Probability

In logic, every predicate is either true or false, and that's that.

- In the real world, that is still true:
  - Every alleged fact is either true or false
  - Every possible percept either appears or it doesn't
  - Every possible event either happens or doesn't
- But very often we just don't know which it is, true or false
- Standard logic can only handle certainty, true or false and nothing else
- Probability theory is one way of dealing with that
- The probability of a predicate, percept, or event
  - is a number on a continuous spectrum from 0 to 1
    - sometimes expressed as a percentage
  - P(event) = 0 means it is absolutely impossible for it to happen
  - P(event) = 1 means it is absolutely certain to happen
- But what do numbers *between* 0 and 1 mean?
  - e.g. weather forecast says 40% chance of rain
  - e.g. P(unfair-coin-lands-heads-up) = 0.4
  - Average result of an "infinite" number of identical experiments
  - Proportion of possible worlds in which it is true
  - 40¢ is a fair price to pay for a bet that would pay out $1
- Stochastic ≈ probability-based

Basic laws (Venn diagrams)

- $P(A \wedge B) = P(A) \times P(B)$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- $P(\neg A) = 1 - P(A)$
- Independent variables

Sometimes probabilities are Conditional

- The chances of an event happening depend on whether some other event has happened
- | means "given" or roughly "if we know that"
- For example, tooth aches do not happen much (if you clean your teeth)
  - P(toothache) = 0.005
  - but
    - P(toothache | cavity) = 0.3
  - This does not denote cause and effect, it could also be that
    - P(cavity | toothache) = 0.75
  - Useful for diagnostic purposes
  - It is still true that P(toothache) = 0.005
    - even after cavity has been observed
- $P(A \mid B) = P(A \wedge B) / P(B)$
  - Observing B rules out all cases where B is false
  - leaving the set of possibilities with total probability of just P(B)
  - within those remaining possibilities,
    - for P(A | B) to be true, A must be true
  - so A and B is true

so the true probability of A | B must be P(A ∧ B)
divided by P(B), to make all the probabilities add up to 1

- P(A ∧ B) = P(A | B) × P(B) is sometimes a more convenient form
  For A and B to be true, we need B to be true, then given that B is true
  we also need A to be true

Notation

- If the range of possible values for Weather is [sunny, rainy, cloudy, snowy]
  we might have
  P(Weather = sunny) = 0.6
  P(Weather = rainy) = 0.1
  P(Weather = cloudy) = 0.29
  P(Weather = snowy) = 0.01
  (these are not independent, so the ∨ rule doesn't work)
  this is often written as
  P(Weather) = < 0.6, 0.1, 0.29, 0.01 >
- Sometimes P(sunny) is written as an abbreviation for P(Weather = sunny)
- For continuous variables, a Probability Density Function is used
- Probability = the area under the curve
- Numeric probabilities only really make sense for ranges of possible values

Joint distributions

- If we have three Boolean variables, Toothache, Cavity, and Catch
  the joint distribution is a 2 × 2 × 2 table, all eight add up to exactly 1
- Use a bold **P** to represent that
- P(some possibility) can be found by adding up the relevant entries
- Marginalisation - eliminate variable(s) by summing entries
  P(some possibility) = Σ for all possible x's of **P**(that possibility ∧ X=x)
- Conditioning
  P(some possibility) = Σ for all possible x's of
  P(that possibility | X=x) × P(X=x)

Bayes' rule

- From P(A ∧ B) = P(A | B) × P(B) and P(A ∧ B) = P(B | A) × P(A)
  we easily get Bayes' rule:
  P(B | A) = P(A | B) × P(B) / P(A)
- When diagnosing an illness, a doctor might know all of
  P(effect | cause)
  P(effect)
  P(cause)
  for a vast collection of different causes and effects,
  just from hundreds of years of observations and studies
  And from them P(cause | effect) can be calculated,
  the probability of a particular disease being the cause of the
  observed effects
- In an epidemic, P(cause) for one particular cause will increase a lot
  previously observed values for P(cause | effect) will become invalid

Combining evidence

- If we observe both toothache and catch, what is the probability distribution for Cavity?
- Assume this joint distribution:

	Toothache		¬Toothache	
	Catch	¬Catch	Catch	¬Catch
Cavity	.108	.012	.072	.008
¬Cavity	.016	.064	.144	.576

(they all add up to 1)

- If we have a joint distribution we can just add up the right entries

  $\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle$ perhaps

  α stands for Normalise:

  multiply by something to make the probabilities add up to 1

  so $\alpha\langle 0.108, 0.016\rangle = \langle 0.871, 0.129\rangle$
- Does not scale up if there are a lot of variables
- Using Bayes' rule

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

  $\alpha\, \mathbf{P}(toothache \wedge catch \mid Cavity) \times \mathbf{P}(Cavity)$
- Again, there are probably too many variables for this to be practical
- If the variables were independent, we'd be better off

  but they aren't. Toothache and Catch are not unrelated
- But if, given knowledge of Cavity, they become independent
- From $P(toothache \wedge catch \mid Cavity) =$

  $P(toothache \mid Cavity) \times P(catch \mid Cavity)$

  we get $P(toothache \wedge catch \mid Cavity) =$

  $\alpha\, P(toothache \mid Cavity) \times P(catch \mid Cavity) \times P(Cavity)$
- Not so many combinations to worry about now
- Conditional independence allows scalability

Naive Bayes Models

- If all the effects are conditionally independent given Cause, then

  $P(Cause \wedge Effect_1 \wedge Effect_2 \wedge \ldots) = P(Cause) \times \Pi\, P(Effect_i \mid Cause)$
- Naive because it relies on independence, but is sometimes used when there is none
- $P(Cause \mid Effect{=}e) = \alpha\, P(Cause) \times \Pi\, P(Effect{=}e_i \mid Cause)$
- Example: text classification

  Given some text, work out which section of the newspaper it came from

  We can know from prior observations these distributions

  $\mathbf{P}(Section)$

  $\mathbf{P}(Hasword_w \mid Section)$

  If 9% of all articles are in the weather section then

  $P(Section{=}weather) = 0.09$

  If 23% of all articles in the weather section use the word "rain" then

  $P(Hasword_{rain}{=}true \mid Section{=}weather) = 0.23$

  From that and the current text itself, we can work out the probabilities of the text being from any given section