Uncertainty

CS 486/686: Introduction to Artificial Intelligence
Introduction

Agents rarely have access to the full truth about their environment

- We are lazy
  - Too much work to write down all antecedents and consequences

- Theoretical ignorance
  - Sometimes there is no complete theory

- Practical ignorance
  - Even if we knew all the rules, we might be uncertain about a particular instance (not enough information yet)
Probability to the Rescue

• Allows us to deal with uncertainty that comes from laziness or ignorance
• Clear semantics
• Provides principled answers for
  - combining evidence, predictive and diagnostic reasoning, incorporation of new evidence
• Can be learned from data
Discrete Random Variables

- Random variable $A$ describes an outcome that cannot be determined in advance (i.e., roll of a dice).
- Discrete random variable: possible values come from a countable domain (sample space).
  - If $X$ is the outcome of a dice throw then $X \in \{1, 2, 3, 4, 5, 6\}$
- **Boolean random variable:** $A \in \{\text{True, False}\}$
  - $A =$ The Canadian PM in 2040 will be male
  - $A =$ 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 an agent is uncertain
  - Subset of the sample space

• Example
  - (Cavity=True) ∧ (Toothache=True)
  - Dice=2

• Events must be
  - Mutually exclusive
  - Exhaustive
Probabilities

- We let $P(A)$ denote the “degree of belief” we have that statement $A$ is true
  - “The fraction of possible worlds in which $A$ is true”
- **Note:** $P(A)$ DOES NOT correspond to a degree of truth
Visualizing $A$

Event space of all possible worlds. It’s area is 1

Worlds in which $A$ is False

Worlds in which $A$ is True

$P(A) = \text{Area of oval}$
Axioms of Probability

• $0 \leq P(A) \leq 1$
• $P(\text{True}) = 1$
• $P(\text{False}) = 0$
• $P(A \lor B) = P(A) + P(B) - P(A \land B)$

• These axioms limit the class of functions that can be considered as probability functions
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

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

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

The area of $A$ can’t be larger than 1

An area of 1 would mean no world could ever have $A$ as false
Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Take the Axioms Seriously

- There have been attempts to use different methodologies for uncertainty
  - Fuzzy logic
  - Three-valued logic
  - Dempster-Shafer
  - ...
- But if you follow the axioms of probability then no one can take advantage of you :)
Theorems from the Axioms

- **Thm:** $P(\sim A) = 1 - P(A)$
- **Proof:**
  
  \[
P(A \lor \sim A) = P(A) + P(\sim A) - P(A \land \sim A)
  
  = P(A) + P(\sim A) - P(False)
  
  = P(A) + P(\sim A) - 0
  
  = 1
  
  \]

  \[
P(\sim A) = 1 - P(A)
  
  \]
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
  - \( P(A=v_i, A=v_j)=0 \) if \( i \) not equal to \( j \)
  - \( P(A=v_1 \text{ or } A=v_2 \text{ or } \ldots \text{ or } A=v_k)=1 \)
Useful Fact

• Given axioms of probability and
  \( P(A=v_i, A=v_j) = 0 \) for \( i \neq j \), and \( P(A=v_1 \text{ or } A=v_2 \text{ or } \ldots \text{ or } A=v_k) = 1 \) then

  - \( P(A=v_1 \text{ or } A=v_2 \text{ or } \ldots \text{ or } A=v_i) = \sum_{j=1}^{i} P(A=v_j) \)

  - \( \sum_{j=1}^{k} P(A=v_j) = 1 \)
Terminology

• **Probability Distribution**
  - A specification of a probability for each event in the sample space

• Assume the world is described by two or more random variables
  - **Joint probability distribution**
    - Specification of probabilities for all combinations of events
Useful Fact

- Given axioms of probability and $P(A=v_i, A=v_j)=0$ for $i \neq j$, and $P(A=v_1$ or $A=v_2$ or ... or $A=v_k)=1$ then
  
  - $P(B, (A=v_1$ or $A=v_2$ or ... or $A=v_i))=\sum_{j=1}^{i} P(B, A=v_j)$
  
  - $\sum_{j=1}^{k} P(B, A=v_j)=1$  

Marginalization
Example: Joint Distribution

<table>
<thead>
<tr>
<th></th>
<th>sunny</th>
<th>~sunny</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cold</td>
<td>~cold</td>
</tr>
<tr>
<td>headache</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>~headache</td>
<td>0.016</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>~cold</td>
</tr>
<tr>
<td>headache</td>
<td>0.072</td>
<td>0.008</td>
</tr>
<tr>
<td>~headache</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>

\[
P(\text{headache} \land \text{sunny} \land \text{cold}) = 0.108 \quad P(\sim \text{headache} \land \text{sunny} \land \sim \text{cold}) = 0.064
\]

\[
P(\text{headache} \lor \text{sunny}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28
\]

\[
P(\text{headache}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2
\]

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 that there is a 50-50 chance you will have a headache.
**Conditional Probability**

\[ P(H|F) = \frac{\text{Fraction of flu inflicted}}{\text{worlds in which you have a headache}} \]

\[ = \frac{\text{(# worlds with flu and headache)}}{\text{(# worlds with flu)}} \]

\[ = \frac{\text{(Area of “H and F” region)}}{\text{(Area of “F” region)}} \]

\[ = \frac{P(H \cap F)}{P(F)} \]

- \( H = \text{“Have headache”} \)
- \( F = \text{“Have Flu”} \)

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

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

- $P(A|B) = \frac{P(A \land B)}{P(B)}$
- Chain Rule:
  - $P(A \land B) = P(A|B)P(B)$

Memorize these!
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”

Is your reasoning correct?

H=“Have headache”
F=“Have Flu”

P(H)=1/10
P(F)=1/40
P(H|F)=1/2
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”

\[
P(H) = \frac{1}{10} \\
P(F) = \frac{1}{40} \\
P(H|F) = \frac{1}{2}
\]

\[
P(F \land H) = \\
P(F|H) =
\]
Example: Joint Distribution

\[
P(\text{headache} \land \text{cold} | \text{sunny}) = \frac{P(\text{headache} \land \text{cold} \land \text{sunny})}{P(\text{sunny})}
\]
\[
= \frac{0.108}{(0.108 + 0.012 + 0.016 + 0.064)}
\]
\[
= 0.54
\]

\[
P(\text{headache} \land \text{cold} | \neg \text{sunny}) = \frac{P(\text{headache} \land \text{cold} \land \neg \text{sunny})}{P(\neg \text{sunny})}
\]
\[
= \frac{0.072}{(0.072 + 0.008 + 0.144 + 0.576)}
\]
\[
= 0.09
\]
Bayes Rule

• Note:
  - \( P(A|B)P(B) = P(A \land B) = P(B \land A) = P(B|A)P(A) \)

• Bayes Rule:
  - \( P(B|A) = \frac{P(A|B)P(B)}{P(A)} \)

**Memorize this!**
General Forms of Bayes Rule

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

\[ P(A|B \land X) = \frac{P(B|A \land X)P(A \land X)}{P(B \land X)} \]

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

\[
P(H|e) = \frac{P(e|H)P(H)}{P(e)}
\]

- **Prior probability**
- **Likelihood**
- **Posterior probability**
- **Normalizing constant**
Example

• A doctor knows that H1N1 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 H1N1. 1 in 100 people suffer from a fever.

• You go to the doctor complaining about a fever. What is the probability that H1N1 is the cause of the fever?
Computing Conditional Probabilities

• Often we are interested in the posterior joint distribution of some query variable Y given specific evidence e for evidence variables E
  - Hidden variables: X-Y-E

• If we had the joint prob. distribution then could marginalize
  - $P(Y|E=e) = \alpha \sum_h P(Y \land (E=e) \land (H=h))$
Computing Conditional Probabilities

- Often we are interested in the posterior joint distribution of some **query variable** $Y$ given specific evidence $e$ for **evidence variables** $E$
  - Hidden variables: $X$-$Y$-$E$

- If we had the joint prob. distribution then could marginalize
  - $P(Y|E=e)=\alpha\sum_h P(Y\land(E=e)\land(H=h))$

**Problem:** Joint distribution is usually too big to handle
Independence

- Two variables A and B are independent if knowledge of A does not change uncertainty of B (and vice versa)
  - \( P(A|B) = P(A) \)
  - \( P(B|A) = P(B) \)
  - \( P(A \land B) = P(A)P(B) \)
  - In general: \( P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i) \)
Conditional Independence

- Full independence is often too strong a requirement

- Two variables A and B are conditionally independent given C if
  - \( P(ab,c) = P(alc) \) 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 dist. 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 $(P(C \mid Fl,Fv)=P(C \mid Fl))$

• If the same condition holds if the patient does not have the Flu then C and Fv are *conditionally independent* given FL $(P(C \mid \sim Fl, Fv)=P(C \mid \sim Fl))$
Conditional Independence

- Full distribution can be written as

\[ P(C, Fl, FC) = P(C, Fv|Fl)P(Fl) \]
\[ = P(C|Fl)P(Fv|Fl)P(Fl) \]

- We only need 5 numbers!
- Huge savings if there are lots of variables
Conditional Independence

- Such a probability distribution is sometimes called a **Naive Bayes model**

- In practice they work well - even when the independence assumption is not true
Summary

• What you **must** know
  - Basic definitions and axioms
  - Marginalization
  - Conditional Probabilities
  - Chain Rule and Bayes Rule
  - Independence and Conditional Independence

Seriously! Otherwise the next few weeks are going to be painful!