Probabilistic Reasoning in AI

COMP-526

Instructor: Doina Precup

Email: dprecup@cs.mcgill.ca

TA: Pablo Samuel Castro

Email: pcastr@cs.mcgill.ca

Class web page

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

Outline

- Administrative details
- Course overview
  - Modeling uncertainty using probabilities
  - Decision making under uncertainty
- Random variables and probabilities
- Conditional probability and Bayes rule

Class reading

- Selected chapters from the following books:
  - M. Jordan, *Introduction to Graphical Models* (copies will be distributed, pending approval)
  - C. Bishop, *Pattern Recognition and Machine Learning*
  - D.J.C.MacKay, *Information theory, inference and learning algorithms*
  - S. Russell and P. Norvig: *Artificial intelligence: A modern approach*
  - R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*
- Class notes: posted on the web page

Evaluation mechanisms

- Seven assignments (70%)
  - Problems related to the class material
  - Programming exercises
- One written examination for the first part of the course (15%)
- One project for the second part of the course (15%)
- Class participation and discussions (up to 5% extra credit)

Intelligent systems have to deal with uncertainty!

E.g. Predicting the behavior of other people (girlfriend / boyfriend / kids)

- Partial knowledge of the state of the world

  E.g. We don't know exactly what is going on in their mind
- Noisy observations

  E.g. Smiling or frowning, making faces
- Inherent stochasticity

  E.g. Today she likes the DVD, tomorrow she does not!
- Phenomena that are not covered by our models

  E.g. Level of hormones, which depend on food, exercise, ...

How do we deal with uncertainty?

- In this course, we will focus on building **models** that capture the uncertainty about the state of the world, the dynamics of the system and about our observations.
- The model will then be used to **reason** about the world, and about the effects of different actions.
- Important questions:
  - What mathematical formalism should we use? What is the meaning of our model?
  - What queries can the model answer? What is the method for answering queries?
  - How do we construct a model? Do we need to ask an expert, or can the model be learned from data?

AI approaches for dealing with uncertainty

- Default reasoning: believe something until evidence to the contrary is found
- Rules with “fudge factors”
- Fuzzy logic: allows events to be “sort of true”
- **Probability**

The dawning of the age of stochasticity

For over two millennia, Aristotle’s logic has ruled over the thinking of western intellectuals. All precise theories, all scientific models, even models of the process of thinking itself, have in principle conformed to the straight-jacket of logic. But from its shady beginnings devising gambling strategies and counting corpses in medieval London, probability theory and statistical inference now emerge as better foundations for scientific models, especially those of the process of thinking, and as essential ingredients of theoretical mathematics, even the foundations of mathematics itself. We propose that this sea of change in our perspective will affect virtually all of mathematics in the next century.

David Mumford, 1999

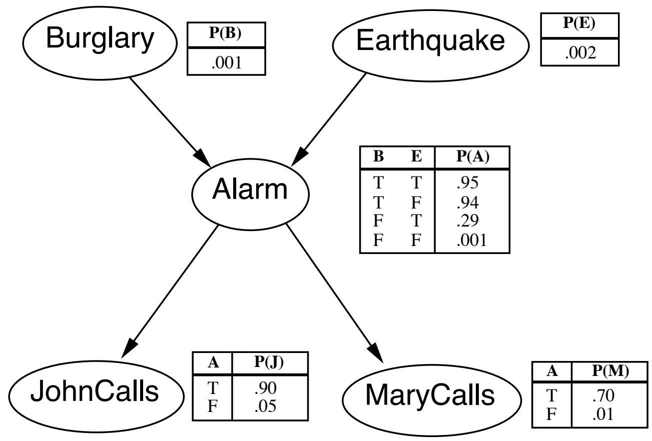
Probability

- 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

Example: A Bayesian (belief) network



Probabilistic reasoning

During the first half of the course we will study:

- Syntax and semantics of graphical models for representing probability distributions:
  - Bayesian networks (built on top of directed graphs)
  - Markov networks (built on top of undirected graphs)
- How to efficiently answer queries in such networks
- How to learn Bayesian and Markov networks from data
- How to model time sequences using graphical models
  - Hidden Markov Models (HMMs)
  - Dynamic Bayesian Networks (DBNs)
  - Kalman filters

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

- Suppose you believe the following:

$$\begin{aligned} p(\text{girlfriend likes Terminator DVD}|\ldots) &= 0.04 \\ p(\text{girlfriend likes cashmere sweater}|\ldots) &= 0.70 \\ p(\text{girlfriend likes diamond necklace}|\ldots) &= 0.9999 \end{aligned}$$

Which action should you choose?

Decision making

- Suppose you believe the following:

$$\begin{aligned} p(\text{girlfriend likes Terminator DVD}|\ldots) &= 0.04 \\ p(\text{girlfriend likes cashmere sweater}|\ldots) &= 0.70 \\ p(\text{girlfriend likes diamond necklace}|\ldots) &= 0.9999 \end{aligned}$$

  Which action should you choose?

  Depends on **preferences** for making her happy vs. your own interest vs. cost

- Probability is not enough for choosing actions
- We also need to consider **risks and payoffs**

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
- A “rational” agent should 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

Fielded applications

- Robot control
- Control of complex, chaotic systems (e.g. helicopters)
- Game playing
- Inventory management
- Allocation of bandwidth in cell phone networks
- Network routing
- ...

What is a random variable?

- Something that has not happened yet:
  - Will a tossed coin come up heads or tails?
  - Will the cancer recur or not?
- Something you do not know ...
  - Did the coin come up heads or tails?
  - How did the protein fold?

  ...because you have not/cannot observe it directly or compute it definitively from what you have observed.

Discrete random variables

A **discrete random variable** X takes values from a discrete set Ω_X, called the **domain** or **sample space** of X.

- X = roll of a die; $\Omega_X = \{1, 2, 3, 4, 5, 6\}$.
- X = nucleotide a position 1, chromosome 1, in a particular person; $\Omega_X = \{A, C, G, T\}$.
- X = does a customer buy a new TV or not

An **event** is a subset of Ω_X.

- $e_1 = \{1\}$ corresponds to a die roll of 1
- $e_2 = \{1, 3, 5\}$ corresponds to an odd value for the roll

Probabilities

- For a discrete r.v. X, each value $x \in \Omega_X$ has a probability of occurring, which we will denote by $p(X = x)$ or, more simply, $p(x)$.
- $p(X)$ denotes the **probability distribution function** (p.d.f.) for X. It can be thought of as a table.

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

- Basic properties:
  - $0 \leq p(x) \leq 1, \forall x \in \Omega_X$
  - $\sum_{x \in \Omega_X} p(x) = 1$

Cumulative distribution functions

- If X takes values from an ordered set Ω_X (such as integers) then the **cumulative distribution function** is

$$\text{c.d.f.}(x) = p(X \leq x) = \sum_{x' \leq x} p(x')$$

- For example, if X is the roll of a die, then:

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6
c.d.f.(x)	1/6	2/6	3/6	4/6	5/6	1

Mean and variance

- If Ω_X is a set of numbers, then the **expected value** or **mean** of X is

$$E(X) = \sum_{x \in \Omega_X} x p(x)$$

- The **variance** of X is

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= \left(\sum_{x \in \Omega_X} x^2 p(x) \right) - \left(\sum_{x \in \Omega_X} x p(x) \right)^2 \end{aligned}$$

- The **standard deviation** of X is the square root of the variance
- Example: If X is a die roll, then the mean value is 3.5 and the standard deviation is approximately 3.4157.

Continuous random variables

- A **continuous random variable** X takes real values.
  - $X =$ expression level for a gene as reported by a microarray.
  - $X =$ price for which a house will sell
  - $X =$ size of a tumor
- Any continuous r.v. X has a **cumulative distribution function**

$$\text{c.d.f.}(x) = p(X \leq x)$$

with the following properties:
  - c.d.f.(x) is a non-decreasing function; c.d.f.$(x) \leq$ c.d.f.(x') whenever $x \leq x'$.
  - $\lim_{x \to -\infty}$ c.d.f$(x) = 0$.
  - $\lim_{x \to +\infty}$ c.d.f.$(x) = 1$.

Probability density functions

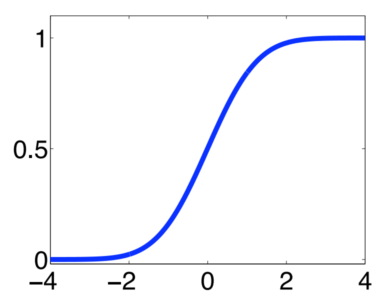
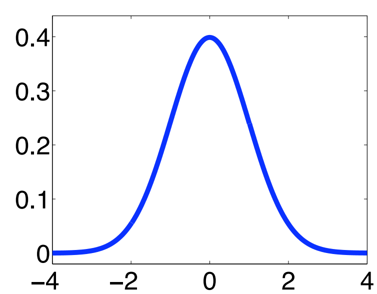
- If c.d.f(x) is continuous and differentiable then its derivative is the **probability density function**, analogous to the probability distribution function of a discrete r.v.

$$\frac{d}{dx}\text{c.d.f}(x) = p(x)$$

- Properties:
  - $0 \le p(x) < \infty$. Note that $p(x) > 1$ is allowed, unlike for discrete r.v.'s.
  - $\int_x p(x)dx = 1$, similar to discrete r.v.'s.

Example: Gaussian random variables

$X \sim N(\mu, \sigma)$ has mean μ and standard deviation σ.

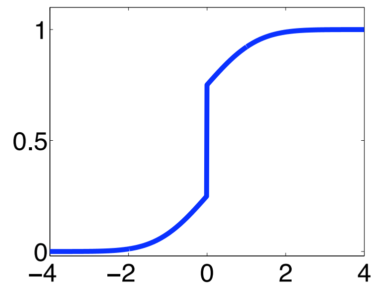
	c.d.f.	p.d.f.
graph		
formula	?	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Not all continuous r.v.'s have p.d.f.'s

- Suppose X equal to zero with probability $\frac{1}{2}$ and otherwise is distributed according to $N(0,1)$.
- Then the c.d.f. is

$$\text{c.d.f}(x) = \begin{cases} \frac{1}{2}f(x) & x < 0 \\ \frac{1}{2}f(x) + \frac{1}{2} & x \geq 0 \end{cases}$$

where $f(x)$ denotes the c.d.f. of a $N(0,1)$ r.v.



- There is no p.d.f. because of the discrete jump in the c.d.f.

Mean and variance

- We will almost always restrict our attention to continuous r.v.'s with p.d.f.'s.
- Then, the expected value is defined as

$$E(X) = \int_x xp(x)dx$$

- Variance is

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= \int_x x^2 p(x)dx - \left(\int_x xp(x)dx\right)^2 \end{aligned}$$

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.

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

Example

- Let $X_1 =$ true iff a rolled die comes out even.
- Let $X_2 =$ true iff the same rolled die comes out odd.

$$p(X_1 = \text{true}) = p(X_1 = \text{false}) = \frac{1}{2}$$
$$p(X_2 = \text{true}) = p(X_2 = \text{false}) = \frac{1}{2}$$

- What is the probability $p(X_1 = \text{true and } X_2 = \text{true})$?

Example

- Let $X_1 =$ true iff a rolled die comes out even.
- Let $X_2 =$ true iff the same rolled die comes out odd.

$$p(X_1 = \text{true}) = p(X_1 = \text{false}) = \frac{1}{2}$$
$$p(X_2 = \text{true}) = p(X_2 = \text{false}) = \frac{1}{2}$$

- What is the probability $p(X_1 = \text{true and } X_2 = \text{true})$?
- We know it is zero, but there is no way of knowing just from $p(X_1)$ and $p(X_2)$!

$\Rightarrow$ There are several ways to specify the **relationships** between variables. They all come down to specifying **joint probability distributions/densities**.

Joint probabilities

- When considering r.v.'s $X_1, X_2, \ldots, X_m$, the joint probability function specifies the probability of every combination of values.

$$p(X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \ldots \text{ and } X_m = x_m)$$

- When the r.v.'s are discrete, the joint probability can be viewed as a table.

	even=true	even=false
odd=true	0	1/2
odd=false	1/2	0

Marginal probabilities

Given r.v.'s $X_1, X_2, \ldots X_m$ with joint probability $p(x_1, x_2, \ldots, x_m)$, the **marginal probability** of a r.v. X_i is obtained by summing (or integrating) over all possible values of the other r.v.'s:

$$p(X_i = x_i) = \sum_{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m} p(x_1, x_2, \ldots, x_m)$$

	die=1	2	3	4	5	6	p(even)
even=true	0	1/6	0	1/6	0	1/6	1/2
even=false	1/6	0	1/6	0	1/6	0	1/2
p(die)	1/6	1/6	1/6	1/6	1/6	1/6	

A similar definition holds for any subset of r.v.'s.

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 binary variables A, B, a state will be a complete assignment of truth values for A and B.
- A **joint probability distribution function** provides the probabilities for each state
- Probabilities for other events which are not listed explicitly are computed by marginalization. This process is called **inference**.

Inference using joint distributions

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

	Happy $= 1$	Happy $= 0$
Rested $= 1$	0.05	0.1
Rested $= 0$	0.6	0.25

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

E.g. $p(\text{Happy} = 1) = p(\text{Happy} = 1, \text{Rested} = 1) + p(\text{Happy} = 1, \text{Rested} = 0) = 0.65$

Example

Suppose we consider medical diagnosis, and there are 100 different symptoms and test results that the doctor could consider. A patient comes in complaining of fever, dry 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!

Independent random variables

- Two random variables X and Y are independent, denoted $X \perp\!\!\!\perp Y$, if and only if

$$p(X = x \text{ and } Y = y) = p(X = x)p(Y = y)$$

  for all values x and y.
- This is often abbreviated as $p(X, Y) = p(X)p(Y)$.
- If this requirement is satisfied, for binary variables, only n numbers are necessary to represent the joint distribution, instead of 2^n!
- But this is a very strict requirement, almost never met.

Conditional probablity

- The basic statements in the Bayesian framework talk about **conditional probabilities**
- $p(X = x|Y = y)$ denotes the belief that event $X = x$ occurs **given that** event $Y = y$ has occurred with absolute certainty.
  - E.g., $p($die=1$|$odd $=$ true$) = 1/3$.
  - E.g., $p($die=1$|$odd $=$ false$) = 0$.
- The conditional probability can be defined (and computed) as

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

as long as $p(y) > 0$.

- An alternative formulation is given by the **product rule**:

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Bayes' Rule

- **Bayes rule** is a reformulation of the product rule:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

.

- Another alternative formulation is the **complete probability formula**:

$$p(x) = \sum_{y} p(x|y)p(y),$$

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

Chain rule

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

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

Why conditional probabilities?

Conditional probabilities are interesting because we often observe something and want to infer something/make a guess about something unobserved but related.

- $p($cancer recurs$|$tumor measurements$)$
- $p($gene expressed $> 1.3|$transcription factor concentrations$)$
- $p($collision to obstacle$|$sensor readings$)$
- $p($word uttered$|$sound wave$)$
- ...