KR-IST Lecture 8b
Bayesian Reasoning

Chris Thornton

November 22, 2013

From rules to conditional probabilities

The standard symbolic rule is treated as fully certain.

$$\text{speedingTicket} \Rightarrow \text{speeding}$$

This is read 'getting a speedingTicket implies that you were speeding'.

But we often need to be able to state consequences probabilistically. This can be done using conditional probabilities.

$$P(\text{speeding}|\text{speedingTicket}) = 0.9$$

This is read 'the probability of speeding given that you got a speeding ticket is 0.9'

Conditional probability distributions

Working in terms of conditional probabililties, we have a distribution of probability values over possible states of affairs.

$$P(\text{speedingTicket}|\text{speeding}) = 0.8$$
$$P(\text{speedingTicket}|\text{sleeping}) = 0.1$$
$$P(\text{speedingTicket}|\text{swimming}) = 0.1$$

Probabilities in a distribution must sum to 1.0.

Uncertainty

The level of uncertainty regarding the state of affairs can be worked out by looking at the distribution.

The more flat it is, the greater the uncertainty.

The more cases there are, the greater the uncertainty (for distributions of a particular flatness).

Entropy

The entropy formula takes both aspects into account.

$$-\sum_i P_i \log_2 P_i$$

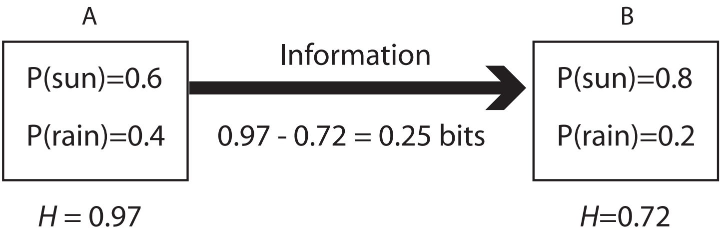
where P_i is the probability of the ith alternative.

The value of the entropy rises with the number of alternatives *and* the uniformity of the attributed probabilities.

More extreme probabilities produce lower evaluations.

$$\begin{aligned}
-(1.0 \log_2 1.0) &= 0 \\
-(0.0 \log_2 0.0) &= 0 \\
-(0.5 \log_2 0.5) &= 0.5 \\
-(0.7 \log_2 0.7) &= 0.36
\end{aligned}$$

Information and knowledge



Using entropy as a measure of uncertainty, we can evaluate how much information is obtained when something happens (e.g., a message) which updates distributions.

Reduction of uncertainty implies an increase of knowledge.

Information in bits

Let's say there are 4 possible states of affairs: speeding, sleeping, swimming, eating.

We have no knowledge about which is the case.

The probability distribution is $\{0.25, 0.25, 0.25, 0.25\}$.

The entropy is 2.0

(It's always $\log_2 n$ with a flat distribution.)

Given we took logs to base 2, the entropy is also the number of bits you need in a binary system to encode 4 values.

The amount of information in a message or event which establishes the state of affairs is then measured as 2 bits.

Probabilistic reasoning

As well as being key for information theory, conditional probabilities are also the basis for methods of probabilistic reasoning.

These methods chain implications together in a way that takes probability into account.

The simplest approach to probabilistic reasoning uses the inference method known as **Bayes' rule**.

Bayes rule

Given evidence E and some conclusion C, it's always the case that

$$P(C|E) \quad = \quad \frac{P(C)P(E|C)}{P(E)}$$

We can plug any values we like into this formula to infer a conditional probability for the conclusion.

P(C) and P(E) are called **prior probabilities**. P(E|C) is the **likelihood**. P(C|E) is called the **posterior probability**.

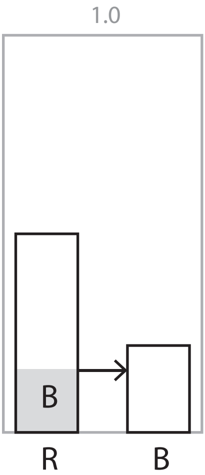
Rich bankers example

50% of people are rich and 20% are bankers.
30% of rich people are bankers.
What are the chances of a random banker being rich?

$P(R) = 0.5$

$P(B) = 0.2$

$P(B|R) = 0.3$



$$P(R|B) = \frac{P(R)\,P(B|R)}{P(B)}$$

$$= \frac{0.5 \times 0.3}{0.2}$$

$$= \frac{0.15}{0.2}$$

$$= 0.75$$

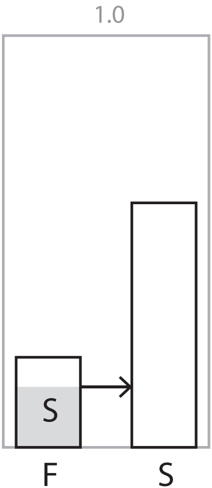
Flu diagnosis example

20% of people have flu and 60% are sneezing.
70% of people with flu are found to be sneezing.
What is the probability someone sneezing has flu?

$P(F) = 0.2$

$P(S) = 0.6$

$P(S|F) = 0.7$



$$P(F|S) = \frac{P(F)\,P(S|F)}{P(S)}$$

$$= \frac{0.2 \times 0.7}{0.6}$$

$$= \frac{0.14}{0.6}$$

$$= 0.23$$

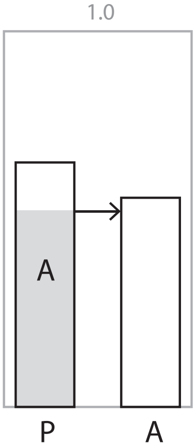
Exam prediction example

60% of people pass the AI exam but only 50% attend lectures.
80% of people who pass the exam attend lectures.
What is the probability of passing the exam given you attend lectures

$P(P) = 0.6$

$P(A) = 0.5$

$P(A|P) = 0.8$



$$P(P|A) = \frac{P(P)\,P(A|P)}{P(A)}$$

$$= \frac{0.6 \text{ x } 0.8}{0.5}$$

$$= \frac{0.48}{0.5}$$

$$= 0.96$$

Bayesian (MAP) inference

Say we have observations D1, D2, and explanatory hypotheses H1 and H2, with all priors (e.g., P(D2)) and likelihoods (e.g., P(D2|H1)) known.

By combining Bayes rule with the product rule, can find the probability of each hypothesis given the data.

$$P(H1|D1,D2) = P(H1|D1) \times P(H1|D2)$$

The most probable hypothesis is called the *maximum a posteriori* (MAP) hypothesis.

Deriving it is called MAP inference (what is usually meant by 'Bayesian inference')

In practice, the process has the problem that probabilities become vanishingly small.

Summary

Summary

- Conditional probabilities are like fuzzy rules

Summary

- Conditional probabilities are like fuzzy rules
- Uncertainty a function of distributional flatness

Summary

- Conditional probabilities are like fuzzy rules
- Uncertainty a function of distributional flatness
- Uncertainty can be measured as entropy

Summary

- Conditional probabilities are like fuzzy rules
- Uncertainty a function of distributional flatness
- Uncertainty can be measured as entropy
- Degree of knowledge corresponds to lack of uncertainty.

Summary

- Conditional probabilities are like fuzzy rules
- Uncertainty a function of distributional flatness
- Uncertainty can be measured as entropy
- Degree of knowledge corresponds to lack of uncertainty.
- Information measured in bits.

Summary

- Conditional probabilities are like fuzzy rules
- Uncertainty a function of distributional flatness
- Uncertainty can be measured as entropy
- Degree of knowledge corresponds to lack of uncertainty.
- Information measured in bits.
- Basic probabilistic reaswoning using Bayes' rule

Summary

- Conditional probabilities are like fuzzy rules
- Uncertainty a function of distributional flatness
- Uncertainty can be measured as entropy
- Degree of knowledge corresponds to lack of uncertainty.
- Information measured in bits.
- Basic probabilistic reaswoning using Bayes' rule

The classic texts

The classic texts

- Claude Shannon's 'A Mathematical Theory of Communication' (1948), which noted that entropy forms a perfect measure of uncertainty and set the foundations of modern information theory.

The classic texts

- Claude Shannon's 'A Mathematical Theory of Communication' (1948), which noted that entropy forms a perfect measure of uncertainty and set the foundations of modern information theory.
- Pearl, J. (1988). 'Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.' San Mateo: Morgan and Kaufman.

The classic texts

- Claude Shannon's 'A Mathematical Theory of Communication' (1948), which noted that entropy forms a perfect measure of uncertainty and set the foundations of modern information theory.
- Pearl, J. (1988). 'Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.' San Mateo: Morgan and Kaufman.