12.

Vagueness, Uncertainty and Degrees of Belief

Noncategorical statements

Ordinary commonsense knowledge quickly moves away from categorical statements like "a P is *always* (*unequivocably*) a Q"

There are many ways in which we can come to less than categorical information

- things are *usually* (*almost never*, *occasionally*, *seldomly*, *rarely, almost always*) a certain way
- judgments about how good an example something is

  e.g., barely rich, a poor example of a chair, not very tall
- imprecision of sensors

  e.g., the best you can do is to get within +/-10%
- reliability of sources of information

  e.g., "most of the time he's right on the money"
- strength/confidence/trust in generic information or deductive rules

Conclusions will not "follow" in the usual sense

Weakening a universal

There are at least 3 ways a universal like $\forall x\, P(x)$ can be made ro be less categorical:

$\forall x\, P(x)$

strength of quantifier
("95% of birds fly")

statistical interpretation/
probabilistic sentences

applicability of predicate/
degree of membership
("fairly tall")

flexible membership
vague predicates

degree of belief in whole
sentence ("80%
confidence in this fact")

uncertain knowledge/
subjective probability

Objective probability

Statistical (frequency) view of sentences

objective: does not depend on who is assessing the probability

Always applied to collections

can not assign probabilities to (random) events that are not members of any obvious repeatable sequence:

- ok for "the probability that I will pick a red face card from the deck"
- not ok for "the probability that the Blue Jays will win the World Series this Fall"
- "the probability that Tweety flies is between .9 and .95" is always false (either Tweety flies or not)

Can use probabilities to correspond to English words like "rarely," "likely," "usually"

generalized quantifiers: "most," "many," "few"

For most x, $Q(x)$ vs. For all x, $Q(x)$

The basic postulates

Numbers between 0 and 1 representing frequency of an event in a (large enough) random sample

extremes: 0 = never happens; 1 = always happens

Start with set U of all possible occurrences. An event a is any subset of U.
A probability measure is any function $\boldsymbol{Pr}$ from events to [0,1] satisfying:

- $\boldsymbol{Pr}(U) = 1$.
- If $a_1, ..., a_n$ are disjoint events, then $\boldsymbol{Pr}(\cup a_i) = \Sigma\, \boldsymbol{Pr}(a_i)$

Conditioning: the probability of one event may depend on its interaction with others

$\boldsymbol{Pr}(a|b)$ = probability of a, given b = $\boldsymbol{Pr}(a \cap b) / \boldsymbol{Pr}(b)$

Conditional independence:

event a is judged independent of event b conditional on background knowledge s if knowing that b happened does not affect the probability of a

$\boldsymbol{Pr}(a|s) = \boldsymbol{Pr}(a|b,s)$ (note: CI is symmetric)

Note: without independence, $\boldsymbol{Pr}(a|s)$ and $\boldsymbol{Pr}(a|b,s)$ can be very different.

Some useful consequences

Conjunction:

$\boldsymbol{Pr}(ab) = \boldsymbol{Pr}(a|b) \cdot \boldsymbol{Pr}(b)$

conditionally independent: $\boldsymbol{Pr}(ab) = \boldsymbol{Pr}(a) \cdot \boldsymbol{Pr}(b)$

Negation:

$\boldsymbol{Pr}(\neg s) = 1 - \boldsymbol{Pr}(s)$

$\boldsymbol{Pr}(\neg s|d) = 1 - \boldsymbol{Pr}(s|d)$

If $b_1, b_2, ..., b_n$ are pairwise disjoint and exhaust all possibilities, then

$\boldsymbol{Pr}(a) = \sum \boldsymbol{Pr}(ab_i) = \sum \boldsymbol{Pr}(a \mid b_i) \cdot \boldsymbol{Pr}(b_i)$

$\boldsymbol{Pr}(a \mid c) = \sum \boldsymbol{Pr}(ab_i \mid c)$

Bayes' rule:

$\boldsymbol{Pr}(a|b) = \boldsymbol{Pr}(a) \cdot \boldsymbol{Pr}(b|a) / \boldsymbol{Pr}(b)$

if a is a disease and b is a symptom, it is usually easier to estimate numbers on RHS of equation (see below, for subjective probabilities)

Subjective probability

It is reasonable to have non-categorical beliefs even in categorical sentences

- confidence/certainty in a sentence
- "your" probability = *subjective*

Similar to defaults

- move from statistical/group observations to belief about individuals
- but not categorical: how certain am I that Tweety flies?

"Prior probability" $\boldsymbol{Pr}(x|s)$ (s = prior state of information or background knowledge)

"Posterior probability" $\boldsymbol{Pr}(x|E,s)$ (E = new evidence)

Need to *combine evidence* from various sources

how to derive new beliefs from prior beliefs and new evidence?

want explanations; probability is just a summary

From statistics to belief

Would like to go from statistical information (e.g., the probability that a bird chosen at random will fly) to a degree of belief (e.g., how certain are we that this particular bird, Tweety, flies)

Traditional approach is to find a reference class for which we have statistical information and use the statistics for that class to compute an appropriate degree of belief for an individual

Imagine trying to assign a degree of belief to the proposition "Eric (an American male) is tall" given facts like these

A) 20% of American males are tall
B) 25% of Californian males are tall
C) Eric is from California

This is called direct inference

Problem: individuals belong to many classes

- with just A $\rightarrow$ .2
- A,B,C - prefer more specific $\rightarrow$ .25
- A,C - no statistics for more specific class $\rightarrow$ .2?
- B - are Californians a representative sample?

Basic Bayesian approach

Would like a more principled way of calculating subjective probabilities

Assume we have n atomic propositions $p_1,..., p_n$ we care about. A logical interpretation I can be thought of as a specification of which p_i are true and which are false.

Notation: for $n=4$, we use $\langle\neg p_1,p_2,p_3,\neg p_4\rangle$ to mean the interpretation where only p_2 and p_3 are true.

A joint probability distribution J, is a function from interpretations to $[0,1]$ satisfying $\Sigma\, J(I) = 1$ (where $J(I)$ is the degree of belief in the world being as per I).

The degree of belief in any sentence α: $\boldsymbol{Pr}(\alpha) = \sum_{I \models \alpha} J(I)$

Example: $\boldsymbol{Pr}(p_2 \wedge \neg p_4) = J(\langle\neg p_1, p_2, p_3,\neg p_4\rangle) + J(\langle\neg p_1, p_2,\neg p_3,\neg p_4\rangle) + J(\langle p_1, p_2, p_3,\neg p_4\rangle) + J(\langle p_1, p_2,\neg p_3,\neg p_4\rangle)$.

Problem with the approach

To calculate the probabilities of arbitrary sentences involving the p_i, we would need to know the full joint distribution function.

For n atomic sentences, this requires knowing 2^n numbers

impractical for all but very small problems

Would like to make plausible assumptions to cut down on what needs to be known.

In the simplest case, all the atomic sentences are independent. This gives us that

$J(\langle P_1,...,P_n\rangle) = \boldsymbol{Pr}(P_1 \wedge ... \wedge P_n) = \prod \boldsymbol{Pr}(P_i)$ (where P_i is either p_i or $\neg p_i$)

and so only n numbers are needed.

Bu this assumption is too strong. A better assumption:

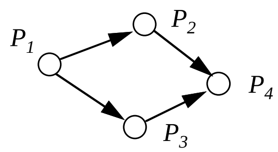
the probability of each P_i only depends on a small number of P_j, and the dependence is acyclic.

Belief networks

Represent all the atoms in a <u>belief network</u> (or Bayes' network).

Assume: $\boxed{J(\langle P_1,...,P_n\rangle) \;=\; \prod \boldsymbol{Pr}(P_i \mid c(P_i)) \quad \text{where } \boldsymbol{Pr}(c(P_i)) > 0}$ $c(P)$ = parents of node P

Example:



$$J(\langle P_1,P_2,P_3,P_4\rangle) \;=\; \mathbf{Pr}(P_1)\cdot\mathbf{Pr}(P_2 \mid P_1)\cdot \mathbf{Pr}(P_3 \mid P_1)\cdot\mathbf{Pr}(P_4 \mid P_2,P_3).$$

So: $J(p_1,\overline{p_2},p_3,\overline{p_4}) \;= \mathbf{Pr}(p_1)\cdot\mathbf{Pr}(\overline{p_2} \mid p_1)\cdot\mathbf{Pr}(p_3 \mid p_1)\cdot\mathbf{Pr}(\overline{p_4} \mid \overline{p_2},p_3)$
$= \;\mathbf{Pr}(p_1)\cdot[1 - \mathbf{Pr}(p_2 \mid p_1)]\cdot\mathbf{Pr}(p_3 \mid p_1)\cdot[\,1 - \mathbf{Pr}(p_4 \mid \overline{p_2},p_3)]$

To fully specify the joint distribution (and therefore probabilities over any subset of the variables), we only need $\boldsymbol{Pr}(P \mid c(P))$ for every node P.

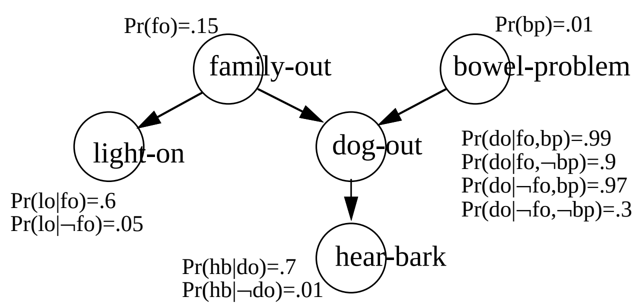
If node P has parents $Q_1, ..., Q_m$, then we need to know the values of $\mathbf{Pr}(p \mid q_1,q_2,...\, q_m)$, $\mathbf{Pr}(p \mid \overline{q_1},q_2 \,...\, q_m)$, $\mathbf{Pr}(p \mid q_1,\overline{q_2},...\, q_m)$, ..., $\mathbf{Pr}(p \mid \overline{q_1},\overline{q_2},...\, \overline{q_m})$.

$n \cdot 2^m$ numbers << 2^n numbers !

Using belief networks

Assign a node to each variable in the domain and draw arrows toward each node P from a select set $c(P)$ of nodes perceived to be "direct causes" of P.

arcs can often be interpreted as causal connections



From the DAG, we get that

$J(\langle \text{FO, LO, BP, DO, HB}\rangle) \;=$
$\boldsymbol{Pr}(\text{FO}) \times \boldsymbol{Pr}(\text{LO} \mid \text{FO}) \times \boldsymbol{Pr}(\text{BP}) \times \boldsymbol{Pr}(\text{DO} \mid \text{FO, BP}) \times \boldsymbol{Pr}(\text{HB} \mid \text{DO})$

Using this formula and the 10 numbers above, we can calculate the full joint distribution

Example calculation

Suppose we want to calculate $\boldsymbol{Pr}(\mathrm{fo} \mid \mathrm{lo}, \neg\mathrm{hb})$

$\boldsymbol{Pr}(\mathrm{fo} \mid \mathrm{lo}, \neg\mathrm{hb}) = \boldsymbol{Pr}(\mathrm{fo}, \mathrm{lo}, \neg\mathrm{hb}) \ / \ \boldsymbol{Pr}(\mathrm{lo}, \neg\mathrm{hb})$ where

$\boldsymbol{Pr}(\mathrm{fo}, \mathrm{lo}, \neg\mathrm{hb}) = \sum J(\langle \mathrm{fo}, \mathrm{lo}, \mathrm{BP}, \mathrm{DO}, \neg\mathrm{hb}\rangle)$ first 4 values below

$\boldsymbol{Pr}(\mathrm{lo}, \neg\mathrm{hb}) = \sum J(\langle \mathrm{FO}, \mathrm{lo}, \mathrm{BP}, \mathrm{DO}, \neg\mathrm{hb}\rangle)$ all 8 values below

$J(\langle \mathrm{fo,lo,bp,do,\neg hb}\rangle) = .15 \cdot .6 \cdot .01 \cdot .99 \cdot .3 = .0002673 +$
$J(\langle \mathrm{fo,lo,bp,\neg do,\neg hb}\rangle) = .15 \cdot .6 \cdot .01 \cdot .01 \cdot .99 = .00000891 +$
$J(\langle \mathrm{fo,lo,\neg bp,do,\neg hb}\rangle) = .15 \cdot .6 \cdot .99 \cdot .9 \cdot .3 = .024057 +$
$J(\langle \mathrm{fo,lo,\neg bp,\neg do,\neg hb}\rangle) = .15 \cdot .6 \cdot .99 \cdot .1 \cdot .99 = .0088209 +$
$J(\langle \mathrm{\neg fo,lo,bp,do,\neg hb}\rangle) = .85 \cdot .05 \cdot .01 \cdot .97 \cdot .3 = .000123675$
$J(\langle \mathrm{\neg fo,lo,bp,\neg do,\neg hb}\rangle) = .85 \cdot .05 \cdot .01 \cdot .03 \cdot .99 = .0000126225 +$
$J(\langle \mathrm{\neg fo,lo,\neg bp,do,\neg hb}\rangle) = .85 \cdot .05 \cdot .99 \cdot .3 \cdot .3 = .00378675$
$J(\langle \mathrm{\neg fo,lo,\neg bp,\neg do,\neg hb}\rangle) = .85 \cdot .05 \cdot .99 \cdot .7 \cdot .99 = .029157975$

$\boldsymbol{Pr}(\mathrm{fo} \mid \mathrm{lo}, \neg\mathrm{hb}) = .03316 \ / \ .06624 = .5$

Bypassing the full calculation

Often it is possible to calculate some probability values without first calculating the full joint distribution

Example: what is $\mathbf{Pr}(\mathrm{fo} \mid \mathrm{lo})$?

by Bayes rule: $\mathbf{Pr}(\mathrm{fo} \mid \mathrm{lo}) = \overset{\checkmark}{\mathbf{Pr}(\mathrm{lo} \mid \mathrm{fo})} \cdot \overset{\checkmark}{\mathbf{Pr}(\mathrm{fo})} \ / \ \overset{?}{\mathbf{Pr}(\mathrm{lo})}$

but: $\mathbf{Pr}(\mathrm{lo}) = \overset{\checkmark}{\mathbf{Pr}(\mathrm{lo} \mid \mathrm{fo})} \cdot \overset{\checkmark}{\mathbf{Pr}(\mathrm{fo})} + \overset{\checkmark}{\mathbf{Pr}(\mathrm{lo} \mid \overline{\mathrm{fo}})} \cdot \overset{\checkmark}{\mathbf{Pr}(\overline{\mathrm{fo}})}$

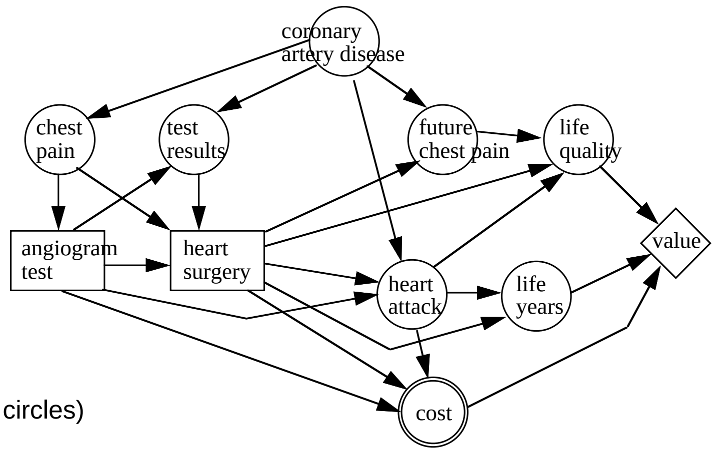
But in general, the problem is NP-hard

- the problem is even hard to approximate in general
- much of the attention on belief networks involves special-purpose procedures that work well for restricted topologies

Influence diagrams

Graphical knowledge representation for decision problems

- nodes represent propositions or quantities of interest, including decision variables, states of the world, and preference values
- arcs represent influence or relevance (probabilistic or deterministic relationships between the variables)



Node types

chance nodes (circles)
value nodes (diamonds)
decision nodes (rectangles)
deterministic nodes (double circles)

Dempster-Shafer theory

Another attempt at evidence-pooling

for cases where there is uncertainty about probability

Uses two-part measure: *belief* and *plausibility*

these are lower and upper bounds on probabilities of a proposition

Relational DB example

Name	Age
a	[22,26]
b	[20,22]
c	[30,35]
d	[20,22]
e	[28,30]

{20,21,22} is the set of possibilities of Age(d), or the *possibility distribution* of Age(d)

Set membership questions like Age(x)$\in Q$ cease to be applicable; more natural to ask about the possibility of Q given the table above of Age(x)

if Q=[20,25], it is *possible* that $Age(a)\in Q$, *not possible* that $Age(c)\in Q$, *certain* that $Age(d)\in Q$

What is the probability that the age of someone is in the range [20,25]?

belief=2/5; plausibility=3/5. So answer is [.4,.6].

DS combination rule $\rightarrow$ multiple sources

Vague predicates

Not every predicate fits every object exactly (nor fails completely)

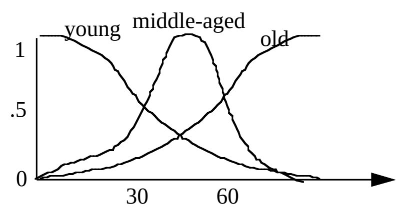
- Categories with degrees of membership
  e.g., fast, old, distant
- Problem: reference sets
  – big fly vs. big elephant

We call predicates that are thought of a holding to a degree vague predicates (or fuzzy predicates).

For each vague predicate, there is a precise base function in terms of which it is understood.

- tall: height
- rich: net worth
- bald: percent hair cover

A degree curve maps the base function to [0,1].



Conjunction and disjunction

As with probabilities, we need boolean combinations of properties

Negation is as with probability:

degree of membership in $\neg P \;=\; 1 -$ degree of membership in P

But handle conjunction with MIN and disjunction with MAX!

Example:

Suppose an individual has very high (.95) degree of membership in predicates Tall, Coordinated, Strong, ... for 20 predicates.

Then want to say very high (.95) degree of membership in (Tall ∧ Coordinated ∧ Strong ∧ ...)

as opposed to

Suppose there is a very high (.95) probability of being Tall, of being Coordinated, of being Strong, ... for 20 predicates.

The probability of being all of them at the same time (Tall ∧ Coordinated ∧ Strong ∧ ...) can be low.

Other operators: "very" = square; "somewhat" = square root

Rules with vague predicates

Imagine degrees of fraud = {high, somewhat high, medium, somewhat low, low}, based on a numeric universe of discourse (to some maximum amount)

Construct a set of rules that indicate degrees of fraud based on authorizations and difference in amount of recorded accountability and actual stock:

1) If number of authorizations is often
   then fraud is somewhat high

2) If amount is larger than usual
   then high fraud

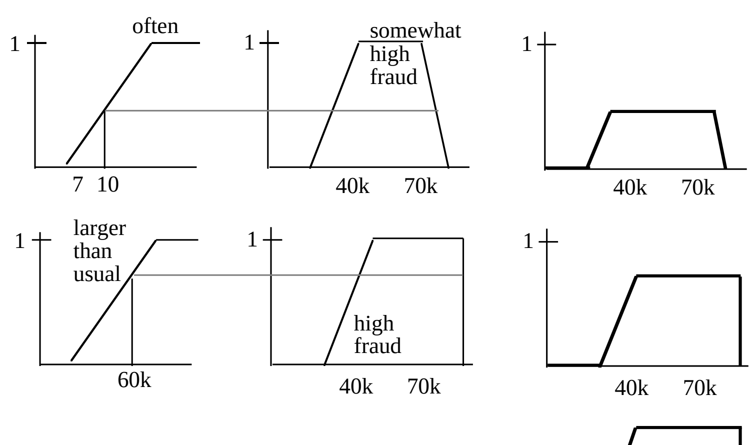
Want to estimate the amount of fraud given inputs

10 authorizations,
amount of $60K

Applying rules

Use degree curves for “somewhat high”, “larger than usual” *etc.*

Can combine with rules in a way that allows conclusion of rule to apply to the degree that the condition of the rule applied.



Given: 10 authorizations
amount of 60k

conclusion:
g =center of gravity