

Probabilistic Modelling and Reasoning

Chris Williams

School of Informatics, University of Edinburgh

September 2008

Course Introduction

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

Relationships between courses

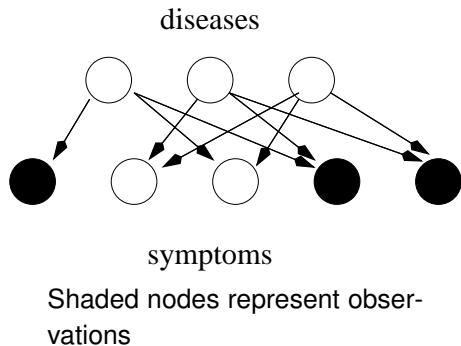
- PMR** Probabilistic modelling and reasoning. Focus on probabilistic modelling. Learning and inference for probabilistic models, e.g. Probabilistic expert systems, latent variable models, Hidden Markov models, Kalman filters, Boltzmann machines.
- IAML** Introductory Applied Machine Learning. Basic introductory course on supervised and unsupervised learning
- MLPR** More advanced course on machine learning, including coverage of Bayesian methods
 - RL** Reinforcement Learning. Focus on Reinforcement Learning (i.e. delayed reward).
- DME** Develops ideas from IAML, 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

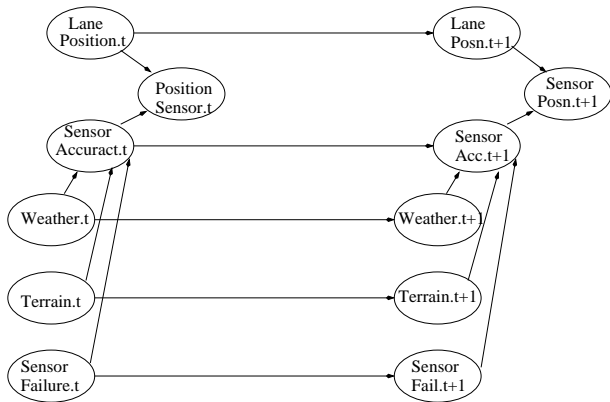
Example 1: QMR-DT

- Diagnostic aid in the domain of internal medicine
- 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. Bishop §1.2; 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 $X = (\text{number of heads})^2$. 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. $P(X = x)$
- $\sum_x P(X = x) = 1$

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:

	<i>Toothache = true</i>	<i>Toothache = false</i>
<i>Cavity = true</i>	0.04	0.06
<i>Cavity = false</i>	0.01	0.89

- Notation

$$P(\textit{Toothache} = \textit{true}, \textit{Cavity} = \textit{false}) = 0.01$$

Marginal Probabilities

The *sum rule*

$$P(X) = \sum_Y P(X, Y)$$

e.g. $P(\textit{Toothache} = \textit{true}) = ?$

Conditional Probability

- Let \mathbf{X} and \mathbf{Y} be two disjoint subsets of variables, such that $P(\mathbf{Y} = \mathbf{y}) > 0$. 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 $P(\text{Cavity} = \text{true} | \text{Toothache} = \text{true})$?
- $\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y}) = 1$ for all \mathbf{y}
- Can we say anything about $\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x} | \mathbf{Y} = \mathbf{y})$?

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

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

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

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

Why is this useful?

- For assessing *diagnostic* probability from causal probability

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

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

Inference from joint distributions

- Typically, we are interested in the posterior joint distribution of the *query variables* \mathbf{X}_F given specific values \mathbf{e} for the *evidence variables* \mathbf{X}_E
- Remaining variables $\mathbf{X}_R = \mathbf{X} \setminus (\mathbf{X}_F \cup \mathbf{X}_E)$
- Sum out over \mathbf{X}_R

$$\begin{aligned} P(\mathbf{X}_F | \mathbf{X}_E = \mathbf{e}) &= \frac{P(\mathbf{X}_F, \mathbf{X}_E = \mathbf{e})}{P(\mathbf{X}_E = \mathbf{e})} \\ &= \frac{\sum_{\mathbf{r}} P(\mathbf{X}_F, \mathbf{X}_R = \mathbf{r}, \mathbf{X}_E = \mathbf{e})}{\sum_{\mathbf{r}, \mathbf{f}} P(\mathbf{X}_F = \mathbf{f}, \mathbf{X}_R = \mathbf{r}, \mathbf{X}_E = \mathbf{e})} \end{aligned}$$

Problems with naïve inference:

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

Decision Theory

DecisionTheory = ProbabilityTheory + 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:
 - Probability theory as calculus of uncertainty
 - Inference in probabilistic graphical models
 - Learning probabilistic models from data
- Events, random variables
- Joint, conditional probability
- Bayes rule, evidence
- Decision theory