Probabilistic Modelling and Reasoning

Chris Williams, Division of Informatics
University of Edinburgh

Course Introduction

- Welcome
- Administration
  - Handout
  - Books
  - Assignments
  - Tutorials
  - Course rep(s)

Relationships between courses

PMR Probabilistic modelling and reasoning. Focus on probabilistic modelling. Learning and inference for probabilistic models, e.g. Probabilistic expert systems, latent variable models, Hidden Markov models, Kalman filters, Boltzmann machines.

LfD1 Learning from Data 1. Basic introductory course on supervised and unsupervised learning

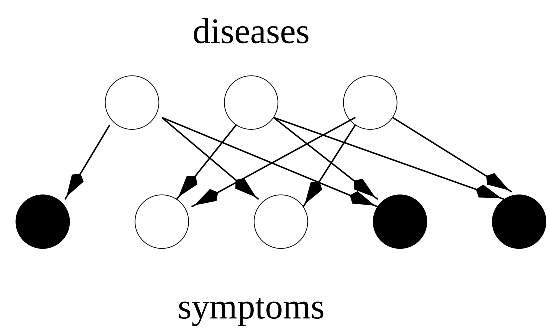
LfD2 Learning from Data 2. Focus on Reinforcement Learning, and advanced supervised learning methods

DME Develops ideas from LfD1, PMR to deal with real-world data sets. Also data visualization and new techniques.

Dealing with Uncertainty

- The key foci of this course are
  1. The use of probability theory as a calculus of uncertainty
  2. The *learning* of probability models from data
- Graphical descriptions are used to define (in)dependence
- Probabilistic graphical models give us a framework for dealing with hidden-cause (or latent variable) models
- Probability models can be used for classification problems, by building a probability density model for each class

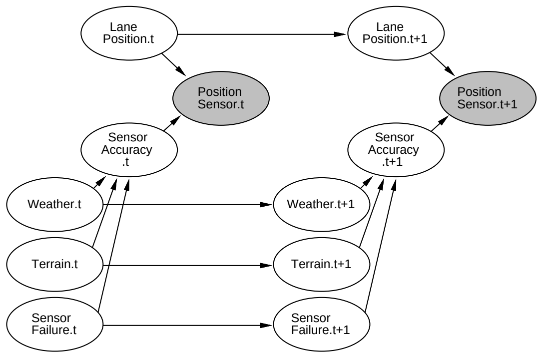
Example 1: QMR-DT



Shaded nodes represent observations

- Diagnostic aid in the domain of internal medecine
- 600 diseases, 4000 symptom nodes
- Task is to infer diseases given symptoms

Example 2: Inference for Automated Driving



- Model of a vision-based lane sensor for car driving
- Dynamic belief network—performing inference through time
- See Russell and Norvig, §17.5

Further Examples

- Automated Speech Recognition using Hidden Markov Models
  acoustic signal → phones → words
- Detecting genes in DNA (Krogh, Mian, Haussler, 1994)
- Tracking objects in images (Kalman filter and extensions)
- Troubleshooting printing problems under Windows 95 (Heckerman et al, 1995)
- Robot navigation: inferring where you are

Probability Theory

- Why probability?
- Events, Probability
- Variables
- Joint distribution
- Conditional Probability
- Bayes' Rule
- Inference
- Reference: e.g. Russell and Norvig, chapter 14

Why probability?

Even if the world were deterministic, probabilistic assertions *summarize* effects of

- **laziness**: failure to enumerate exceptions, qualifications etc.
- **ignorance**: lack of relevant facts, initial conditions etc.

Other approaches to dealing with uncertainty

- Default or non-monotonic logics
- Certainty factors (as in MYCIN) – *ad hoc*
- Dempster-Shafer theory
- Fuzzy logic handles degree of truth, not uncertainty

Events

- The set of all possible outcomes of an experiment is called the *sample space*, denoted by Ω
- Events are subsets of Ω
- If A and B are events, $A \cap B$ is the event "A and B"; $A \cup B$ is the event "A or B"; A^c is the event "not A"
- A probability measure is a way of assigning probabilities to events s.t
  - $P(\emptyset) = 0$, $P(\Omega) = 1$
  - If $A \cap B = \emptyset$
    $$P(A \cup B) = P(A) + P(B)$$
    i.e. probability is additive for disjoint events
- **Example**: when two fair dice are thrown, what is the probability that the sum is 4?

Variables

- A variable takes on values from a collection of mutually exclusive and collectively exhaustive states, where each state corresponds to some event
- A variable X is a map from the sample space to the set of states
- Examples of variables
  - Colour of a car $blue,\ green,\ red$
  - Number of children in a family $0, 1, 2, 3, 4, 5, 6, > 6$
  - Toss two coins, let $X =$ (number of heads)^2. X can take on the values 0, 1 and 4.
- Random variables can be *discrete* or *continuous*
- Use capital letters to denote random variables and lower case letters to denote values that they take, e.g. $P(X = x)$

Probability: Frequentist and Bayesian

- **Frequentist** probabilities are defined in the limit of an infinite number of trials
- Example: "The probability of a particular coin landing heads up is 0.43"
- **Bayesian** (subjective) probabilities quantify *degrees of belief*
- Example: "The probability of it raining tomorrow is 0.3"
- Not possible to repeat "tomorrow" many times
- Frequentist interpretation is a special case

Joint distributions

- Properties of several random variables are important for modelling complex problems
- Suppose $Toothache$ and $Cavity$ are the variables:

	$Toothache = true$	$Toothache = false$
$Cavity = true$	0.04	0.06
$Cavity = false$	0.01	0.89

- Notation
  $P(Toothache = true, Cavity = false) = 0.01$
- Marginal probabilities, the *sum rule*
  $$P(X) = \sum_Y P(X, Y)$$
  e.g. $P(Toothache = true) = ?$

Conditional Probability

- Let $\mathbf{X}$ and $\mathbf{Y}$ be two disjoint subsets of variables, such that $P(\mathbf{y}) > 0$. Then the *conditional probability distribution* (CPD) of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is given by

$$P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{y})}$$

- Product rule

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y}|\mathbf{X}) = P(\mathbf{Y})P(\mathbf{X}|\mathbf{Y})$$

- **Example**: In the dental example, what is $P(Cavity = true | Toothache = true)$?

- Chain rule is derived by repeated application of the product rule

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

Bayes' Rule

- From the product rule,

$$P(\mathbf{X}|\mathbf{Y}) = \frac{P(\mathbf{Y}|\mathbf{X})P(\mathbf{X})}{P(\mathbf{Y})}$$

- From the sum rule the denominator is

$$P(\mathbf{Y}) = \sum_{X} P(\mathbf{Y}|\mathbf{X})P(\mathbf{X})$$

- Why is this useful?
- For assessing *diagnostic* probability from causal probability

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

- **Example**: 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

Evidence: from Prior to Posterior

- Prior probability $P(Cavity = true) = 0.1$
- After we observe $Toothache = true$, we obtain the *posterior* probability $P(Cavity = true | Toothache = true)$
- This statement is dependent on the fact that $Toothache = true$ is all I know
- Revised probability of toothache if, say, I have a dental examination....
- Some information may be irrelevant, e.g. $P(Cavity = true | Toothache = true, DiceRoll = 5)$ $= P(Cavity = true | Toothache = true)$

Inference from joint distributions

- Typically, we are interested in the posterior joint distribution of the *query variables* $\mathbf{Y}$ given specific values $\mathbf{e}$ for the *evidence variables* $\mathbf{E}$

- Hidden variables $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

- Sum out over hidden variables

$$P(\mathbf{Y}|\mathbf{E}=\mathbf{e}) \propto P(\mathbf{Y},\mathbf{E}=\mathbf{e})$$
$$= \sum_{\mathbf{h}} P(\mathbf{Y},\mathbf{H}=\mathbf{h},\mathbf{E}=\mathbf{e})$$

- Obvious problems:
  1) Worst-case time complexity $O(d^n)$ where d is the largest arity
  2) Space complexity $O(d^n)$ to store the joint distribution
  3) How to find the numbers for $O(d^n)$ entries???

Decision Theory

$$\text{DecisionTheory} = \text{ProbabilityTheory} + \text{UtilityTheory}$$

- When making actions, an agent will have preferences about different possible outcomes

- Utility theory can be used to represent and reason with preferences

- A rational agent will select the action with the highest expected utility

Summary

- Course foci:
  - Probabilty theory as calculus of uncertainty
  - Learning probabilistic models form data

- Events, random variables

- Joint, conditional probability

- Bayes rule, evidence

- Decision theory