Uncertainty

Introduction

- The world is not a well-defined place.
- There is uncertainty in the facts we know:
  - What's the temperature? Imprecise measures
  - Is Bush a good president? Imprecise definitions
  - Where is the pit? Imprecise knowledge
- There is uncertainty in our inferences
  - If I have a blistery, itchy rash and was gardening all weekend I probably have poison ivy
- People make successful decisions all the time anyhow.

Sources of Uncertainty

- Uncertain data
  - missing data, unreliable, ambiguous, imprecise representation, inconsistent, subjective, derived from defaults, noisy…
- Uncertain knowledge
  - Multiple causes lead to multiple effects
  - Incomplete knowledge of causality in the domain
  - Probabilistic/stochastic effects
- Uncertain knowledge representation
  - restricted model of the real system
  - limited expressiveness of the representation mechanism
- inference process
  - Derived result is formally correct, but wrong in the real world
  - New conclusions are not well-founded (eg, inductive reasoning)
  - Incomplete, default reasoning methods

Reasoning Under Uncertainty

- So how do we do reasoning under uncertainty and with inexact knowledge?
  - heuristics
    - ways to mimic heuristic knowledge processing methods used by experts
  - empirical associations
    - experiential reasoning
    - based on limited observations
  - probabilities
    - objective (frequency counting)
    - subjective (human experience)

Decision making with uncertainty

- **Rational** behavior:
  - For each possible action, identify the possible outcomes
  - Compute the **probability** of each outcome
  - Compute the **utility** of each outcome
  - Compute the probability-weighted **(expected) utility** over possible outcomes for each action
  - Select the action with the highest expected utility (principle of **Maximum Expected Utility**)

Some Relevant Factors

- expressiveness
  - can concepts used by humans be represented adequately?
  - can the confidence of experts in their decisions be expressed?
- comprehensibility
  - representation of uncertainty
  - utilization in reasoning methods
- correctness
  - probabilities
  - relevance ranking
  - long inference chains
- computational complexity
  - feasibility of calculations for practical purposes
- reproducibility
  - will observations deliver the same results when repeated?

Basics of Probability Theory

- mathematical approach for processing uncertain information
  - sample space set
    $X = \{x1, x2, \ldots, xn\}$
    - collection of all possible events
    - can be discrete or continuous
  - probability number $P(xi)$: likelihood of an event xi to occur
    - non-negative value in [0,1]
    - total probability of the sample space is 1
    - for mutually exclusive events, the probability for at least one of them is the sum of their individual probabilities
    - *experimental probability*
      - based on the frequency of events
    - *subjective probability*
      - based on expert assessment

Compound Probabilities

- describes *independent* events
  - do not affect each other in any way
- *joint* probability of two independent events A and B

  $P(A \cap B) = P(A) * P(B)$
- *union* probability of two independent events A and B

  $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
  $$= P(A) + P(B) - P(A) * P(B)$$

Probability theory

- **Random variables**
  - Domain
- **Atomic event**: complete specification of state
- **Prior probability**: degree of belief without any other evidence
- **Joint probability**: matrix of combined probabilities of a set of variables

- Alarm, Burglary, Earthquake
  - Boolean (like these), discrete, continuous
- Alarm=True $\wedge$ Burglary=True $\wedge$ Earthquake=False
  alarm $\wedge$ burglary $\wedge$ earthquake
- P(Burglary) = .1
- P(Alarm, Burglary) =

	alarm	¬alarm
burglary	.09	.01
¬burglary	.1	.8

Probability theory (cont.)

- **Conditional probability**: probability of effect given causes
- **Computing conditional probs**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **normalizing** constant
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b) P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a) P(a)$ (**conditioning**)

- $P(\text{burglary} \mid \text{alarm}) = .47$
  $P(\text{alarm} \mid \text{burglary}) = .9$
- $P(\text{burglary} \mid \text{alarm}) =$
  $P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm})$
  $= .09 / .19 = .47$
- $P(\text{burglary} \wedge \text{alarm}) =$
  $P(\text{burglary} \mid \text{alarm}) P(\text{alarm}) =$
  $.47 * .19 = .09$
- $P(\text{alarm}) =$
  $P(\text{alarm} \wedge \text{burglary}) +$
  $P(\text{alarm} \wedge \neg\text{burglary}) =$
  $.09+.1 = .19$

Independence

- When two sets of propositions do not affect each others' probabilities, we call them **independent**, and can easily compute their joint and conditional probability:
  - Independent (A, B) if $P(A \wedge B) = P(A)\,P(B)$, $P(A \mid B) = P(A)$
- For example, {moon-phase, light-level} might be independent of {burglary, alarm, earthquake}
  - Then again, it might not: Burglars might be more likely to burglarize houses when there's a new moon (and hence little light)
  - But if we know the light level, the moon phase doesn't affect whether we are burglarized
  - Once we're burglarized, light level doesn't affect whether the alarm goes off
- We need a more complex notion of independence, and methods for reasoning about these kinds of relationships

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

- Queries:
  - Is *smart* independent of *study*?
  - Is *prepared* independent of *study*?

Conditional independence

- Absolute independence:
  - A and B are **independent** if $P(A \wedge B) = P(A)\, P(B)$; equivalently, $P(A) = P(A \mid B)$ and $P(B) = P(B \mid A)$
- A and B are **conditionally independent** given C if
  - $P(A \wedge B \mid C) = P(A \mid C)\, P(B \mid C)$
- This lets us decompose the joint distribution:
  - $P(A \wedge B \wedge C) = P(A \mid C)\, P(B \mid C)\, P(C)$
- Moon-Phase and Burglary are ***conditionally independent given*** Light-Level
- Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint probability distribution

Exercise: Conditional independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

- Queries:
  - Is *smart* conditionally independent of *prepared*, given *study*?
  - Is *study* conditionally independent of *prepared*, given *smart*?

Conditional Probabilities

- describes *dependent* events
  - affect each other in some way
- *conditional probability* of event a given that event B has already occurred
  $P(A|B) = P(A \cap B) / P(B)$

Bayesian Approaches

- derive the probability of an event given another event
- Often useful for diagnosis:
  - If X are (observed) effects and Y are (hidden) causes,
  - We may have a model for how causes lead to effects ($P(X \mid Y)$)
  - We may also have prior beliefs (based on experience) about the frequency of occurrence of effects ($P(Y)$)
  - Which allows us to reason abductively from effects to causes ($P(Y \mid X)$).
- has gained importance recently due to advances in efficiency
  - more computational power available
  - better methods

Bayes' Rule for Single Event

- single hypothesis H, single event E
  $P(H|E) = (P(E|H) * P(H)) / P(E)$
  or
- $P(H|E) = (P(E|H) * P(H) /$
  $(P(E|H) * P(H) + P(E|\neg H) * P(\neg H))$

Bayes Example: Diagnosing Meningitis

- Suppose we know that
  - Stiff neck is a symptom in 50% of meningitis cases
  - Meningitis (m) occurs in 1/50,000 patients
  - Stiff neck (s) occurs in 1/20 patients
- Then
  - $P(s|m) = 0.5$, $P(m) = 1/50000$, $P(s) = 1/20$
  - $P(m|s) = (P(s|m)\ P(m))/P(s)$

    $= (0.5 \times 1/50000) / 1/20 = .0002$
- So we expect that one in 5000 patients with a stiff neck to have meningitis.

Advantages and Problems Of Bayesian Reasoning

- advantages
  - sound theoretical foundation
  - well-defined semantics for decision making
- problems
  - requires large amounts of probability data
    - sufficient sample sizes
  - subjective evidence may not be reliable
  - independence of evidences assumption often not valid
  - relationship between hypothesis and evidence is reduced to a number
  - explanations for the user difficult
  - high computational overhead

Some Issues with Probabilities

- Often don't have the data
  - Just don't have enough observations
  - Data can't readily be reduced to numbers or frequencies.
- Human estimates of probabilities are notoriously inaccurate. In particular, often add up to >1.
- Doesn't always match human reasoning well.
  - P(x) = 1 - P(-x). Having a stiff neck is strong (.9998!) evidence that you don't have meningitis. True, but counterintuitive.

- Several other approaches for uncertainty address some of these problems.

Dempster-Shafer Theory

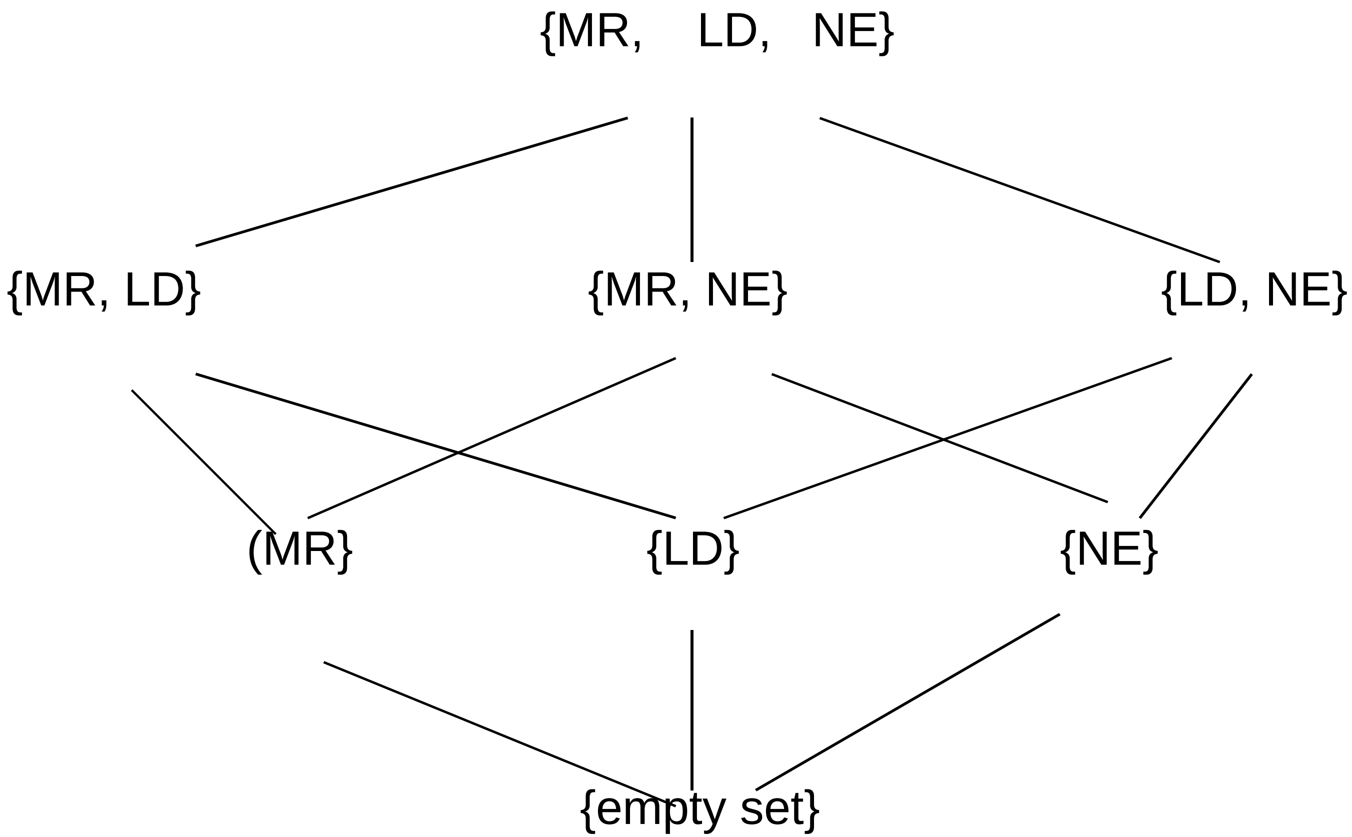
- mathematical theory of evidence
- Notations
  - Environment T: set of objects that are of interest
  - *frame of discernment FD*
    - power set of the set of possible elements
  - mass probability function m
    - assigns a value from [0,1] to every item in the frame of discernment
  - *mass probability* m(A)
    - portion of the total mass probability that is assigned to an element A of FD

D-S Underlying concept

- The most basic problem with uncertainty is often with the axiom that P(X) +P(not X) = 1
  - If the probability that you have poison ivy when you have a rash is .3, this means that a rash is strongly suggestive (.7) that you don’t have poison ivy.
  - True, in a sense, but neither intuitive nor helpful.
- What you really mean is that the probability is .3 that you have poison ivy and .7 that we *don’t know yet* what you have.
- So we initially assign all of the probability to the total set of things you *might* have: the frame of discernment.

Example: Frame of Discernment

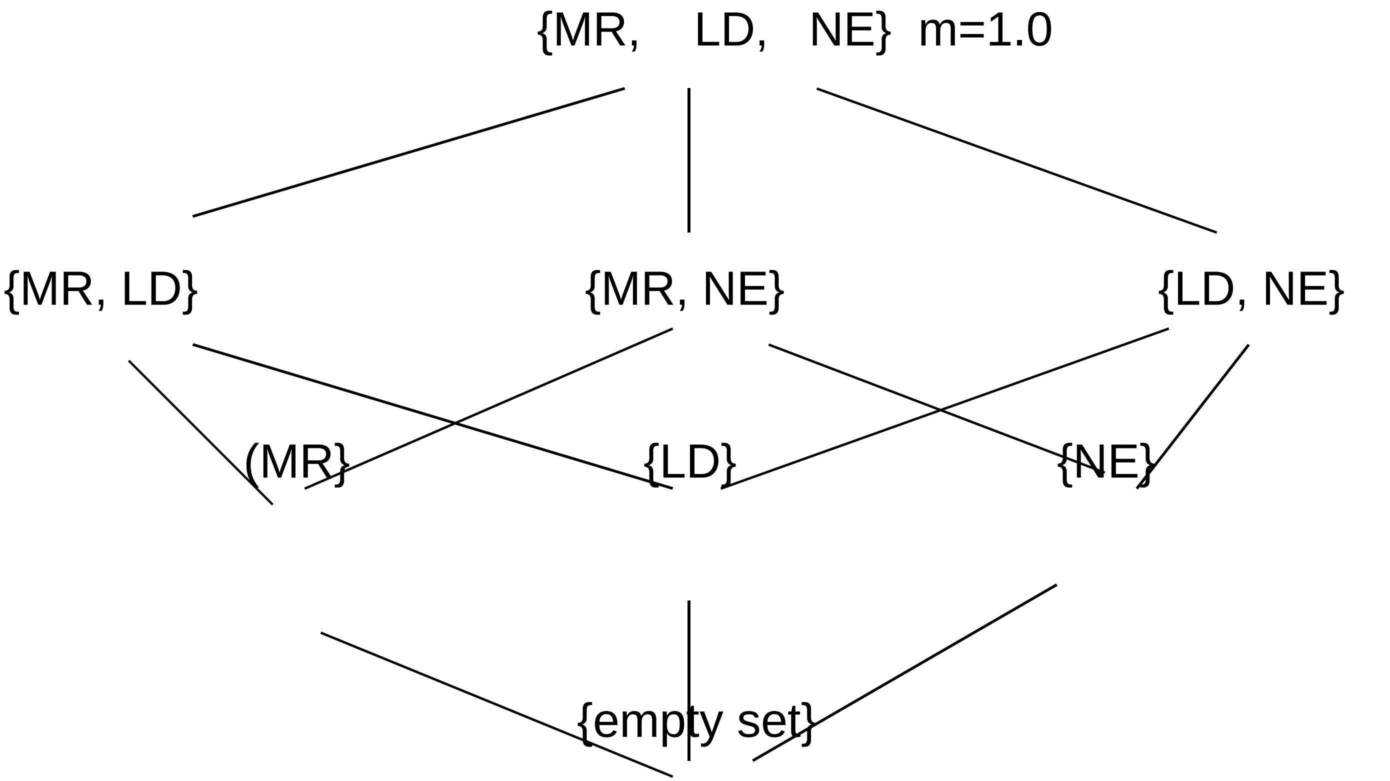
Environment: Mentally retarded (MR), Learning disabled (LD), Not Eligible (NE)



Example: We don't know anything

Frame of Discernment:

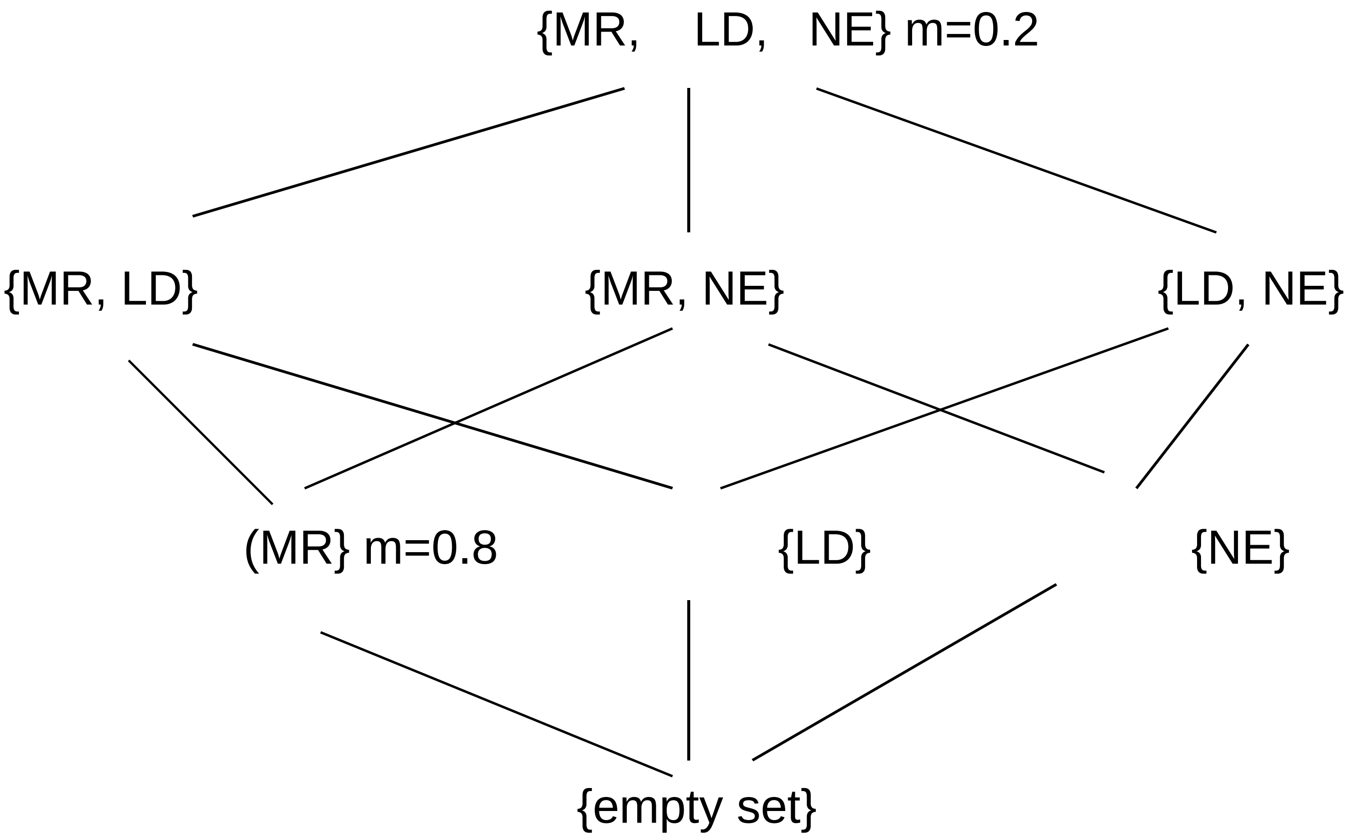
Mentally retarded (MR), Learning disabled (LD), Not Eligible (NE)



Example: We believe MR at 0.8

Frame of Discernment:

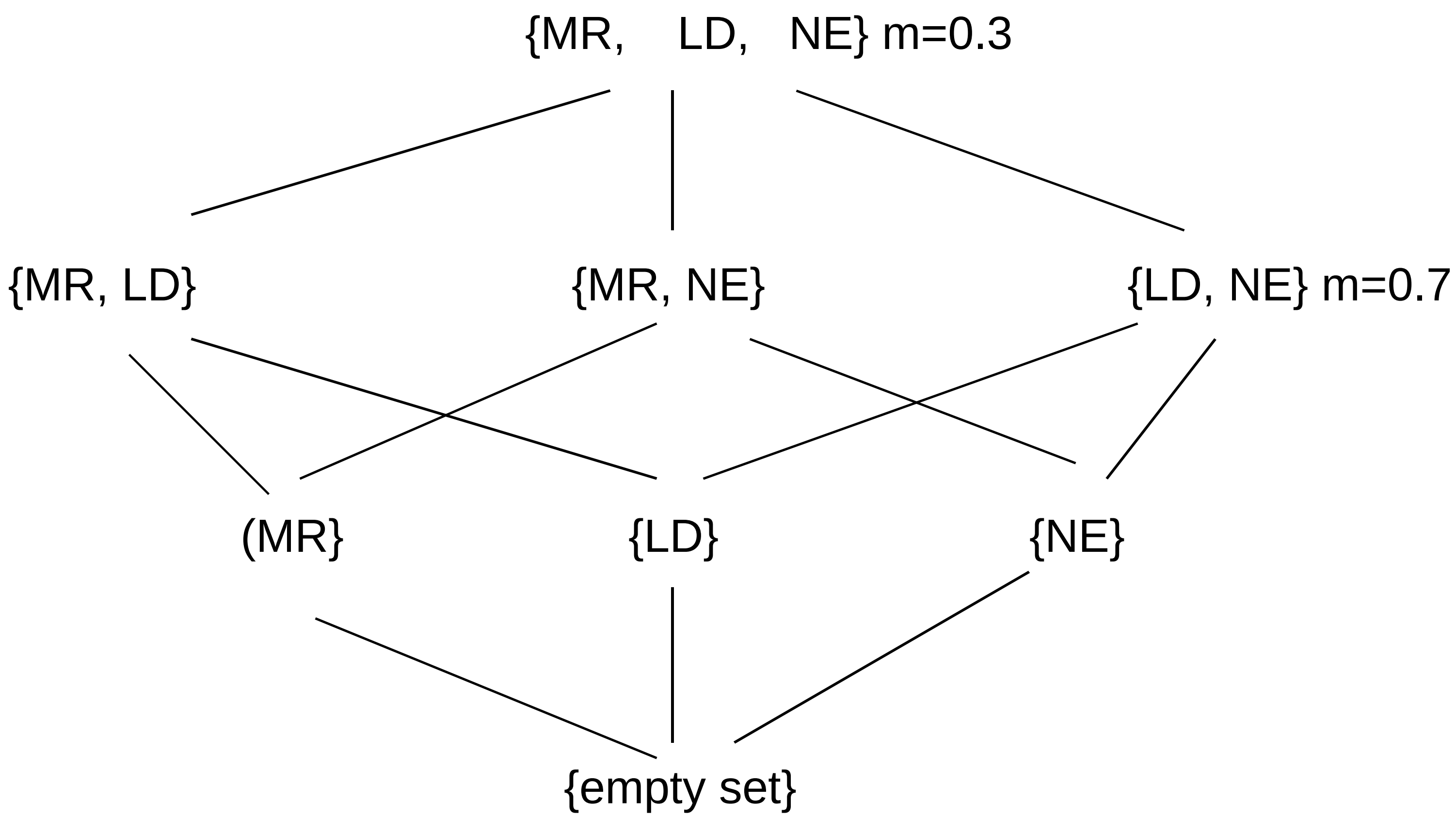
Mentally retarded (MR), Learning disabled (LD), Not Eligible (NE)



Example: We believe NOT MR at 0.7

Frame of Discernment:

Mentally retarded (MR), Learning disabled (LD), Not Eligible (NE)



Belief and Certainty

- belief Bel(A) in a subset A
  - sum of the mass probabilities of all the proper subsets of A
  - likelihood that one of its members is the conclusion
- plausibility Pls(A)
  - maximum belief of A, upper bound
  - 1 – Bel(not A)
- certainty Cer(A)
  - interval [Bel(A), Pls(A)]
  - expresses the range of belief

Example: Bel, Pls

Frame of Discernment:

Mentally retarded (MR), Learning disabled (LD), Not Eligible (NE)

{MR, LD, NE} m=0, Bel=1

{MR, LD}

m=.3, Bel=.6

{MR, NE}

m=.2, Bel = .4

{LD, NE}

m=.1, Bel=.4

(MR}

m=.1, Bel=.1

{LD}

m=.2, Bel=.2

{NE}

m=.1, Bel=.1

{empty set}

m=0, Bel=0

Interpretation: Some Evidential Intervals

- Completely true: [1,1]
- Completely false: [0,0]
- Completely ignorant: [0,1]
- Doubt -- disbelief in X: Dbt = Bel(not X)
- Ignorance -- range of uncertainty: Igr =Pls-Bel
- Tends to support: [Bel, 1] (0<Bel<1)
- Tends to refute: [0, Pls] (0>Pls<1)
- Tends to both support and refute: [Bel, Pls] (0<Bel<Pls<1)

Advantages and Problems of Dempster-Shafer

- advantages
  - clear, rigorous foundation
  - ability to express confidence through intervals
    - certainty about certainty
- problems
  - non-intuitive determination of mass probability
  - very high computational overhead
  - may produce counterintuitive results due to normalization when probabilities are combined
  - Still hard to get numbers

Certainty Factors

- shares some foundations with Dempster-Shafer theory, but more practical
- denotes the belief in a hypothesis H given that some pieces of evidence are observed
- *no statements* about the belief is *no evidence is present*
  - in contrast to Bayes' method

Belief and Disbelief

- measure of belief
  - degree to which hypothesis H is supported by evidence E
  - $MB(H,E) = 1 \quad \text{IF } P(H) = 1$

    $(P(H|E) - P(H)) / (1 - P(H)) \quad \text{otherwise}$
- measure of disbelief
  - degree to which doubt in hypothesis H is supported by evidence E
  - $MB(H,E) = 1 \quad \text{IF } P(H) = 0$

    $(P(H) - P(H|E)) / P(H)) \quad \text{otherwise}$

Certainty Factor

- certainty factor CF
  - ranges between -1 (denial of the hypothesis H) and 1 (confirmation of H)
- $CF = (MB - MD) / (1 - \min(MD, MB))$
- combining antecedent evidence
  - use of premises with less than absolute confidence
    - $E1 \wedge E2 = \min(CF(H, E1), CF(H, E2))$
    - $E1 \vee E2 = \max(CF(H, E1), CF(H, E2))$
    - $\neg E = \neg\ CF(H, E)$

Combining Certainty Factors

- certainty factors that support the same conclusion
- several rules can lead to the same conclusion
- applied incrementally as new evidence becomes available

- Cfrev(CFold, CFnew) =
  - CFold + CFnew(1 - CFold) if both > 0
  - CFold + CFnew(1 + CFold) if both < 0
  - CFold + CFnew / (1 - min(|CFold|, |CFnew|)) if one < 0

Advantages of Certainty Factors

- Advantages
  - simple implementation
  - reasonable modeling of human experts' belief
    - expression of belief and disbelief
  - successful applications for certain problem classes
  - evidence relatively easy to gather
    - no statistical base required

Problems of Certainty Factors

- Problems
  - partially ad hoc approach
    - theoretical foundation through Dempster-Shafer theory was developed later
  - combination of non-independent evidence unsatisfactory
  - new knowledge may require changes in the certainty factors of existing knowledge
  - certainty factors can become the opposite of conditional probabilities for certain cases
  - not suitable for long inference chains

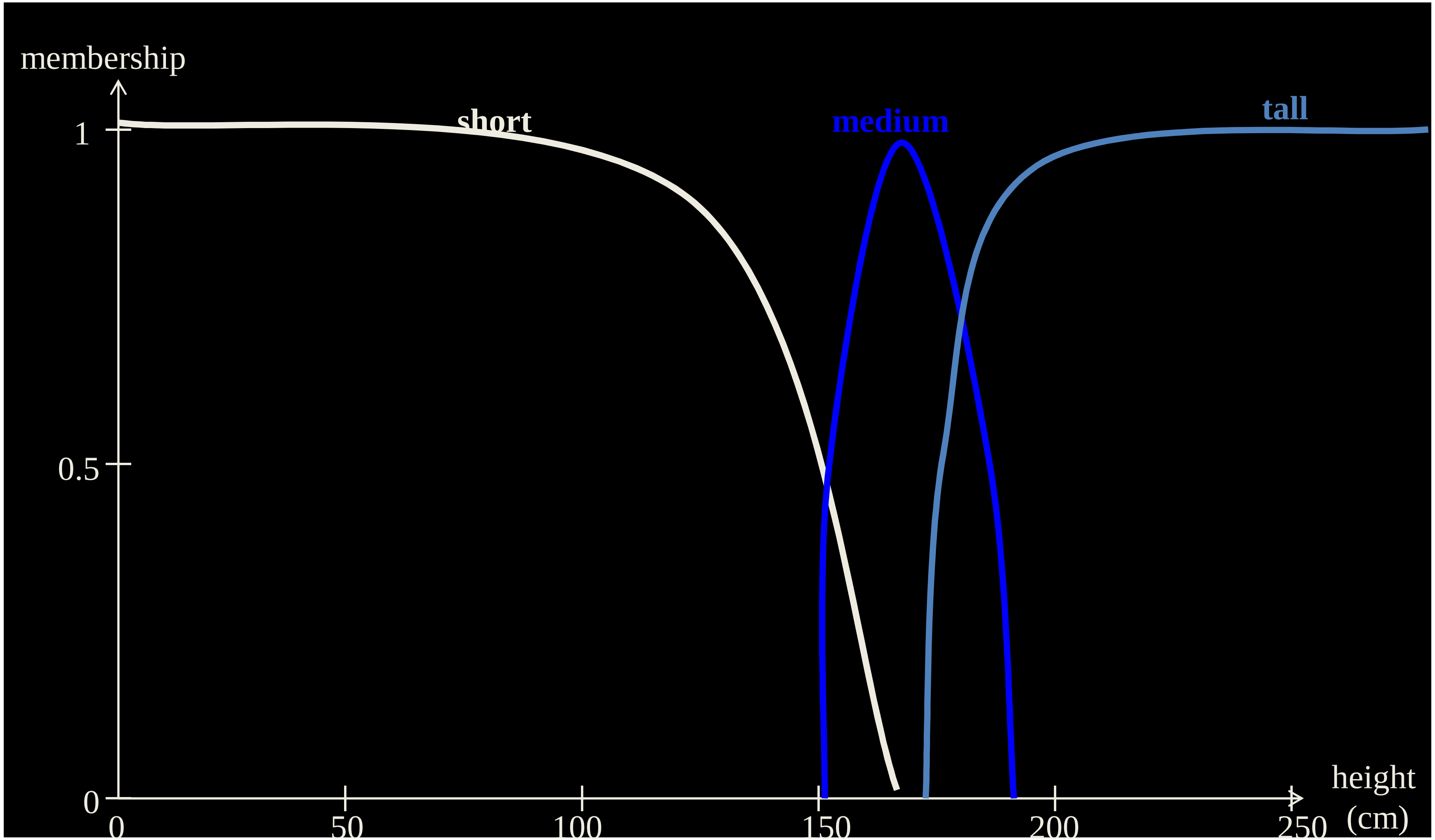
Fuzzy Logic

- approach to a formal treatment of uncertainty
- relies on quantifying and reasoning through natural (or at least non-mathematical) language
- Rejects the underlying concept of an excluded middle: things have a degree of membership in a concept or set
  – Are you tall?
  – Are you rich?
- As long as we have a way to formally describe degree of membership and a way to combine degrees of memberships, we can reason.

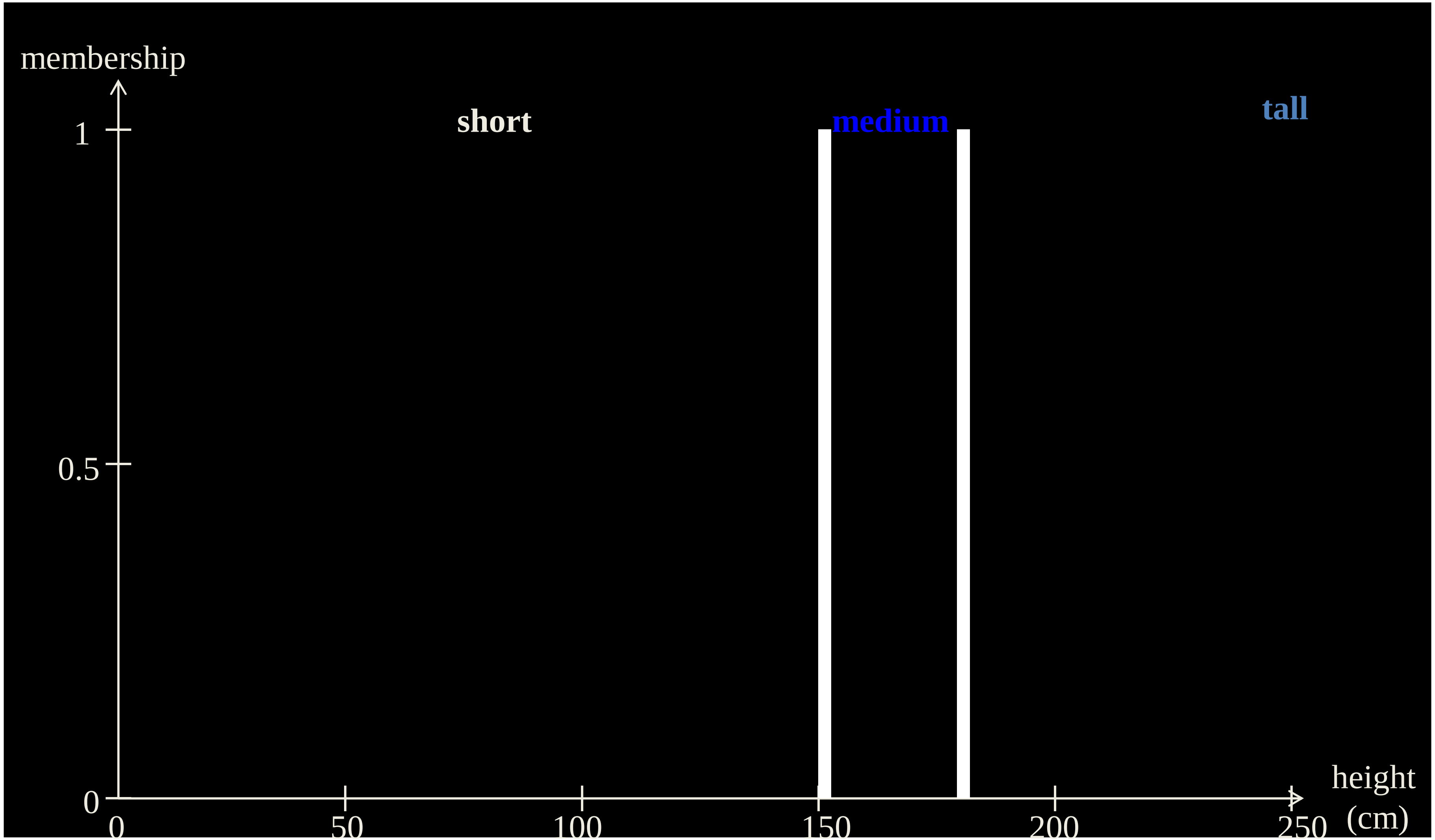
Fuzzy Set

- categorization of elements xi into a set S
  - described through a membership function m(s)
  - associates each element xi with a degree of membership in S
- possibility measure Poss{$x \in S$}
  - degree to which an individual element x is a potential member in the fuzzy set S
  - combination of multiple premises
    - Poss($A \wedge B$) = min(Poss(A),Poss(B))
    - Poss($A \vee B$) = max(Poss(A),Poss(B))

Fuzzy Set Example



Fuzzy vs. Crisp Set



Fuzzy Reasoning

- In order to implement a fuzzy reasoning system you need
  - For each variable, a defined set of values for membership
    - Can be numeric (1 to 10)
    - Can be linguistic
      - really no, no, maybe, yes, really yes
      - tiny, small, medium, large, gigantic
      - good, okay, bad
  - And you need a set of rules for combining them
    - Good and bad = okay.

Fuzzy Inference Methods

- Lots of ways to combine evidence across rules
  - Poss(B|A) = min(1, (1 - Poss(A)+ Poss(B)))
    - implication according to Max-Min inference
  - also Max-Product inference and other rules
  - formal foundation through Lukasiewicz logic
    - extension of binary logic to infinite-valued logic
- Can be enumerated or calculated.

Some Additional Fuzzy Concepts

- Support set: all elements with membership > 0
- Alpha-cut set: all elements with membership greater than alpha
- Height: maximum grade of membership
- Normalized: height = 1

Some typical domains

- Control (subways, camera focus)
- Pattern Recognition (OCR, video stabilization)
- Inference (diagnosis, planning, NLP)

Advantages and Problems of Fuzzy Logic

- advantages
  - general theory of uncertainty
  - wide applicability, many practical applications
  - natural use of vague and imprecise concepts
    - helpful for commonsense reasoning, explanation
- problems
  - membership functions can be difficult to find
  - multiple ways for combining evidence
  - problems with long inference chains

Uncertainty: Conclusions

- In AI we must often represent and reason about uncertain information
- This is no different from what people do all the time!
- There are multiple approaches to handling uncertainty.
- Probabilistic methods are most rigorous but often hard to apply; Bayesian reasoning and Dempster-Shafer extend it to handle problems of independence and ignorance of data
- Fuzzy logic provides an alternate approach which better supports ill-defined or non-numeric domains.
- Empirically, it is often the case that the main need is some way of expressing "maybe". Any system which provides for at least a three-valued logic tends to yield the same decisions.