Uncertainty
[RN2 Sec. 13.1-13.6]
[RN3 Sec. 13.1-13.5]

CS 486/686
University of Waterloo
Lecture 7: October 2, 2012

A Decision Making Scenario

- You are considering to buy a used car…
  - Is it in good condition?
  - How much are you willing to pay?
  - Should you get it inspected by a mechanics?
  - Should you buy the car?

In the next few lectures

- Probability theory
  - Model uncertainty
- Utility theory
  - Model preferences
- Decision theory
  - Combine probability theory and utility theory

Introduction

- Logical reasoning breaks down when dealing with uncertainty
- Example: Diagnosis
  - $\forall$p Symptom(p,Toothache) $\Rightarrow$ Disease(p, Cavity)
    - But not all people with toothaches have cavities…
  - $\forall$p Symptom(p, Toothache) $\Rightarrow$ Disease(p,Cavity) v Disease(p,Gumdisease) v Disease(p, Hit in the Jaw) v …
    - Can't enumerate all possible causes and not very informative
  - $\forall$p Disease(p, Cavity) $\Rightarrow$ Symptom(p,Toothache)
    - Does not work since not all cavities cause toothaches…

Introduction

- Logic fails because
  - We are lazy
    - Too much work to write down all antecedents and consequences
  - Theoretical ignorance
    - Sometimes there is just no complete theory
  - Practical ignorance
    - Even if we knew all the rules, we might be uncertain about a particular instance (not collected enough information yet)

Probabilities to the rescue

- For many years AI danced around the fact that the world is an uncertain place
- Then a few AI researchers decided to go back to the 18th century
  - Revolutionary
  - Probabilities allow us to deal with uncertainty that comes from our laziness and ignorance
  - Clear semantics
  - Provide principled answers for
    - Combining evidence, predictive and diagnostic reasoning, incorporation of new evidence
  - Can be learned from data
  - Intuitive for humans (?)

Discrete Random Variables

- Random variable *A* describes an outcome that cannot be determined in advance (i.e. roll of a dice)
  - Discrete random variable means that its possible values come from a countable domain (sample space)
    - E.G If X is the outcome of a dice throw, then X $\in$ {1,2,3,4,5,6}
  - **Boolean random variable** *A* $\in$ {True, False}
    - *A* = The Canadian PM in 2040 will be female
    - *A* = You have Ebola
    - *A* = You wake up tomorrow with a headache

Events

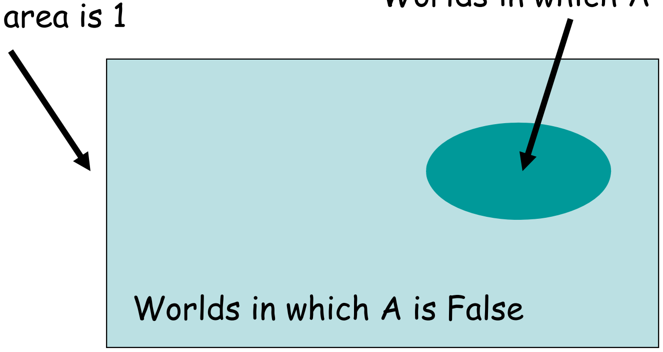
- An event is a complete specification of the state of the world in which the agent is uncertain
  - Subset of the sample space
- Example:
  - Cavity=True ∧ Toothache=True
  - Dice=2
- Events must be
  - Mutually exclusive
  - Exhaustive (at least one event must be true)

Probabilities

- We let P(A) denote the "degree of belief" we have that statement A is true
  - Also "fraction of worlds in which A is true"
    - Philosophers like to discuss this (but we won't)

- Note:
  - P(A) DOES NOT correspond to a degree of truth
  - Example: Draw a card from a shuffled deck
    - The card is of some type (e.g ace of spades)
    - Before looking at it P(ace of spades) = 1/52
    - After looking at it P(ace of spades) = 1 or 0

Visualizing A



P(A) = Area of oval

The Axioms of Probability

- $0 \leq P(A) \leq 1$
- P(True) = 1
- P(False) = 0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

- These axioms limit the class of functions that can be considered as probability functions

Interpreting the axioms

- $0 \leq P(A) \leq 1$
- P(True) = 1
- P(False) = 0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

The area of A can't be smaller than 0

A zero area would mean no world could ever have A as true

Interpreting the axioms

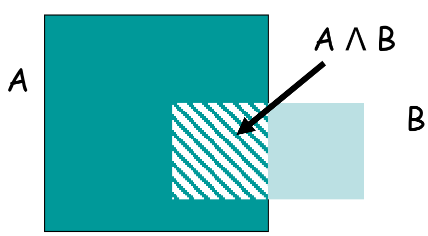
- $0 \leq P(A) \leq 1$
- P(True) = 1
- P(False) = 0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

The area of A can't be larger than 1

An area of 1 would mean all possible worlds have A as true

Interpreting the axioms

- $0 \leq P(A) \leq 1$
- P(True) = 1
- P(False) = 0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Take the axioms seriously!

- There have been attempts to use different methodologies for uncertainty
  - Fuzzy logic, three valued logic, Dempster-Shafer, non-monotonic reasoning,…
- But if you follow the axioms of probability then no one can take advantage of you ☺

A Betting Game [di Finetti 1931]

- Propositions A and B
- Agent 1 announces its "degree of belief" in A and B (P(A) and P(B))
- Agent 2 chooses to bet for or against A and B at stakes that are consistent with P(A) and P(B)
- If Agent 1 does not follow the axioms, it is guaranteed to lose money

Agent 1 Proposition	Belief	Agent 2 Bet	Odds	Outcome for Agent 1 A∧B	A∧~B	~A∧B	~A∧~B
A	0.4	A	4 to 6	-6	-6	4	4
B	0.3	B	3 to 7	-7	3	-7	3
AVB	0.8	~(AVB)	2 to 8	2	2	2	-8
				-11	**-1**	**-1**	**-1**

Theorems from the axioms

- Thm: P(~A)=1-P(A)
- Proof: P(AV~A)=P(A)+P(~A)-P(A∧~A)
  P(True)=P(A)+P(~A)-P(False)
  1 = P(A)+P(~A)-0
  P(~A)=1-P(A)

Theorems from axioms

- Thm: P(A) = P(A∧B) + P(A∧~B)
- Proof: For you to do

Why? Because it is good for you

Multivalued Random Variables

- Assume domain of A (sample space) is $\{v_1, v_2, \ldots, v_k\}$
- A can take on exactly one value out of this set
  - $P(A=v_i \wedge A=v_j) = 0$ if $i \neq j$
  - $P(A=v_1 \vee A=v_2 \vee \ldots \vee A=v_k) = 1$

Terminology

- Probability distribution:
  - A specification of a probability for each event in our sample space
  - Probabilities must sum to 1
- Assume the world is described by two (or more) random variables
  - Joint probability distribution
    - Specification of probabilities for all combinations of events

Joint distribution

- Given two random variables A and B:
- Joint distribution:
  - $\Pr(A=a \wedge B=b)$ for all a,b
- Marginalisation (sumout rule):
  - $\Pr(A=a) = \Sigma_b \Pr(A=a \wedge B=b)$
  - $\Pr(B=b) = \Sigma_a \Pr(A=a \wedge B=b)$

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

$P(headache \wedge sunny \wedge cold) = 0.108$ $P(\sim headache \wedge sunny \wedge \sim cold) = 0.064$

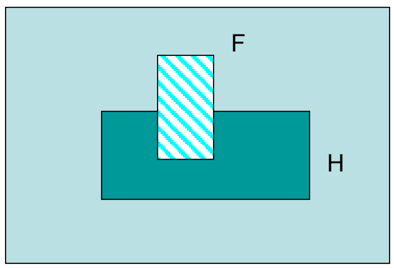
$P(headache \vee sunny) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

$P(headache) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$

marginalization

Conditional Probability

- P(A|B) fraction of worlds in which B is true that also have A true

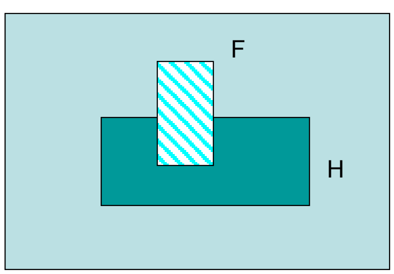


H="Have headache"
F="Have Flu"

P(H)=1/10
P(F)=1/40
P(H|F)=1/2

Headaches are rare and flu is rarer, but if you have the flu, then there is a 50-50 chance you will have a headache

Conditional Probability



H="Have headache"
F="Have Flu"

P(H)=1/10
P(F)=1/40
P(H|F)=1/2

P(H|F)= Fraction of flu inflicted worlds in which you have a headache

=(# worlds with flu and headache)/ (# worlds with flu)

= (Area of "H and F" region)/ (Area of "F" region)

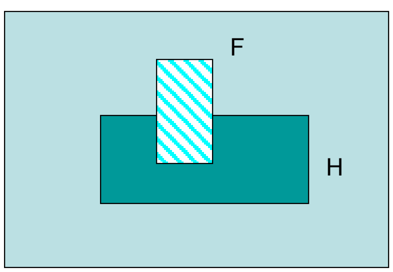
= $P(H \wedge F)/ P(F)$

Conditional Probability

- Definition:
  - P(A|B) = P(A∧B) / P(B)
- Chain rule:
  - P(A∧B) = P(A|B) P(B)

Memorize these!

Inference



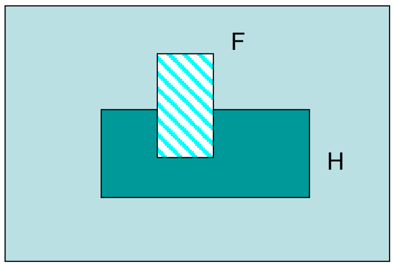
One day you wake up with a headache. You think "Drat! 50% of flues are associated with headaches so I must have a 50-50 chance of coming down with the flu"

H="Have headache"
F="Have Flu"

P(H)=1/10
P(F)=1/40
P(H|F)=1/2

Is your reasoning correct?

Inference



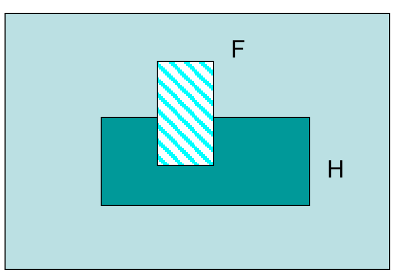
One day you wake up with a headache. You think "Drat! 50% of flues are associated with headaches so I must have a 50-50 chance of coming down with the flu"

H="Have headache"
F="Have Flu"

$P(F \wedge H)=P(F)P(H|F)=1/80$

$P(H)=1/10$
$P(F)=1/40$
$P(H|F)=1/2$

Inference



One day you wake up with a headache. You think "Drat! 50% of flues are associated with headaches so I must have a 50-50 chance of coming down with the flu"

H="Have headache"
F="Have Flu"

$P(F \wedge H)=P(F)P(H|F)=1/80$

$P(H)=1/10$
$P(F)=1/40$
$P(H|F)=1/2$

$P(F|H) = P(F \wedge H)/P(H) = 1/8$

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

P(headache ∧ cold | sunny) = P(headache ∧ cold ∧ sunny) / P(sunny)
= 0.108/(0.108+0.012+0.016+0.064)
= 0. 54

P(headache ∧ cold | ~sunny) = P(headache ∧ cold ∧ ~sunny) / P(~sunny)
= 0.072/(0.072+0.008+0.144+0.576)
= 0.09

Bayes Rule

- Note
  - P(A|B)P(B) = P(A∧B) = P(B∧A)=P(B|A)P(A)
- Bayes Rule
  - P(B|A)= [P(A|B)P(B)]/P(A)

Memorize this!

Using Bayes Rule for inference

- Often we want to form a hypothesis about the world based on what we have observed
- Bayes rule is vitally important when viewed in terms of stating the belief given to hypothesis H, given evidence e

Likelihood

Prior probability

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

Posterior probability

Normalizing constant

More General Forms of Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

$$P(A|B \wedge X) = \frac{P(B|A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

$$P(A = v_i|B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{k=1}^{n} P(B|A = v_k)P(A = v_k)}$$

Example

- A doctor knows that Asian flu causes a fever 95% of the time. She knows that if a person is selected at random from the population, they have a 10^{-7} chance of having Asian flu. 1 in 100 people suffer from a fever.
- You go to the doctor complaining about the symptom of having a fever. What is the probability that Asian flu is the cause of the fever?

Example

- A doctor knows that Asian flu causes a fever 95% of the time. She knows that if a person is selected at random from the population, they have a 10^{-7} chance of having Asian flu. 1 in 100 people suffer from a fever.
- You go to the doctor complaining about the symptom of having a fever. What is the probability that Asian flu is the cause of the fever?

A=Asian flu

F= fever

Evidence = Symptom (F)

Hypothesis = Cause (A)

$$
\begin{aligned}
P(A|F) &= \frac{P(F|A)P(A)}{P(F)} \\
&= \frac{0.95 \times 10^{-7}}{0.01} \\
&= 0.95 \times 10^{-5}
\end{aligned}
$$

Computing conditional probabilities

- Often we are interested in the posterior joint distribution of some query variables Y given specific evidence e for evidence variables E
- Set of all variables: X
- Hidden variables: H=X-Y-E
- If we had the joint probability distribution then could marginalize
- $P(Y|E=e) = \alpha \Sigma_h P(Y \wedge E=e \wedge H=h)$
  - α is the normalization factor

Computing conditional probabilities

- Often we are interested in the posterior joint distribution of some query variables Y given specific evidence e for evidence variables E
- Set of all variables: X
- Hidden variables: H=X-Y-E
- If we had the joint probability distribution then could marginalize
- $P(Y|E=e) = \alpha \Sigma_h P(Y \wedge E=e \wedge H=h)$
  - α is the normalization factor

Problem: Joint distribution is usually too big to handle

CS486/686 Lecture Slides (c) 2012 K. Larson and P. Poupart

Independence

- Two variables A and B are independent if knowledge of A does not change uncertainty of B (and vice versa)
  - $P(A|B)=P(A)$
  - $P(B|A)=P(B)$
  - $P(A \wedge B)=P(A)P(B)$
  - In general $P(X_1,X_2,\ldots,X_n)=\Pi_{i=1}^{n} P(X_i)$

Need only n numbers to specify a joint distribution!

Conditional Independence

- Absolute independence is often too strong a requirement

- Two variables A and B are conditionally independent given C if
  - $P(a|b,c)=P(a|c)$ for all a,b,c
  - i.e. knowing the value of B does not change the prediction of A if the value of C is known

Conditional Independence

- Diagnosis problem
  - Fl = Flu, Fv = Fever, C=Cough
- Full joint distribution has 2^3-1=7 independent entries
- If someone has the flu, we can assume that the probability of a cough does not depend on having a fever
  - P(C|Fl,Fv)=P(C|Fl)
- If the patient does not have the Flu, then C and Fv are again conditionally independent
  - P(C|~Fl, Fv)=P(C|~Fl)

Conditional Independence

- Full distribution can be written as
  - P(C,Fl,Fv)= P(C,Fv|Fl)P(Fl)
    = P(C|Fl)P(Fv|Fl)P(Fl)
  - That is we only need 5 numbers now!
  - Huge savings if there are lots of variables

Conditional Independence

- Full distribution can be written as
  - P(C,Fl,Fv)= P(C,Fv|Fl)P(Fl)
    = P(C|Fl)P(Fv|Fl)P(Fl)
  - That is we only need 5 numbers now!
  - Huge savings if there are lots of variables

Such a probability distribution is sometimes called a naïve Bayes model.

In practice, they work well – even when the independence assumption is not true

Next class

- Bayesian networks