Probabilistic Reasoning in AI

COMP-526

Instructor: Doina Precup

Email: dprecup@cs.mcgill.ca

TA: Rohan Shah

Email: rshah3@cs.mcgill.ca

Class web page

http://www.cs.mcgill.ca/∼dprecup/courses/prob.html

Outline

- Administrative details
- Dealing with uncertainty
- Probability
- Probabilistic reasoning
- Decision making under uncertainty

Administrative issues

- Class material:
  - M. Jordan, Introduction to Graphical Models
  - Notes by D. Koller and N. Friedman
  - R.S.Sutton and A.G.Barto,
    Reinforcement Learning: An Introduction
  - Class notes: posed on the web page
  - Additional readings: TBA
- Evaluation mechanisms:
  - Roughly 10-11 assignments (70%)
  - Two written examinations (15% each)
  - Class participation and discussions (up to 5% extra credit)
- Programming assignments must **function** to get credit

Uncertainty

Uncertainty is inherent in many tasks

E.g. Will leaving home t minutes before the flight get me to the airport on time?

- Partial knowledge of the state of the world

  E.g. We do not know the road state, other drivers' plans etc.
- Noisy observations

  E.g. Traffic reports
- Inherent stochasticity

  E.g. Flat tires, accidents etc.
- Phenomena that are not covered by our models

  E.g. the complexity of predicting traffic

How do we deal with uncertainty?

- Implicit methods: ignore uncertainty as much as possible, build procedures that are robust to uncertainty
- *Explicit (model-based) methods*
  - Build a *model of the world* that describes the uncertainty about the system state, dynamics and about our observations
  - Reason about the effect of actions given the model

We will focus mainly on explicit model-based methods

How do we 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?

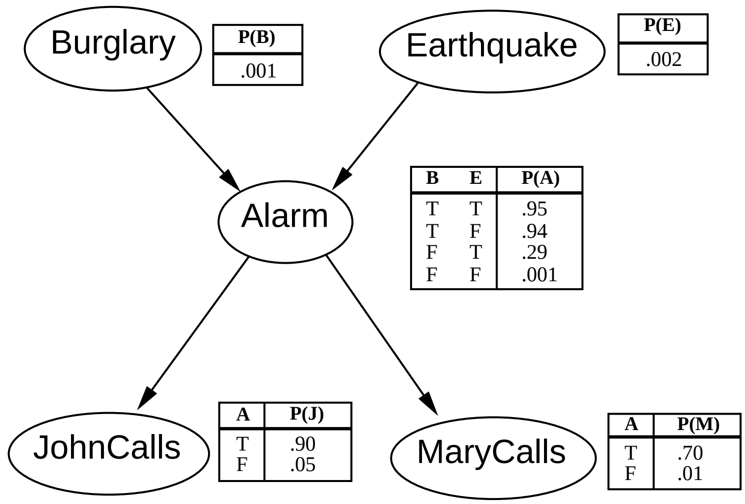
Probability

- 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
- Arguably intuitive to human experts

Representing probabilities efficiently

- Naive representations of probability are hopelessly inefficient
  E.g. consider patients described by several attributes:
  - background: age, gender, medical history,...
  - Symptoms: fever, blood pressure, headache,...
  - Diseases: pneumonia, hepatitis,...
- A probability distribution needs to assign a number to each combination of values of these attributes!
- Real examples involve hundreds of attributes
- **Key idea:** exploit regularities and structure of the domain
- We will focus mainly on exploiting **conditional independence** properties

A Bayesian network



P(B)
.001

P(E)
.002

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

A	P(J)
T	.90
F	.05

A	P(M)
T	.70
F	.01

Probabilistic Reasoning

During the first half of the course we will study:

- Syntax and semantics of Bayesian networks
- How to efficiently answer queries in a Bayesian network
- How to learn Bayesian networks from data
- How to extend Bayesian networks in order to represent properties of sequences and temporal processes

Fielded applications

- Expert systems
  - Medical diagnosis (e.g. Pathfinder)
  - Fault diagnosis (e.g. jet-engines)
- Monitoring
  - Space shuttle engines (Vista project)
  - Freeway traffic
- Sequence analysis and classification
  - Speech recognition
  - Biological sequences
- Information access
  - Collaborative filtering
  - Information retrieval

Decision making

- Probability is not enough for choosing actions
- We also need to consider *risks and payoffs*
  E.g. Suppose I believe the following:

$$\begin{aligned} P(A_{25} \text{ gets me there on time} | \ldots) &= 0.04 \\ P(A_{90} \text{ gets me there on time} | \ldots) &= 0.70 \\ P(A_{120} \text{ gets me there on time} | \ldots) &= 0.95 \\ P(A_{1440} \text{ gets me there on time} | \ldots) &= 0.9999 \end{aligned}$$

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

Practical decision making

- We need to represent both probabilities and utilities
- The **expected utility** of actions is computed given evidence and past actions
- We choose the action that maximizes expected utility
- **Value of information:** is it worth acquiring more information in order to choose better actions?

Decision making

In the second half of the course we will study:

- Utility theory
- Models of repeated decision: Markov Decision Processes
- Partially Observable Markov Decision Processes
- Learning to act optimally

Related fields

- Artificial Intelligence
- Machine learning
- Operations research
- Decision theory
- Statistics
- Information theory
- ...

Discrete random variables and probability

- A **discrete random variable** X describes an outcome that cannot be determined in advance (e.g. the roll of a die)
- The **sample space** S of a random variable X is the set of all possible values of the variable
  E.g. For a die, $S = \{1, 2, 3, 4, 5, 6\}$
- An **event** is a subset of S. E.g. $e = \{1\}$ corresponds to a die roll of 1
- 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, $p(1) = p(2) = \ldots = p(6) = 1/6$
  Note that we still cannot determine the value of X, just the chance of encountering a given value

Discrete random variables

If X is a discrete variable, then a probability space $p(x)$ has the following properties:

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

Terminology

- The **n-th moment of a random variable** X is defined as:

$$M_n = \sum_{x \in S} x^n p(x)$$

- The first moment is called the **expectation** or **mean**:

$$E\{x\} = M_1 = \sum_{x \in S} x p(x)$$

E.g. for a roll with a fair die, the expectation is:

$$M_1 = \sum_{x \in \{1,2,3,4,5,6\}} x \frac{1}{6} = 3.5$$

As illustrated above, the expectation is **not** the value we expect to see the most!

And more terminology...

- The **variance** is defined as:

$$Var\{x\} = M_2 - M_1^2 = E\{x^2\} - E\{x\}^2$$

- The **standard deviation** $\sigma = \sqrt{Var\{x\}}$ evaluates the "spread" of x with respect to its mean

Beliefs

- We will use probability in order to describe the world and the existing uncertainties
- **Beliefs** (also called Bayesian or subjective probabilities) relate logical propositions to the current state of knowledge
- Beliefs are **subjective** assertions about the world, given one's 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
- **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. $P(True) = 1$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$, or equivalently,
   $P(A \vee B) = P(A) + P(B)$ if A and B are mutually exclusive

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

Using functions that disobey these laws as probabilities can force suboptimal decisions (de Finetti, 1931).

Defining probabilistic models

- We define the world as a set of random variables $\Omega = \{X_1 \ldots X_n\}$.
- A **probabilistic model** is an encoding of probabilistic information that allows us to compute the probability of any event in the world

A simple probabilistic model:

- We divide the world 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.

$$P(Happy) = P(Happy, Rested) + P(Happy, \neg Rested) = 0.65$$