Uncertainty and Bayesian Networks

Tutorial 3
Outline

◊ Uncertainty
◊ Probability
◊ Syntax and Semantics for Uncertainty
◊ Inference
◊ Independence and Bayes’ Rule
◊ Syntax and Semantics for Bayesian Networks
◊ Parameterized distributions
◊ Exact inference by enumeration
◊ Exact inference by variable elimination
Uncertainty

Let action $A_t = \text{leave for airport } t \text{ minutes before flight}$
Will $A_t$ get me there on time?

Problems:
1) partial observability (road state, other drivers’ plans, etc.)
2) noisy sensors (KCBS traffic reports)
3) uncertainty in action outcomes (flat tire, etc.)
4) immense complexity of modelling and predicting traffic

Hence a purely logical approach either
1) risks falsehood: “$A_{25}$ will get me there on time”
or 2) leads to conclusions that are too weak for decision making:
   “$A_{25}$ will get me there on time if there’s no accident on the bridge
   and it doesn’t rain and my tires remain intact etc etc.”

($A_{1440}$ might reasonably be said to get me there on time
but I’d have to stay overnight in the airport . . .)
Methods for handling uncertainty

Probabilistic assertions summarize effects of
laziness: failure to enumerate exceptions, qualifications, etc.
ignorance: lack of relevant facts, initial conditions, etc.

Subjective or Bayesian probability:
Probabilities relate propositions to one’s own state of knowledge
e.g., \( P(A_{25} | \text{no reported accidents}) = 0.06 \)
These are not claims of a “probabilistic tendency” in the current situation
(but might be learned from past experience of similar situations)

Probabilities of propositions change with new evidence:
e.g., \( P(A_{25} | \text{no reported accidents, 5 a.m.}) = 0.15 \)
(Analogous to logical entailment status \( KB \models \alpha \), not truth.)
Making decisions under uncertainty

Suppose I believe the following:

\[ P(A_{25} \text{ gets me there on time} | \ldots) = 0.04 \]
\[ P(A_{90} \text{ gets me there on time} | \ldots) = 0.70 \]
\[ P(A_{120} \text{ gets me there on time} | \ldots) = 0.95 \]
\[ P(A_{1440} \text{ gets me there on time} | \ldots) = 0.9999 \]

Which action to 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
Probability basics

Begin with a set $\Omega$—the sample space
e.g., 6 possible rolls of a die.
$\omega \in \Omega$ is a sample point/possible world/atomic event

A probability space or probability model is a sample space
with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

\[
0 \leq P(\omega) \leq 1 \\
\sum_\omega P(\omega) = 1
\]
e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event $A$ is any subset of $\Omega$

\[
P(A) = \sum_{\omega \in A} P(\omega)
\]
E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$
Propositions

Think of a proposition as the event (set of sample points) where the proposition is true

Given Boolean random variables $A$ and $B$:
- event $a$ = set of sample points where $A(\omega) = true$
- event $\neg a$ = set of sample points where $A(\omega) = false$
- event $a \land b$ = points where $A(\omega) = true$ and $B(\omega) = true$

Often in AI applications, the sample points are defined by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables.

With Boolean variables, sample point = propositional logic model
- e.g., $A = true$, $B = false$, or $a \land \neg b$.

Proposition = disjunction of atomic events in which it is true
- e.g., $(a \lor b) \equiv (\neg a \land b) \lor (a \land \neg b) \lor (a \land b)$
- $\Rightarrow P(a \lor b) = P(\neg a \land b) + P(a \land \neg b) + P(a \land b)$
Syntax for propositions

Propositional or Boolean random variables
  e.g., $Cavity$ (do I have a cavity?)
  $Cavity = \text{true}$ is a proposition, also written cavity

Discrete random variables (finite or infinite)
  e.g., $Weather$ is one of $\langle \text{sunny, rain, cloudy, snow} \rangle$
  $Weather = \text{rain}$ is a proposition
  Values must be exhaustive and mutually exclusive

Continuous random variables (bounded or unbounded)
  e.g., $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.

Arbitrary Boolean combinations of basic propositions
Prior or unconditional probabilities of propositions

e.g., \( P(Cavity = true) = 0.1 \) and \( P(Weather = sunny) = 0.72 \)
correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

\[ P(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle \text{ (normalized, i.e., sums to 1)} \]

Joint probability distribution for a set of r.v.s gives the
probability of every atomic event on those r.v.s (i.e., every sample point)

\[ P(Weather, Cavity) = \text{a } 4 \times 2 \text{ matrix of values:} \]

\[
\begin{array}{c|cccc}
Weather = & sunny & rain & cloudy & snow \\
Cavity = true & 0.144 & 0.02 & 0.016 & 0.02 \\
Cavity = false & 0.576 & 0.08 & 0.064 & 0.08 \\
\end{array}
\]

Every question about a domain can be answered by the joint
distribution because every event is a sum of sample points
Definition of conditional probability:

\[ P(a|b) = \frac{P(a \land b)}{P(b)} \quad \text{if} \quad P(b) \neq 0 \]

**Product rule** gives an alternative formulation:

\[ P(a \land b) = P(a|b)P(b) = P(b|a)P(a) \]

A general version holds for whole distributions, e.g.,

\[ P(\text{Weather, Cavity}) = P(\text{Weather}|\text{Cavity})P(\text{Cavity}) \]

(View as a \(4 \times 2\) set of equations, **not** matrix mult.)

**Chain rule** is derived by successive application of product rule:

\[
P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1}) P(X_n|X_1, \ldots, X_{n-1}) \\
= P(X_1, \ldots, X_{n-2}) P(X_{n-1}|X_1, \ldots, X_{n-2}) P(X_n|X_1, \ldots, X_{n-1}) \\
= \ldots \\
= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})
\]
Independence

$A$ and $B$ are independent iff

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A)P(B)$$

$$P(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather})$$

$$= P(\text{Toothache}, \text{Catch}, \text{Cavity})P(\text{Weather})$$

$\text{Catch}$ is conditionally independent of $\text{Toothache}$ given $\text{Cavity}$:

$$P(\text{Catch}|\text{Toothache}, \text{Cavity}) = P(\text{Catch}|\text{Cavity})$$

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in $n$ to linear in $n$.

Conditional independence is our most basic and robust form of knowledge about uncertain environments.
Bayes’ Rule

Product rule \( P(a \land b) = P(a|b)P(b) = P(b|a)P(a) \)

\[ \Rightarrow \text{Bayes’ rule} \quad P(a|b) = \frac{P(b|a)P(a)}{P(b)} \]

or in distribution form

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y) \]

Useful for assessing diagnostic probability from causal probability:

\[ P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)} \]

E.g., let \( M \) be meningitis, \( S \) be stiff neck:

\[ P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008 \]

Note: posterior probability of meningitis still very small!
Bayesian networks

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

Syntax:
- a set of nodes, one per variable
- a directed, acyclic graph (link $\approx$ “directly influences”)
- a conditional distribution for each node given its parents:

\[ P(X_i|Parents(X_i)) \]

In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over $X_i$ for each combination of parent values
A CPT for Boolean $X_i$ with $k$ Boolean parents has $2^k$ rows for the combinations of parent values.

Each row requires one number $p$ for $X_i = true$ (the number for $X_i = false$ is just $1 - p$).

If each variable has no more than $k$ parents, the complete network requires $O(n \cdot 2^k)$ numbers.

I.e., grows linearly with $n$, vs. $O(2^n)$ for the full joint distribution.

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)
"Global" semantics defines the full joint distribution as the product of the local conditional distributions:

\[ P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{parents}(X_i)) \]

e.g., \[ P(j \land m \land a \land \neg b \land \neg e) \]
\[ = P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \]
\[ = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \]
\[ \approx 0.00063 \]
**Global semantics**

- **Burglary**
  - $P(B)$ = 0.001

- **Earthquake**
  - $P(E)$ = 0.002

- **Alarm**
  - $P(A)$
    - $t$ $t$ = 0.95
    - $t$ $f$ = 0.94
    - $f$ $t$ = 0.29
    - $f$ $f$ = 0.001

- **JohnCalls**
  - $P(J)$
    - $t$ = 0.90
    - $f$ = 0.05

- **MaryCalls**
  - $P(M)$
    - $t$ = 0.70
    - $f$ = 0.01
Local semantics: each node is conditionally independent of its nondescendants given its parents

Theorem: Local semantics ⇔ global semantics
Markov blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children’s parents
Constructing Bayesian networks

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables \(X_1, \ldots, X_n\)
2. For \(i = 1\) to \(n\)
   - add \(X_i\) to the network
   - select parents from \(X_1, \ldots, X_{i-1}\) such that
     \[
     P(X_i|\text{Parents}(X_i)) = P(X_i|X_1, \ldots, X_{i-1})
     \]

This choice of parents guarantees the global semantics:

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}
\]
\[
= \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \quad \text{(by construction)}
\]
Inference by enumeration

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:
\[ P(B|j, m) = \frac{P(B, j, m)}{P(j, m)} = \alpha P(B, j, m) = \alpha \sum_e \sum_a P(B, e, a, j, m) \]

Rewrite full joint entries using product of CPT entries:
\[ P(B|j, m) = \alpha \sum_e \sum_a P(B)P(e)P(a|B, e)P(j|a)P(m|a) = \alpha P(B) \sum_e P(e) \sum_a P(a|B, e)P(j|a)P(m|a) \]

Recursive depth-first enumeration: \( O(n) \) space, \( O(d^m) \) time
Variable elimination: carry out summations right-to-left, storing intermediate results (factors) to avoid recomputation

\[
\begin{align*}
P(B|j, m) &= \alpha P(B) \sum_{e} P(e) \sum_{a} P(a|B, e) P(j|a) P(m|a) \\
&= \alpha P(B) \sum_{e} P(e) \sum_{a} P(a|B, e) P(j|a) f_M(a) \\
&= \alpha P(B) \sum_{e} P(e) \sum_{a} P(a|B, e) f_J(a) f_M(a) \\
&= \alpha P(B) \sum_{e} P(e) \sum_{a} f_A(a, b, e) f_J(a) f_M(a) \\
&= \alpha P(B) \sum_{e} P(e) f_{\bar{A}JM}(b, e) \quad \text{(sum out } A) \\
&= \alpha P(B) f_{\bar{E}AJM}(b) \quad \text{(sum out } E) \\
&= \alpha f_B(b) \times f_{\bar{E}AJM}(b)
\end{align*}
\]
Complexity of exact inference

Singly connected networks (or polytrees):
- any two nodes are connected by at most one (undirected) path
- time and space cost of variable elimination are $O(d^kn)$

Multiply connected networks:
- can reduce 3SAT to exact inference $\Rightarrow$ NP-hard
- equivalent to counting 3SAT models $\Rightarrow$ #P-complete

\[
\begin{array}{cccc}
A & B & C & D \\
1 & 2 & 3 & \\
3 & v & C \\
\top & v & \neg A \\
\top & v & \neg D \\
\end{array}
\]
Summary

Probability is a rigorous formalism for uncertain knowledge

Joint probability distribution specifies probability of every atomic event

Queries can be answered by summing over atomic events

For nontrivial domains, we must find a way to reduce the joint size

Independence and conditional independence provide the tools

Bayes nets provide a natural representation for (causally induced) conditional independence

Topology + CPTs = compact representation of joint distribution

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Continuous variables ⇒ parameterized distributions (e.g., linear Gaussian)
Exercises

♦ Show from first principles that $P(a|b \land a) = 1$.

♦ Prove that any 3-SAT problem can be reduced to exact inference in a Bayesian network constructed to represent the particular problem and hence that exact inference is NP-hard.
1. The "first principles" needed here are the definition of conditional probability, \( P(X|Y) = \frac{P(X \land Y)}{P(Y)} \), and the definitions of the logical connectives. It is not enough to say that if \( B \land A \) is "given" then \( A \) must be true! From the definition of conditional probability, and the fact that \( A \land A \iff A \) and that conjunction is commutative and associative, we have

\[
P(A|B \land A) = \frac{P(A \land (B \land A))}{P(B \land A)} = \frac{P(B \land A)}{P(B \land A)} = 1
\]

2. Consider a SAT problem such as the following:

\[
(\neg A \lor B) \land (\neg B \lor C) \land (\neg C \lor D) \land (\neg C \lor \neg D \lor E)
\]

The idea is to encode this as a Bayes net, such that doing inference in the Bayes net gives the answer to the SAT problem.
The figure shows the Bayes net corresponding to this SAT problem. The general construction method is as follows:

The root nodes correspond to the logical variables of the SAT problem. They have a prior probability of 0.5.

Each clause $C_i$ is a node. Its parents are the variables in the clause. The CPT is deterministic and implements the disjunction given in the clause. (Negative literals in the clause are indicated by negation symbols on the links in the figure.)
A single sentence node $S$ has all the clauses as parents and a CPT that implements deterministic conjunction.

It is clear that $P(S) > 0$ iff the SAT problem is satisfiable. Hence, we have reduced SAT to Bayes net inference. Since SAT is NP-complete, we have shown that Bayes net inference is NP-hard (even without evidence).