Lecture 10: Introduction to reasoning under uncertainty

- Introduction to reasoning under uncertainty
- Review of probability
  - Axioms and inference
  - Conditional probability
  - Probability distributions

Uncertainty

- Back to planning:
  - Let action *A(t)* denote leaving for the airport *t* minutes before the flight
  - For a given value of *t*, will *A(t)* get me there on time?
- Problems:
  - Partial observability (roads, other drivers' plans, etc.)
  - Noisy sensors (traffic reports)
  - Uncertainty in action outcomes (flat tire, etc.)
  - Immense complexity of modeling and predicting traffic

How To Deal With Uncertainty

- Implicit methods:
  - Ignore uncertainty as much as possible
  - Build procedures that are robust to uncertainty
  - This is the approach in the planning methods studied so far (e.g. monitoring and replanning)
- *Explicit methods*
  - Build a *model* of the world that *describes the uncertainty* (about the system's state, dynamics, sensors, model)
  - Reason about the effect of actions given the model

How To Represent Uncertainty

- What language should we use? What are the semantics of our representations?
- What queries can we answer with our representations? How do we answer them?
- How do we construct a representation? Do we need to ask an expert or can we learn from data?

Why Not Use First-Order Logic?

- A purely logical approach has two main problems:
  - Risks falsehood
    - A(25) will get me there on time.
  - Leads to conclusions that are too weak:
    - *A(25)* will get me there on time if there is no accident on the bridge and it does not rain and my tires remain intact, etc. etc.
    - *A(1440)* might reasonably be said to get me there on time (but I would have to stay overnight at the airport!)

Methods for Handling Uncertainty

- *Default (non-monotonic) logic*: make assumptions unless contradicted by evidence.
  - E.g. "Assume my car doesnt have a flat tire.

  What assumptions are reasonable? What about contradictions?
- *Rules with fudge factor:*
  - E.g. "Sprinkler $\rightarrow_{0.99}$ WetGrass", "WetGrass $\rightarrow_{0.7}$ Rain"

  But: Problems with combination (e.g. Sprinkler causes rain?)
- *Probability*:
  - E.g. Given what I know, *A(25)* succeed with probability 0.2
- *Fuzzy logic*:
  - E.g. WetGrass is true to degree 0.2

  But: Handles degree of truth, NOT uncertainty.

Probability

- A well-known and well-understood framework for dealing with uncertainty
- Has a clear semantics
- Provides principled answers for:
  - Combining evidence
  - Predictive and diagnostic reasoning
  - Incorporation of new evidence
- Can be learned from data
- Intuitive to human experts (arguably?)

Beliefs (Bayesian Probabilities)

- We use probability to describe uncertainty due to:
  - Laziness: failure to enumerate exceptions, qualifications etc.
  - Ignorance: lack of relevant facts, initial conditions etc.
  - True randomness? Quantum effects? ...
- *Beliefs (Bayesian or subjective probabilities)* relate propositions to one's current state of knowledge
  - E.g. $P(A(25)|\ \textit{no reported accident}) = 0.1$

  These are *not assertions about the world / absolute truth*
- Beliefs change with new evidence:
  - E.g. $P(A(25)|\ \textit{no reported accident, 5am}) = 0.2$
- This is analogous to logical entailment: $KB \vdash \alpha$ means that α is true *given* the KB, but may not be true in general.

Making Decisions Under Uncertainty

- Suppose I believe 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*
- Which action should I choose?

Making Decisions Under Uncertainty

- Suppose I believe 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*
- Which action should I choose?
  - Depends on my preferences for missing flight vs. airport cuisine, etc.
  - *Utility theory* is used to represent and infer preferences.
  - *Decision theory* $=$ utility theory $+$ probability theory

Random Variables

- A *random variable* X describes an outcome that cannot be determined in advance
  - E..g. The roll of a die
  - E.g. Number of e-mails received in a day
- The *sample space (domain)* S of a random variable X is the set of all possible values of the variable
  - E.g. For a die, $S = \{1, 2, 3, 4, 5, 6\}$
  - E.g. For number of emails received in a day, S is the natural numbers
- An *event* is a subset of S.
  - E.g. $e = \{1\}$ corresponds to a die roll of 1
  - E.g. number of e-mails in a day more than 100

Probability for Discrete Random Variables

- Usually, random variables are governed by some "law of nature", described as a *probability function* P defined on S.
- $P(x)$ defines the chance that variable X takes value $x \in S$.
  - E.g. for a die roll with a fair die, $P(1) = P(2) = \cdots = P(6) = 1/6$
- Note that we still cannot determine the value of X, just the chance of encountering a given value
- If X is a discrete variable, then a probability space $P(x)$ has the following properties:

$$0 \leq P(x) \leq 1, \forall x \in S \text{ and } \sum_{x \in S} P(x) = 1$$

Beliefs

- We use probability to describe the world and existing uncertainty
- Agents will have *beliefs* based on their current state of knowledge
  - E.g. *P(Some day AI agents will rule the world)=0.2* reflects a personal belief, based on one's state of knowledge about current AI, technology trends etc.
- Different agents may hold different beliefs, as these are *subjective*
- Beliefs may change over time as agents get new evidence
- *Prior (unconditional) beliefs* denote belief prior to the arrival of any new evidence.

Axioms of Probability

- Beliefs satisfy the axioms of probability.
- For any propositions A, B:
  1. $0 \leq P(A) \leq 1$
  2. $P(\text{True}) = 1$ (hence $P(\text{False}) = 0$)
  3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
  4. Alternatively, if A and B are mutually exclusive $(A \wedge B = F)$ then:

$$P(A \vee B) = P(A) + P(B)$$

- The axioms of probability limit the class of functions that can be considered probability functions.

Defining Probabilistic Models

- We define the world as a set of random variables $\Omega = \{X_1 \dots X_n\}$.
- A *probabilistic model* is an encoding of probabilistic information that allows us to compute the probability of any event in the world
- The world is divided into a set of elementary, mutually exclusive events, called *states*
  - E.g. If the world is described by two Boolean variables A, B, a state will be a complete assignment of truth values for A and B.
- A *joint probability distribution function* assigns non-negative weights to each event, such that these weights sum to 1.

Inference using Joint Distributions

E.g. Suppose *Happy* and *Rested* are the random variables:

	Happy$=$ *true*	*Happy*$=$ *false*
Rested $=$ *true*	0.05	0.1
Rested $=$ *false*	0.6	0.25

The *unconditional probability* of any proposition is computable as the sum of entries from the full joint distribution

- E.g. $P(Happy) = P(Happy, Rested) + P(Happy, \neg Rested) = 0.65$

Conditional Probability

- The basic statements in the Bayesian framework talk about *conditional probabilities*.
  - $P(A|B)$ is the belief in event A given that event B is known with certainty
- The *product rule* gives an alternative formulation:

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

  - Note: we often write $P(A, B)$ as a shorthand for $P(A \wedge B)$

Chain Rule

Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
& P(X_1, \ldots, X_n) = \\
= \quad & P(X_1, \ldots, X_{n-1}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & P(X_1, \ldots, X_{n-2}) P(X_{n-1} | X_1, \ldots, X_{n-2}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & \ldots \\
= \quad & \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\end{aligned}
$$

Bayes Rule

- *Bayes rule* is another alternative formulation of the product rule:

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

- The *complete probability formula* states that:

$$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$$

or more generally,

$$P(A) = \sum_i P(A|b_i)P(b_i),$$

where b_i form a set of exhaustive and mutually exclusive events.

Conditional vs unconditonal probability

- Bertrand's coin box problem
- What are we supposed to compute?

Using Bayes Rule for Inference

- Suppose we want to form a hypothesis about the world based on observable variables.
- Bayes rule tells us how to calculate the belief in a hypothesis H given evidence e: $P(H|e) = (P(e|H)P(H))/P(e)$
  - $P(H|e)$ is the *posterior probability*
  - $P(H)$ is the *prior probability*
  - $P(e|H)$ is the *likelihood*
  - $P(e)$ is a *normalizing constant*, which can be computed as:

$$P(e) = P(e|H)P(H) + P(e|\neg H)P(\neg H)$$

Sometimes we write $P(H|e) \propto P(e|H)P(H)$

Example: Medical Diagnosis

- You go to the doctor complaining about the symptom of having a fever (*evidence*).
- The doctor knows that bird flu causes a fever 95% of the time.
- The doctor knows that if a person is selected randomly from the population, there is a 10^{-7} chance of the person having bird flu.
- In general, 1 in 100 people in the population suffer from fever.
- What is the probability that you have bird flu (*hypothesis*)?

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)} = \frac{0.95 \times 10^{-7}}{0.01} = 0.95 \times 10^{-5}$$

Computing Conditional Probabilities

- Typically, we are interested in the *posterior joint distribution* of some *query variables* Y given specific values e for some *evidence variables* E
- Let the *hidden variables* be $Z = X - Y - E$
- If we have a joint probability distribution, we can compute the answer by "summing out" the hidden variables:

$$P(Y|e) \propto P(Y,e) = \sum_z P(Y,e,z) \quad \leftarrow \text{summing out}$$

Problem: the joint distribution is too big to handle!

Example

- Consider medical diagnosis, where there are 100 different symptoms and test results that the doctor could consider.
- A patient comes in complaining of fever, cough and chest pains.
- The doctor wants to compute the probability of pneumonia.
  - The probability table has $>= 2^{100}$ entries!
  - For computing the probability of a disease, we have to sum out over 97 hidden variables!

Independence of Random Variables

- Two random variables X and Y are *independent* if knowledge about X does not change the uncertainty about Y (and vice versa):

$$\begin{aligned} P(x|y) &= P(x) \quad \forall x \in S_X, y \in S_Y \\ P(y|x) &= P(y) \quad \forall x \in S_X, y \in S_Y \end{aligned}$$

  or equivalently, $P(x, y) = P(x)P(y)$

- If n Boolean variables are independent, the whole joint distribution can be computed as:

$$P(x_1, \ldots x_n) = \prod_i P(x_i)$$

- Only n numbers are needed to specify the joint, instead of 2^n

Problem: Absolute independence is a very strong requirement

Example: Dice

- Let A be the event that one die comes to 1
- Let B be the event that a different die comes to 4
- Let C be the event of the sum of the two dice being 5
- Are A and B independent?
- Are A and C independent?
- Are A, B and C all independent?

Conditional Independence

- Two variables X and Y are *conditionally independent* given Z if:

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

- This means that knowing the value of Y does not change the prediction about X *if the value of Z is known*.

Example

- Consider a patient with three random variables: B (patient has bronchitis), F (patient has fever), C (patient has a cough)
- The full joint distribution has $2^3 - 1 = 7$ independent entries
- If someone has bronchitis, we can assume that, the probability of a cough does *not* depend on whether they have a fever:

$$P(C|B,F) = P(C|B) \tag{1}$$

  I.e., C is *conditionally independent* of F given B
- The same independence holds if the patient does not have bronchitis:

$$P(C|\neg B,F) = P(C|\neg B) \tag{2}$$

  therefore C and F are conditionally independent given B

Example (continued)

- The full joint distribution can now be written as:

$$\begin{aligned} & P(C,F,B) = \\ = \quad & P(C,F|B)P(B) \\ = \quad & P(C|B)P(F|B)P(B) \end{aligned}$$

I.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove two numbers)

Much more important savings happen if the system has lots of variables!

Naive Bayesian Model

- A common assumption in early diagnosis is that the symptoms are independent of each other given the disease
- Let $s_1, \ldots s_n$ be the symptoms exhibited by a patient (e.g. fever, headache etc)
- Let D be the patient's disease
- Using the naive Bayes assumption:

$$P(D, s_1, \ldots s_n) = P(D)P(s_1|D) \cdots P(s_n|D)$$

Recursive Bayesian Updating

The naive Bayes assumption allows incremental updating of beliefs as more evidence is gathered

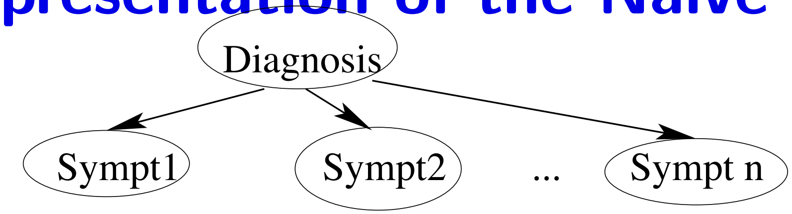
- Suppose that after knowing symptoms $s_1, \dots s_n$ the probability of D is: $P(D|s_1 \dots s_n) = P(D) \prod_{i=1}^{n} P(s_i|D)$
- What happens if a new symptom s_{n+1} appears?

$$\begin{aligned} P(D|s_1 \dots s_n, s_{n+1}) &= P(D) \prod_{i=1}^{n+1} P(s_i|D) \\ &= P(D|s_1 \dots s_n) P(s_{n+1}|D) \end{aligned}$$

- An even nicer formula can be obtained by taking logs:

$$\log P(D|s_1 \dots s_n, s_{n+1}) = \log P(D|s_1 \dots s_n) + \log P(s_{n+1}|D)$$

A Graphical Representation of the Naive Bayesian Model



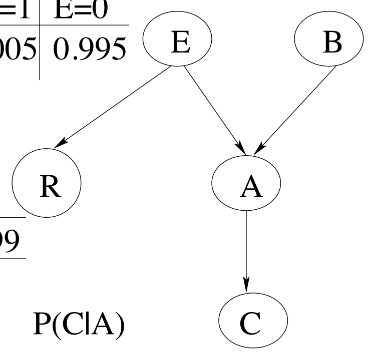
- The nodes represent random variables
- The arcs represent "influences"
- The *lack of arcs* represents conditional independence

How about models represented as more general graphs?

Example (Adapted from Pearl)

- I'm at work, and my neighbor calls to say the burglar alarm is ringing in my house. Sometimes the alarm is set off by minor earthquakes. Earthquakes are sometimes reported on the radio. Should I rush home?
- What random variables are there?
- Are there any independence relationships between these variables?

Example: A Belief (Bayesian) Network



P(E)

E=1	E=0
0.005	0.995

P(B)

B=1	B=0
0.01	0.99

P(R|E)

	R=1	R=0
E=0	0.0001	0.9999
E=1	0.65	0.35

P(A|B,E)

	A=1	A=0
B=0,E=0	0.001	0.999
B=0,E=1	0.3	0.7
B=1,E=0	0.8	0.2
B=1,E=1	0.95	0.05

P(C|A)

	C=1	C=0
A=0	0.05	0.95
A=1	0.7	0.3

- The nodes represent random variables
- The arcs represent "direct influences"
- At each node, we have a conditional probability distribution (CPD) for the corresponding variable *given its parents*

Using a Bayes Net for Reasoning (1)

Computing any entry in the joint probability table is easy:

$$\begin{aligned} P(B, \neg E, A, C, \neg R) &= P(B)P(\neg E)P(A|B, \neg E)P(C|A)P(\neg R|\neg E) \\ &= 0.01 \cdot 0.995 \cdot 0.8 \cdot 0.7 \cdot 0.9999 \approx 0.0056 \end{aligned}$$

What is the probability that a neighbor calls?

$$P(C = 1) = \sum_{e,b,r,a} P(C = 1, e, b, r, a) = \ldots$$

What is the probability of a call in case of a burglary?

$$P(C = 1|B = 1) = \frac{P(C = 1, B = 1)}{P(B = 1)} = \frac{\sum_{e,r,a} P(C = 1, B = 1, e, r, a)}{\sum_{c,e,r,a} P(c, B = 1, e, r, a)}$$

This is *causal reasoning* or *prediction*

Using a Bayes Net for Reasoning (2)

- Suppose we got a call.
- What is the probability of a burglary?

$$P(B|C) = \frac{P(C|B)P(B)}{P(C)} = ...$$

- What is the probability of an earthquake?

$$P(E|C) = \frac{P(C|E)P(E)}{P(C)} = ...$$

This is *evidential reasoning* or *explanation*

Using a Bayes Net for Reasoning (3)

- What happens to the probabilities if the radio announces an earthquake?

$$P(E|C,R) \gg P(E|C) \text{ and } P(B|C,R) \ll P(B|C)$$

This is called *explaining away*

Example: Pathfinder (Heckerman, 1991)

- Medical diagnostic system for lymph node diseases
- Large net!
  - 60 diseases, 100 symptoms and test results, 14000 probabilities
- Network built by medical experts
  - 8 hours to determine the variables
  - 35 hours for network topology
  - 40 hours for probability table values
- Experts found it easy to invent causal links and probabilities
- Pathfinder is now <u>*outperforming world experts*</u> in diagnosis
- Being extended now to other medical domains

Using Graphs for Knowledge Representation

- Graphs have been proposed as models of human memory and reasoning on many occasions (e.g. semantic nets, inference networks, conceptual dependencies)
- There are many efficient algorithms that work with graphs, and efficient data structures
- Recently there has been an explosion of *graphical models* for representing probabilstic knowledge
- Lack of edges is assumed to represent conditional independence assumptions
- Graph properties are used to do *inference* (i.e. compute conditional probabilities) efficiently using such representations