derive the probability of some **event**, *given* **another event**
  - Assumption of attribute independency
    (AKA the Naïve assumption)
  - Naïve Bayes assumes that all *attributes* are independent.
- Also the basis of modern machine learning
- Bayes' rule is derived from the product rule

R&N 495

33

Bayes' Rule

- $P(Y \mid X) = P(X \mid Y)\, P(Y) / P(X)$
- Often useful for diagnosis.
- If we have:
  - X = (observable) effects, e.g., symptoms
  - Y = (hidden) causes, e.g., illnesses
  - A model for how causes lead to effects: $P(X \mid Y)$
  - Prior beliefs about frequency of occurrence of effects: $P(Y)$
- We can reason from effects to causes: $P(Y \mid X)$

34

Naïve Bayes Algorithm

- Estimate the probability of each class:
  - Compute the posterior probability (Bayes rule)

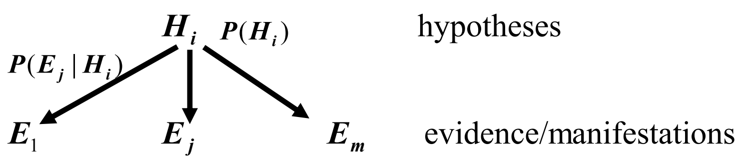
$$P(c_i \mid D) = \frac{P(c_i)P(D \mid c_i)}{P(D)}$$

  - Choose the class with the highest probability
- Assumption of attribute independency (Naïve assumption): Naïve Bayes assumes that all of the attributes are independent.

35

Bayesian Inference

- In the setting of diagnostic/evidential reasoning



  - Know: 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): $P(H_i | E_j) = P(H_i)P(E_j | H_i) / P(E_j)$

36

Simple Bayesian Diagnostic Reasoning

- We know:
  - 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): $E_1, \dots, E_m$
- Goal: Find the hypothesis H_i with the highest posterior
  - $\text{Max}_i\ P(Hi | E_1, \dots, E_m)$

37

Priors

- Four values total here:
  - $P(H \mid E) = (P(E \mid H) * P(H)) / P(E)$
- $P(H \mid E)$ — what we want to compute
- Three we already know, called the priors
  - $P(E \mid H)$
  - $P(H)$
  - $P(E)$

(In ML later, we will use the training set to estimate the priors)

38

Bayesian Diagnostic Reasoning II

- Bayes' rule says that
  - $P(H_i \mid E_1, \ldots, E_m) = P(E_1, \ldots, E_m \mid H_i)\, P(H_i) / P(E_1, \ldots, E_m)$
- Assume each piece of evidence E_i is **conditionally independent** of the others, **given** a hypothesis H_i, then:
  - $P(E_1, \ldots, E_m \mid H_i) = \prod^{l}_{j=1} P(E_j \mid H_i)$
- If we only care about relative probabilities for the H_i, then we have:
  - $P(H_i \mid E_1, \ldots, E_m) = \alpha\, P(H_i) \prod^{l}_{j=1} P(E_j \mid H_i)$

39

Bayes Exercise: Diagnosing Meningitis

$$P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$$

- Your patient comes in with a stiff neck.
- Is it meningitis?
- Suppose we know that
  - Stiff neck is a symptom in 50% of meningitis cases
  - Meningitis (m) occurs in 1/50,000 patients
  - Stiff neck (s) occurs in 1/20 patients
- So probably not. But specifically?

40

Bayes Exercise: Diagnosing Meningitis

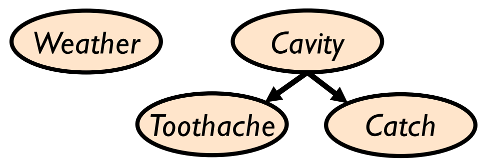
$$P(H_i \mid E_j) = P(H_i)P(E_j \mid H_i) / P(E_j)$$

- Stiff neck is a symptom in 50% of meningitis cases
- Meningitis (m) occurs in 1/50,000 patients
- Stiff neck (s) occurs in 1/20 patients

41

Bayes' Nets: Big Picture

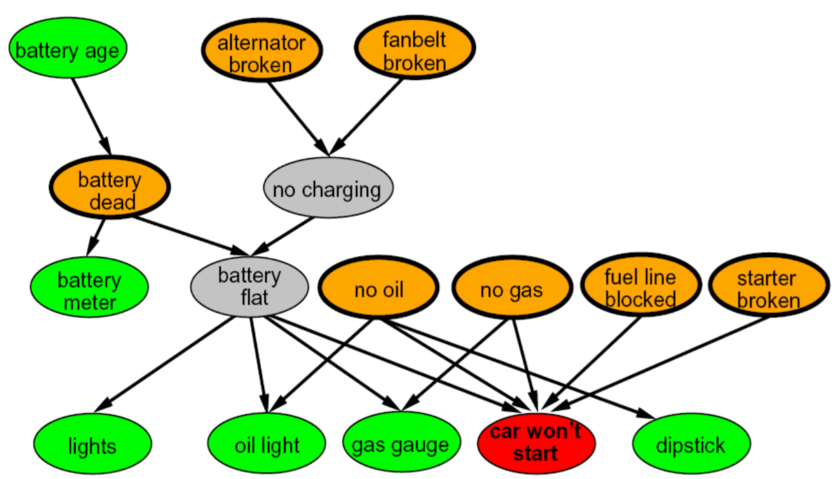
- **Bayes' nets:** a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - A type of graphical models
- We describe how variables interact **locally**
  - Local interactions chain together to give global, indirect interactions



Slides derived from Matt E. Taylor, U Alberta

42

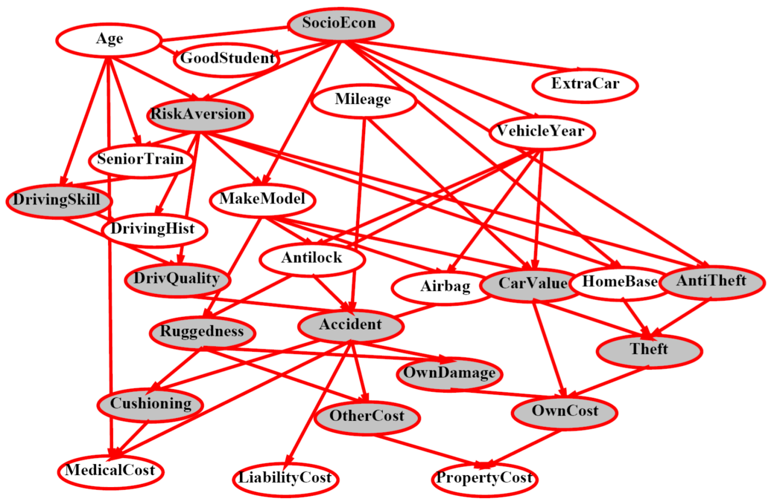
Example: Car Won't Start



Slides derived from Matt E. Taylor, U Alberta

43

Example: Insurance

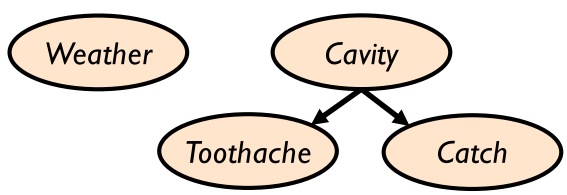


Slides derived from Matt E. Taylor, U Alberta

44

Example: Toothache

- Random variables:
  - How's the weather?
  - Do you have a toothache?
  - Does the dentist's probe catch when she pokes your tooth?
  - Do you have a cavity?

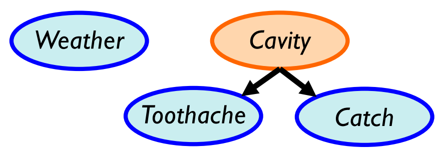


Slides derived from Matt E. Taylor, U Alberta

45

Graphical Model Notation

- Nodes: variables (with domains)
- Can be assigned (observed) or unassigned (hidden)
- Arcs: interactions
  - Indicate "direct influence" between
  - Formally: **encode conditional independence**
    - Toothache and Catch are **conditionally independent, given Cavity**
- For now: imagine that arrows mean *causation*
  - (in general, they don't!)

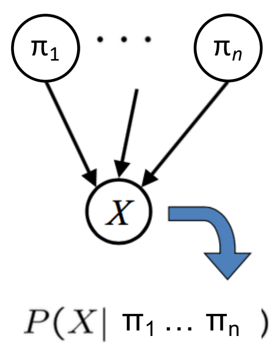


Slides derived from Matt E. Taylor, U Alberta

46

Bayesian Belief Networks (BNs)

- Let's formalize the semantics of a BN
  - A set of nodes, one per variable *X*
- A directed arc between each co-influential node
  - $X \rightarrow Y$ means *X* **has an influence on** *Y*
- A directed, acyclic graph



Slides derived from Matt E. Taylor, U Alberta

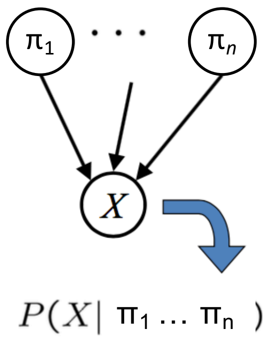
47

Bayesian Belief Networks (BNs)

- Each node X has a **conditional probability distribution**:

  $P(X_i \mid Parents(X_i))$

  - A collection of distributions over X
  - One for each combination of parents' values
  - Quantifies the effects of the parents on a node
- CPT: conditional probability table
  - Description of a noisy "causal" process



Slides derived from Matt E. Taylor, U Alberta

48

Conditional Probability Tables

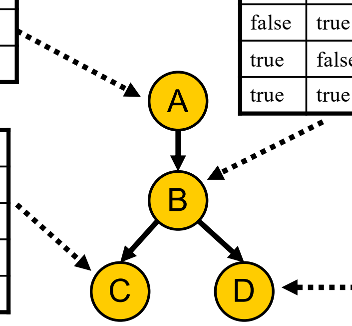
- For X_i, CPD $P(X_i \mid Parents(X_i))$ quantifies effect of parents on X_i
- **Parameters** are probabilities in conditional probability tables (CPTs):

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95



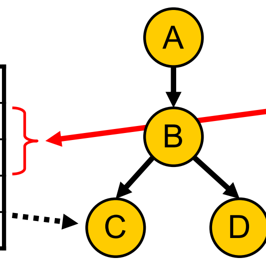
Example from web.engr.oregonstate.edu/~wong/slides/BayesianNetworksTutorial.ppt

49

CPTs cont'd

- Conditional Probability Distribution for C given B
- If you have a Boolean variable with *k* Boolean parents, this table has 2^{k+1} probabilities

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

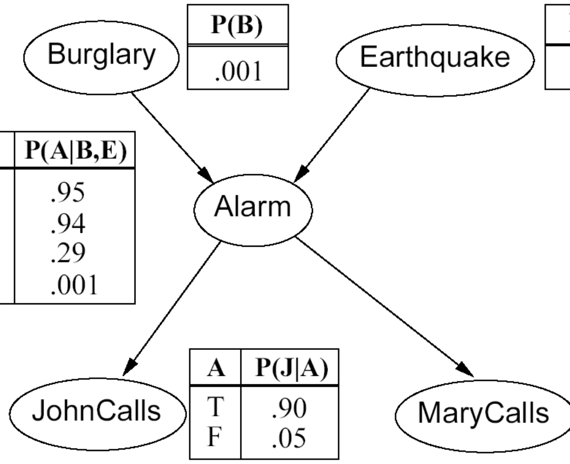


For a given combination of values of the parents (B in this example), the entries for P(C=true | B) and P(C=false | B) must sum to 1
Example:
P(C=true | B=false) +
P(C=false |B=false) = 1

Example from web.engr.oregonstate.edu/~wong/slides/BayesianNetworksTutorial.ppt

50

Bayesian Belief Networks (BNs)



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

51

Bayesian Belief Networks (BNs)

Making a Bayesian Network BN: **BN = (DAG, CPD)**

- **DAG**: directed acyclic graph (BN's structure)
  - **Nodes**: random variables
    - Typically binary or discrete
    - Methods exist for continuous variables
  - **Arcs**: indicate probabilistic dependencies between nodes
    - Lack of link signifies conditional independence
- **CPD**: conditional probability distribution (BN's parameters)
    - Conditional probabilities at each node, usually stored as a table (conditional probability table, or **CPT**)

52

Bayesian Belief Networks (BNs)

Making a Bayesian Network BN: **BN = (DAG, CPD)**

- **DAG**: directed acyclic graph (BN's structure)
- **CPD**: conditional probability distribution (BN's parameters)
  - Conditional probabilities at each node, usually stored as a table (conditional probability table, or **CPT**)

$$P(x_i \mid \pi_i) \text{ where } \pi_i \text{ is the set of all parent nodes of } x_i$$

- Root nodes are a special case
  - No parents, so use priors in CPD:

$$\pi_i = \varnothing, \text{ so } P(x_i \mid \pi_i) = P(x_i)$$

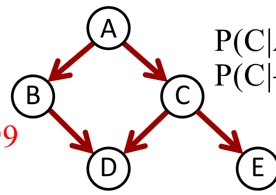
53

Example BN

- We only specify P(A) etc., not P(¬A), since they have to sum to one

P(A) = 0.001

P(B|A) = 0.3
P(B|¬A) = 0.001
P(¬B|A) = 0.7
P(¬B|¬A) = 0.999



P(C|A) = 0.2
P(C|¬A) = 0.005

P(E|C) = 0.4
P(E|¬C) = 0.002

P(D|B,C) = 0.1
P(D|B,¬C) = 0.01
P(D|¬B,C) = 0.01
P(D|¬B,¬C) = 0.00001

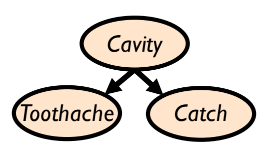
54

Probabilities in BNs

- Bayes' nets implicitly encode **joint distributions** as a **product of local conditional distributions**.
- To see probability of a full assignment, multiply all the relevant conditionals together:

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

  - Example: P(+cavity, +catch, ¬toothache) = ?



- This lets us reconstruct any entry of the full joint

Slides derived from Matt E. Taylor, U Alberta

55

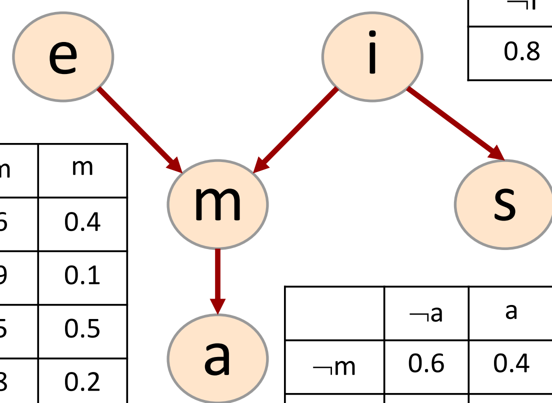
P(a, m, i, e, s) =

P(¬a, ¬m, i, ¬e, s) =

P(a, m, ¬i, e, ¬s) =

¬e	e
0.7	0.3

¬i	i
0.8	0.2



	¬m	m
¬i, ¬e	0.6	0.4
¬i, e	0.9	0.1
i, ¬e	0.5	0.5
i, e	0.8	0.2

	¬s	s
¬i	0.75	0.25
i	0.4	0.6

	¬a	a
¬m	0.6	0.4
m	0.9	0.1

www.upgrad.com/blog/bayesian-network-example/

56

Summary

- Probability review
  - Distributions, conditional probability, marginalizing
  - Independence
  - Bayes' rule
- Bayes' nets (Bayesian Belief Networks)
  - Graphical notation
  - Conditional probability tables
  - Probability distributions
- Next time
  - Inference using Bayes' nets

57