Informatics 2D – Reasoning and Agents

Semester 2, 2019–2020

Alex Lascarides
alex@inf.ed.ac.uk

Lecture 22 – Probabilities and Bayes' Rule

10th March 2020

Where are we?

Last time ...

- Introduced basics of decision theory (probability theory + utility)
- Talked about random variables, probability distributions
- Introduced basic probability notation and axioms

Today ...

- **Probabilities and Bayes' Rule**

Inference with joint probability distributions

- Last time we talked about joint probability distributions (JPDs) but didn't present a method for **probabilistic inference** using them
- Problem: Given some observed evidence and a query proposition, how can we compute the **posterior probability** of that proposition?
- We will first discuss a simple method using a JPD as "knowledge base"
- Although not very useful in practice, it helps us to discuss interesting issues along the way

Example

- Domain consisting only of Boolean variables *Toothache*, *Cavity* and *Catch* (steel probe catches in tooth)
- Consider the following JPD:

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- Probabilities (table entries) sum to 1
- We can compute probability of any proposition, e.g.
  $P(catch \vee cavity) =$
  $0.108 + 0.016 + 0.072 + 0.144 + 0.012 + 0.008 = 0.36$

Marginalisation, conditioning & normalisation

- Extracting distribution of subset of variables is called **marginalisation**: $\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})$
- Example:

$$\begin{aligned} P(cavity) = {} & P(cavity, toothache, catch) + P(cavity, toothache, \neg catch) \\ & + P(cavity, \neg toothache, catch) + P(cavity, \neg toothache, \neg catch) \\ & = 0.108 + 0.012 + 0.072 + 0.008 = 0.2 \end{aligned}$$

- **Conditioning** – variant using the product rule:

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

Marginalisation, conditioning & normalisation

- Computing conditional probabilities:

$$
\begin{aligned}
P(cavity|toothache) &= \frac{P(cavity \wedge toothache)}{P(toothache)} \\
&= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6
\end{aligned}
$$

- **Normalisation** ensures probabilities sum to 1, normalisation constants often denoted by α
- Example:

$$
\begin{aligned}
&\mathbf{P}(Cavity|toothache) = \alpha\mathbf{P}(Cavity, toothache) \\
&= \alpha[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}
$$

A general inference procedure

- Let X be a query variable (e.g. *Cavity*), $\mathbf{E}$ set of evidence variables (e.g. $\{Toothache\}$) and $\mathbf{e}$ their observed values, $\mathbf{Y}$ remaining unobserved variables
- Query evaluation: $\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$
- Note that X, $\mathbf{E}$, and $\mathbf{Y}$ constitute complete set of variables, i.e. $\mathbf{P}(x, \mathbf{e}, \mathbf{y})$ simply a subset of probabilities from the JPD
- For every value x_i of X, sum over all values of every variable in $\mathbf{Y}$ and normalise the resulting probability vector
- Only theoretically relevant, it requires $O(2^n)$ steps (and entries) for n Boolean variables
- Basically, all methods we will talk about deal with tackling this problem!

Independence

- Suppose we extend our example with the variable *Weather*
- What is the relationship between old and new JPD?
- Can compute $P(toothache, catch, cavity, Weather = cloudy)$ as:

$$P(Weather = cloudy|toothache, catch, cavity)P(toothache, catch, cavity)$$

- And since the weather does not depend on dental stuff, we expect that

$$P(Weather = cloudy|toothache, catch, cavity) = P(Weather = cloudy)$$

- So

$$P(toothache, catch, cavity, Weather = cloudy) = P(Weather = cloudy)P(toothache, catch, cavity)$$

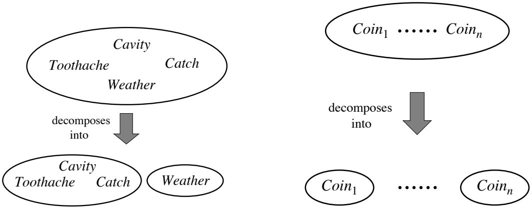
- One 8-element and one 4-element table rather than a 32-table!

Independence

- This is called **independence**, usually written as

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

- Depends on domain knowledge; can factor distributions



- Such independence assumptions can help to dramatically reduce complexity
- Independence assumptions are sometimes *necessary* even when not entirely justified, so as to make probabilistic reasoning in the domain practical (more later).

Bayes' rule

- **Bayes' rule** is derived by writing the product rule in two forms and equating them:

$$\left.\begin{array}{l} P(a \wedge b) = P(a|b)P(b) \\ P(a \wedge b) = P(b|a)P(a) \end{array}\right\} \Rightarrow P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

- General case for multivaried variables using background evidence **e**:

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

- Useful because often we have good estimates for three terms on the right and are interested in the fourth

Applying Bayes' rule

- Example: meningitis causes stiff neck with 50%, probability of meningitis (m) 1/50000, probability of stiff neck (s) 1/20

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{\frac{1}{2} \times \frac{1}{50000}}{\frac{1}{20}} = \frac{1}{5000}$$

- Previously, we were able to avoid calculating probability of evidence ($P(s)$) by using normalisation
- With Bayes' rule: $\mathbf{P}(M|s) = \alpha\langle P(s|m)P(m), P(s|\neg m)P(\neg m)\rangle$
- Usefulness of this depends on whether $P(s|\neg m)$ is easier to calculate than $P(s)$
- Obvious question: why would conditional probability be available in one direction and not in the other?
- Diagnostic knowledge (from symptoms to causes) is often fragile (e.g. $P(m|s)$ will go up if $P(m)$ goes up due to epidemic)

Combining evidence

- Attempting to use additional evidence is easy in the JPD model

$$\mathbf{P}(Cavity|toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \approx \langle 0.871, 0.129\rangle$$

but requires additional knowledge in Bayesian model:

$$\mathbf{P}(Cavity|toothache \wedge catch) = \alpha\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$$

- This is basically almost as hard as JPD calculation
- Refining idea of independence: *Toothache* and *Catch* are independent given presence/absence of *Cavity* (both caused by cavity, no effect on each other)

$$\mathbf{P}(toothache \wedge catch|Cavity) = \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)$$

Conditional independence

- Two variables X and Y are conditionally independent given Z if $\mathbf{P}(X, Y|Z) = \mathbf{P}(X|Z)\mathbf{P}(Y|Z)$
- Equivalent forms $\mathbf{P}(X|Y, Z) = \mathbf{P}(X|Z)$, $\mathbf{P}(Y|X, Z) = \mathbf{P}(Y|Z)$
- So in our example:

$$\mathbf{P}(Cavity|toothache \wedge catch) = \alpha \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$$

- As before, this allows us to decompose large JPD tables into smaller ones, grows as $O(n)$ instead of $O(2^n)$
- This is what makes probabilistic reasoning methods scalable at all!

Conditional independence

- Conditional independence assumptions much more often reasonable than absolute independence assumptions
- **Naive Bayes model**:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i | Cause)$$

- Based on the idea that all effects are conditionally independent given the cause variable
- Also called **Bayesian classifier** or (by some) even "**idiot Bayes** model"
- Works surprisingly well in many domains despite its simplicity!

Summary

- Probabilistic inference with full JPDs
- Independence and conditional independence
- Bayes' rule and its applications problems with fairly simple techniques
- Next time: **Probabilistic Reasoning with Bayesian Networks**