Vagueness, uncertainty and degrees of belief

In some situations, the logical universals are not appropriate to express general statements.

Besides the intrinsic imprecision of statements like “someone is somewhat tall”, the way the conclusions are formulated may also be imprecise (e.g., in the medical field, a rule may not be applicable in 100% of cases).

We distinguish three ways to “relax” the categorical nature of classical logic, in order to make the universal $\forall x.P(x)$ more flexible:

These slides are mainly based on Ronald J. Brachman, Hector J. Levesque. Knowledge representation and reasoning, Morgan Kaufmann, 2004.

Noncategorical reasoning

1. Relaxation of the strength of the quantifier – instead of “for all x” we say “for % of x”

   95% of the persons in this group are master students.

   This is a statistical interpretation. The use of probability in these sentences is objective (it is an assertion about the frequency of an event, it is not a subjective interpretation or a degree of confidence).

Noncategorical reasoning

1. Relaxation of the strength of the quantifier – instead of “for all x” we say “for % of x”

   95% of the persons in this group are master students.

   This is a statistical interpretation. The use of probability in these sentences is objective (it is an assertion about the frequency of an event, it is not a subjective interpretation or a degree of confidence).

2. Relaxation of the applicability of a predicate – instead of a statement like “Everyone in my group is (absolutely) tall”, we say

   “Everyone in my group is (moderately) tall.”

   The predicate “tall” applies to an individual to a greater or lesser extent. These predicates are called vague.

   A person may be considered both tall (moderately) and short (weakly) at the same time.

Noncategorical reasoning

1. Relaxation of the strength of the quantifier – instead of “for all x” we say “for % of x”

   95% of the persons in this group are master students.

   This is a statistical interpretation. The use of probability in these sentences is objective (it is an assertion about the frequency of an event, it is not a subjective interpretation or a degree of confidence).

2. Relaxation of the applicability of a predicate – instead of a statement like “Everyone in my group is (absolutely) tall”, we say

   “Everyone in my group is (moderately) tall.”

   The predicate “tall” applies to an individual to a greater or lesser extent. These predicates are called vague.

   A person may be considered both tall (moderately) and short (weakly) at the same time.

3. Relaxation of the degree of belief in a sentence – instead of having the statement “Everyone in this group is a master student” we say

   “I believe that everyone in this group is a master student, but I am not sure.”

   This is uncertain knowledge and it can be quantified by using the concept of subjective probabilities.

Objective/subjective probability

Objective probability

It refers to the frequency of a single event happening and it does not depend on who is assessing the probability.

It is best applied to situations like “coin flipping” or “card drawing”.

Subjective probability

The degree of confidence (also called subjective probability) in a sentence is separable from the content of the sentence. The degree of belief in a sentence can vary, regardless of how vague or categorical the sentence may be.

For example, we may be absolutely certain that Bill is quite tall, while we may only suspect that he is married.

With subjective beliefs, we express degrees of confidence rather than black-and -white conclusions.

Subjective probability

Subjective probabilities can be mechanically computed like the objective ones, but they are used in a different way. We are interested in how evidences combine to change our degree of confidence in a belief, rather than simply deriving new conclusions.

Def. The prior probability of a sentence α involves the prior state of information (or background knowledge) β. We write $Pr(\alpha|\beta)$.

For example, suppose that we know that 0.2% of the population has hepatitis. Based on just that, our degree of belief that John (a randomly chosen individual) has hepatitis is 0.002.

Def. A posterior probability is derived when new evidence is considered:

$$Pr(\alpha|\beta\wedge\gamma), \text{ where } \gamma \text{ is the new evidence.}$$

For example, if John is yellowish, given the symptoms and the prior probability, we may conclude that the posterior probability of John having hepatitis is 0.65.

The key problem is how to combine evidences from different sources to reevaluate our beliefs.

A basic Bayesian approach

Suppose that we have a number of atomic sentences of interest $p_1, \ldots, p_n$ (e.g., Eric is tall, Anna is married, George is a teacher). In different interpretations, different combinations of these sentences will be true.

Let $\mathcal{I}$ be an interpretation that specifies which sentences are true/false.

Def. The joint probability distribution $\mathcal{J}$ is the specification of the degree of belief for each of the 2^n truth assignments

$$\sum_{\mathcal{I}} \mathcal{J}(\mathcal{I})=1 \text{ and } \mathcal{J}(\mathcal{I})\in[0,1].$$

The degree of belief in any sentence α is defined as

$$\Pr(\alpha)= \sum_{\mathcal{I}\models\alpha} \mathcal{J}(\mathcal{I})$$

Knowing that $\Pr(\alpha|\beta) = \frac{\Pr(\alpha\wedge\beta)}{\Pr(\beta)}$, the degree of belief that John is tall given that he is male from California is

$$\frac{\Pr(\text{John is tall, John is male, John is from California})}{\Pr(\text{John is male, John is from California})}$$

For n atomic sentences, we need to specify 2^n-1 values (unachievable for most of the practical applications).

Belief (or Bayesian) networks

Suppose that we have the atomic sentences $p_1,\ldots,p_n$.

We can specify an interpretation using the notation $<P_1,\ldots,P_n>$, where each P_i is p_i (when the sentence is true) or $\neg p_i$ (when the sentence is false). We have that

$$\mathcal{J}(<P_1,\ldots,P_n>)=Pr(P_1\wedge\ldots\wedge P_n)$$

because there is only one interpretation that satisfies $P_1\wedge\ldots\wedge P_n$.

We represent all the variables P_i in a directed **acyclic** graph, called a belief (or Bayesian) network.

There is an arc from P_i to P_j if the truth of P_i directly affects the truth of P_j (the former is a parent of the latter).

We assume that the variables are numbered such that the parents of any P_j appears earlier in the sequence than P_j (we can do that because the graph is acyclic).

Belief (or Bayesian) networks

According to the chain rule in probabilities, we have:

$$\mathcal{J}(\langle P_1,\ldots,P_n\rangle)=Pr(P_1)\cdot Pr(P_2|P_1)\cdot\ldots\cdot Pr(P_n|P_1\wedge\ldots\wedge P_{n-1})$$

To compute the joint probability distribution, we still need 2^n-1 values because for each term $Pr(P_{j+1}|P_1\wedge\ldots\wedge P_j)$ there are 2^j conditional probabilities to specify and $\sum_{j=0}^{n-1} 2^j = 2^n - 1$.

In order to reason about subjective probabilities, some simplifying assumptions are necessary.

Belief (or Bayesian) networks

We will assume that each propositional variable in the belief network is conditionally independent from the nonparent variables, given the parent variables.

$$\Pr(P_{j+1} | P_1 \wedge \ldots \wedge P_j) = \Pr(P_{j+1} | \mathrm{parents}(P_{j+1}))$$

With these independence assumptions, it follows that

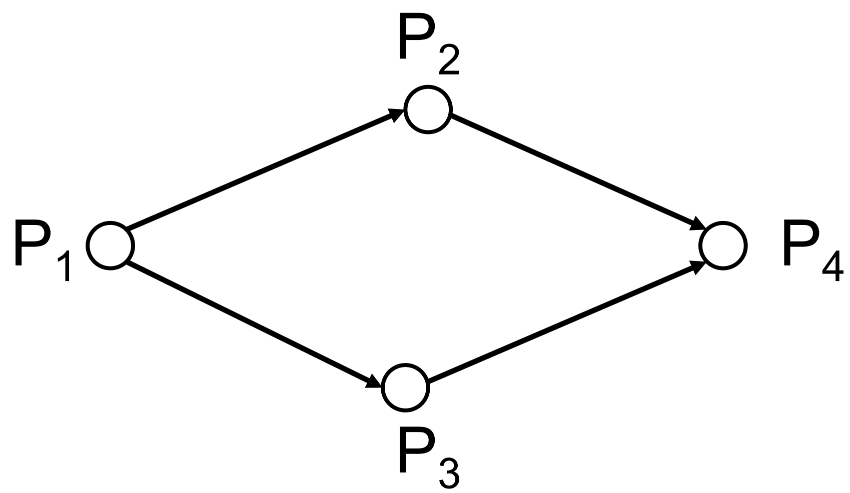
$$\mathcal{J}(<P_1, \ldots, P_n>) = \Pr(P_1) \cdot \Pr(P_2 | \mathrm{parents}(P_2)) \cdot \ldots \cdot \Pr(P_n | \mathrm{parents}(P_n))$$

To fully specify $\mathcal{J}$, we need to know $\Pr(P|\mathrm{parents}(P))$ for each variable P.

If k is the maximum number of parents for any node, then we have no more than $n \cdot 2^k$ values to specify.

Belief (or Bayesian) networks

For the following belief network



we have $\mathcal{J}(\langle P_1,P_2,P_3,P_4\rangle)=Pr(P_1)\cdot Pr(P_2|P_1)\cdot Pr(P_3|P_1)\cdot Pr(P_4|P_2 \wedge P_3)$

We now need (1+2+2+4)=9 values rather than 15 (without the independence assumption) to compute $\mathcal{J}$.

Belief (or Bayesian) networks

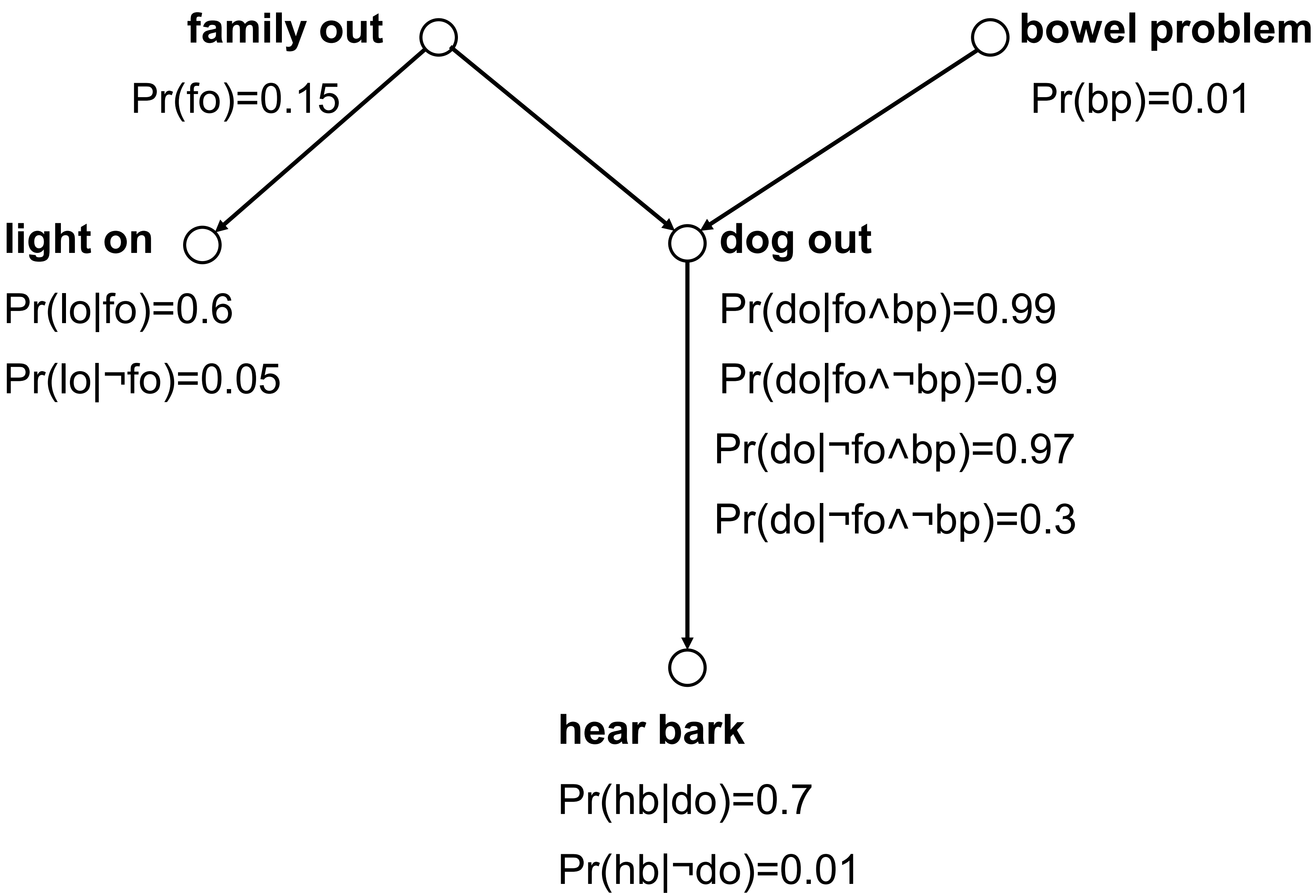
Example (due to Eugene Charniak)

We have a family with a dog. We usually put the dog out (do) when the family is out (fo). We also put the dog out when it has bowel problem (bp). A reasonable proportion of time when the dog is out, you can hear it barking (hb) when you approach the house. We usually leave the light on (lo) outside the house when the family is out.

We are required to calculate the probability that the family is out, given that the light is on and we don’t hear barking.

Belief (or Bayesian) networks

Using the facts, we can construct the following belief network:



We assume the following about the joint probability distribution:

$$\mathcal{J}(\langle FO,LO,BP,DO,HB\rangle)=Pr(FO)\cdot Pr(LO|FO)\cdot Pr(BP)\cdot Pr(DO|FO\wedge BP)\cdot Pr(HB|DO)$$

We need 1+2+1+4+2=10 values to specify the joint probability distribution.

Belief (or Bayesian) networks

The probability that the family is out, given that the light is on and we don't hear barking is:

$$\Pr(\mathrm{fo}|\mathrm{lo}\wedge\neg\mathrm{hb}) = \frac{\Pr(\mathrm{fo}\wedge\mathrm{lo}\wedge\neg\mathrm{hb})}{\Pr(\mathrm{lo}\wedge\neg\mathrm{hb})} = \frac{\sum_{BP,DO} J(<\mathrm{fo},\mathrm{lo},\mathrm{BP},\mathrm{DO},\neg\mathrm{hb}>)}{\sum_{FO,BP,DO} J(<\mathrm{FO},\mathrm{lo},\mathrm{BP},\mathrm{DO},\neg\mathrm{hb}>)}$$

1. $\mathcal{J}$ (<fo,lo,bp,do,¬hb>)=Pr(fo)·Pr(lo|fo)·Pr(bp)·Pr(do|fo ∧bp)·(1-Pr(hb|do))

   =0.15·0.6·0.01·0.99·0.3

2. $\mathcal{J}$ (<fo,lo,bp,¬do,¬hb>)=Pr(fo)·Pr(lo|fo)·Pr(bp)·(1-Pr(do|fo ∧bp))·(1-Pr(hb|¬do))

   =0.15·0.6·0.01·0.01·0.99

3. $\mathcal{J}$ (<fo,lo,¬bp,do,¬hb>)=Pr(fo)·Pr(lo|fo)·(1-Pr(bp))·Pr(do|fo ∧¬bp)·(1-Pr(hb|do))

   =0.15·0.6·0.99·0.9·0.3

4. $\mathcal{J}$ (<fo,lo,¬bp,¬do,¬hb>)=Pr(fo)·Pr(lo|fo)·(1-Pr(bp))·(1-Pr(do|fo ∧¬bp))·(1-Pr(hb|¬do))

   =0.15·0.6·0.99·0.1·0.99

$\mathcal{J}$ (<FO,LO,BP,DO,HB>)=Pr(FO)·Pr(LO|FO)·Pr(BP)·Pr(DO|FO ∧BP)·Pr(HB|DO)

Belief (or Bayesian) networks

5. $\mathcal{J}$ (<¬fo,lo,bp,do,¬hb>)=(1-Pr(fo))·Pr(lo|¬fo)·Pr(bp)·Pr(do|¬fo ∧bp)·(1-Pr(hb|do))

=0.85·0.05·0.01·0.97·0.3

6. $\mathcal{J}$ (<¬fo,lo,bp,¬do,¬hb>)=(1-Pr(fo))·Pr(lo|¬fo)·Pr(bp)·(1-Pr(do|¬fo ∧bp))·(1-Pr(hb|¬do))

=0.85·0.05·0.01·0.03·0.99

7. $\mathcal{J}$ (<¬fo,lo,¬bp,do,¬hb>)=(1-Pr(fo))·Pr(lo|¬fo)·(1-Pr(bp)) Pr(do|¬fo ∧¬bp)·(1-Pr(hb|do))

=0.85·0.05·0.99·0.3·0.3

8. $\mathcal{J}$ (<¬fo,lo,¬bp, ¬do,¬hb>)=(1-Pr(fo))·Pr(lo|¬fo)·(1-Pr(bp))·(1-Pr(do|¬fo ∧¬bp))·
·(1-Pr(hb|¬do))

=0.85·0.05·0.99·0.7·0.99

So, $\mathrm{Pr}(\mathrm{fo}|\mathrm{lo}\wedge\neg\mathrm{hb}) = \frac{1.+2.+3.+4.}{1.+2.+3.+4.+5.+6.+7.+8.}$

$\mathcal{J}$ (<FO,LO,BP,DO,HB>)=Pr(FO)·Pr(LO|FO)·Pr(BP)·Pr(DO|FO ∧BP)·Pr(HB|DO)

Belief (or Bayesian) networks

To do:

We are given the following KB:

> Mary's apartment has an alarm. The alarm is quite reliable in case of a burglary but also if there is smoke in the apartment. Mary has a neighbor, John, who promised to call 112 when he hears the alarm.

Using these facts, draw the Bayesian belief network, express the joint probability distribution indicating which and how many values are necessary to specify it. Write all these values on the graph.

Compute the probability of John calling 112, given that there is smoke in the apartment.

Decision networks (influence diagrams)

They are general decision mechanisms that combine Bayesian networks with additional node types for actions and utilities.

A decision network represents information about an agent's current state, its possible actions, the resulting state from the agent's action and the utility (value) of that state.

There are three types of nodes:

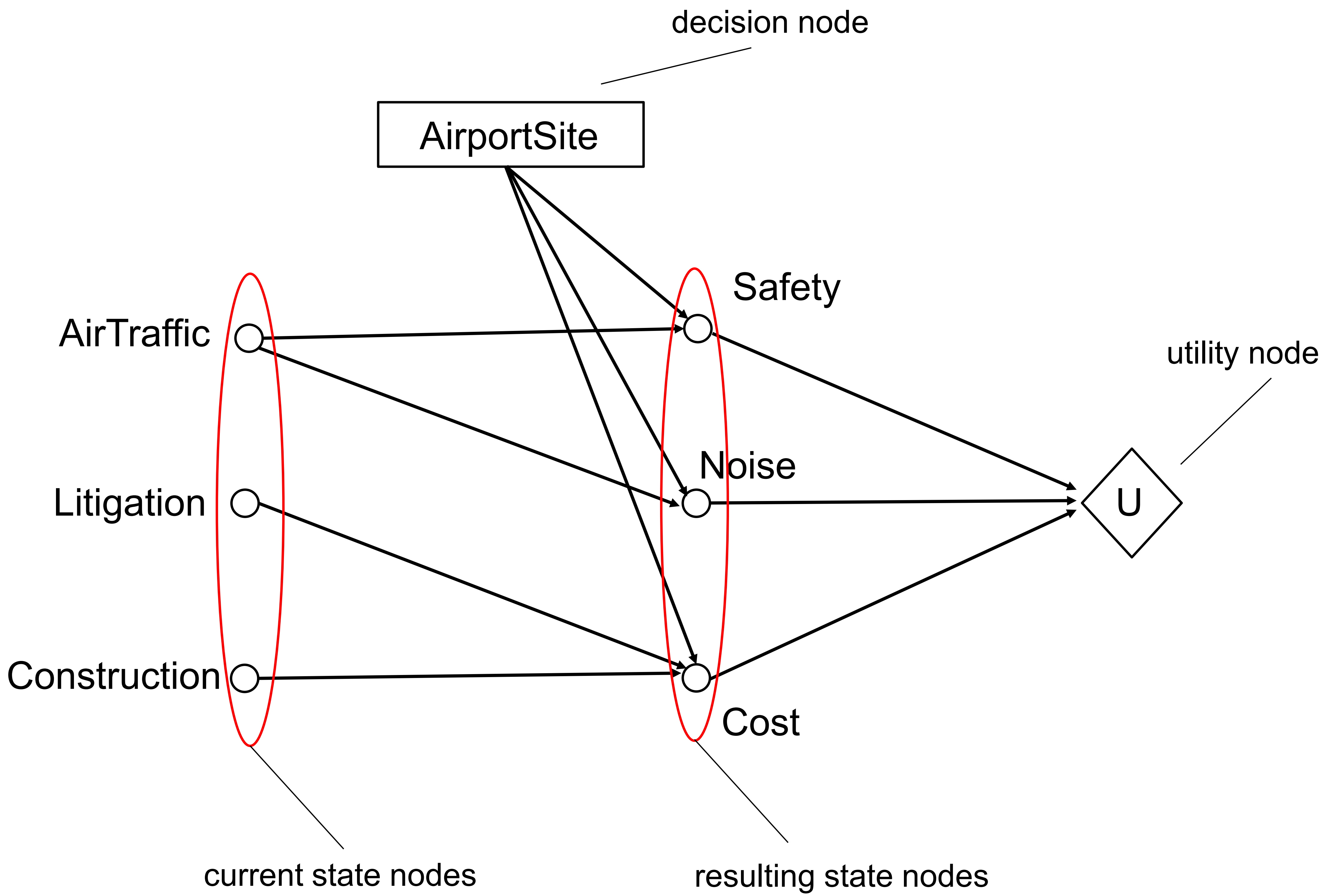
-<u>chance nodes</u> (as circles) represent probabilistic variables, just like they do in Bayesian networks. In decision networks, the parent nodes can include chance nodes as well as decision nodes.

-<u>decision nodes</u> (as rectangles) represent decisions made by the agent. In the following figure, the AirportSite action can take on a different value for each site under consideration. The choice influences the cost, safety and noise that will result.

-<u>utility (value) node</u> (as diamond) represents the agent's utility function – there is only one such node. It has parents all variables describing the outcome that directly affect utility. The utility is expressed as a function of the parents attributes.

See also Ch. 16.5 – Decision Networks in Stuart J. Russell, Peter Norvig. Artificial Intelligence – a modern approach, 3rd edition, Pearson, 2010.

Decision networks (influence diagrams)



Decision networks (influence diagrams)

Evaluating decision networks

For each possible value of the decision node, the resulting utility is calculated. The action (decision) with the highest utility (value) will be chosen.

The algorithm is the following:

1. set the evidence variables for the current state
2. for each possible value of the decision node:
   - -calculate the posterior probabilities for the parent nodes of the utility node
   - -calculate the resulting utility for that action
3. return the action with the highest utility

Representing ignorance

The Dempster-Shafer theory

It is designed to deal with the distinction between uncertainty and ignorance. Rather than computing the probability of a proposition, it computes a lower and an upper bound on the probability of a proposition.

If we have an unbiased coin, the degree of belief that we get heads if we flip it is 0.5.

If we have a biased coin, due to lack of information, we may want to say only that the degree of belief lies between some limits within [0,1].

The limits are called belief and plausibility.

For an unbiased coin, we have 0.5 belief and 0.5 plausibility that the result is heads.
For an unknown coin, we have 0 belief that we get heads and 1 plausibility.

Thus, the value of a propositional variable is represented by a range, called the possibility distribution of the variable.

Representing ignorance

Example – suppose that we have a database with names of people and their believed ages. In the case of complete knowledge, the ages would be values. But if we don't know the exact age, we may specify it by a range.

Mary	[18,22]
Anna	[20,24]
George	[35,40]
David	[27,33]
Cris	[20,23]

Given an interval Q, we ask about the possibility of age(x)∈Q.

For example, if Q=[19,24] then it is possible that age(Mary)∈Q; it is not possible that age(George)∈Q; and it is certain that age(Cris)∈Q.

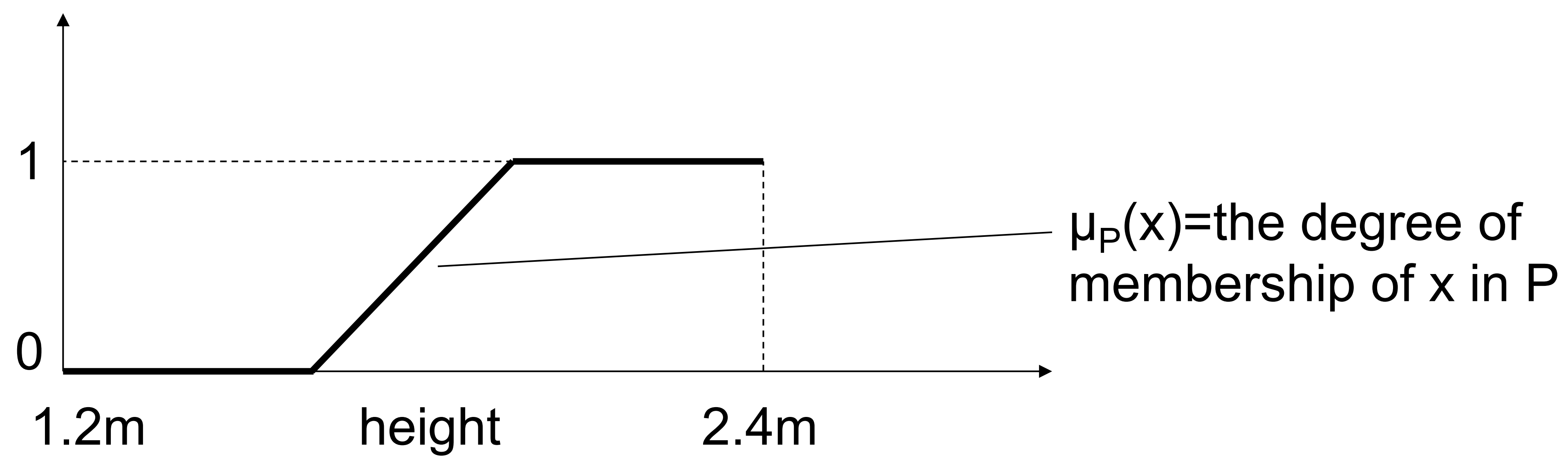
Consider now that we ask what is the probability that the age of a randomly selected individual is in Q?

The belief in this proposition is 2/5 (because of Anna and Cris) and the plausibility is 3/5 (Mary, Anna and Cris). So, the answer is [0.4,0.6].

Vagueness

It refers to the degree to which certain predicates are satisfied. For each vague predicate, there is a corresponding base function in terms of which the predicate is understood (for “tall” the base function is “height”).

Def. The degree curve is a function that capture the relationship between a vague predicate and its base function.



Vagueness

An important thing is that an object's degree of satisfaction can be nonzero for multiple predicates over the same base function (e.g. "short" and "tall").

Negation, conjunction and disjunction of vague predicates:

$$\mu_{\neg P}=1-\mu_P$$

$$\mu_{P\wedge Q}=\min(\mu_P, \mu_Q)$$

$$\mu_{P\vee Q}=\max(\mu_P, \mu_Q)$$

In a typical application, called fuzzy control, vague predicates are used in production rules.

Unlike standard production systems, where a rule either applies or not, here the antecedent of a rule will apply to some degree and the action will be affected to a proportional degree. Such a system enables inferences even when the antecedent conditions are only partially satisfied.

Vagueness

Example – we are given the following rules:

1. If the service is poor or the food is rancid then the tip is cheap.
2. If the service is good then the tip is normal.
3. If the service is excellent or the food is delicious then the tip is generous.

Assume that service and food quality are described by numbers on a linear scale (e.g. a number from 0 to 10). The amount of tip is represented as a percentage of the cost of the meal (e.g.10%).

For each of the eight predicates in the example, we are given a degree curve. The base functions are: service, food quality or tip.

Problem: Given the ratings for the service and for the food, calculate the tip, subject to the rules above.

e.g. service=3,food=8, tip=?

Vagueness

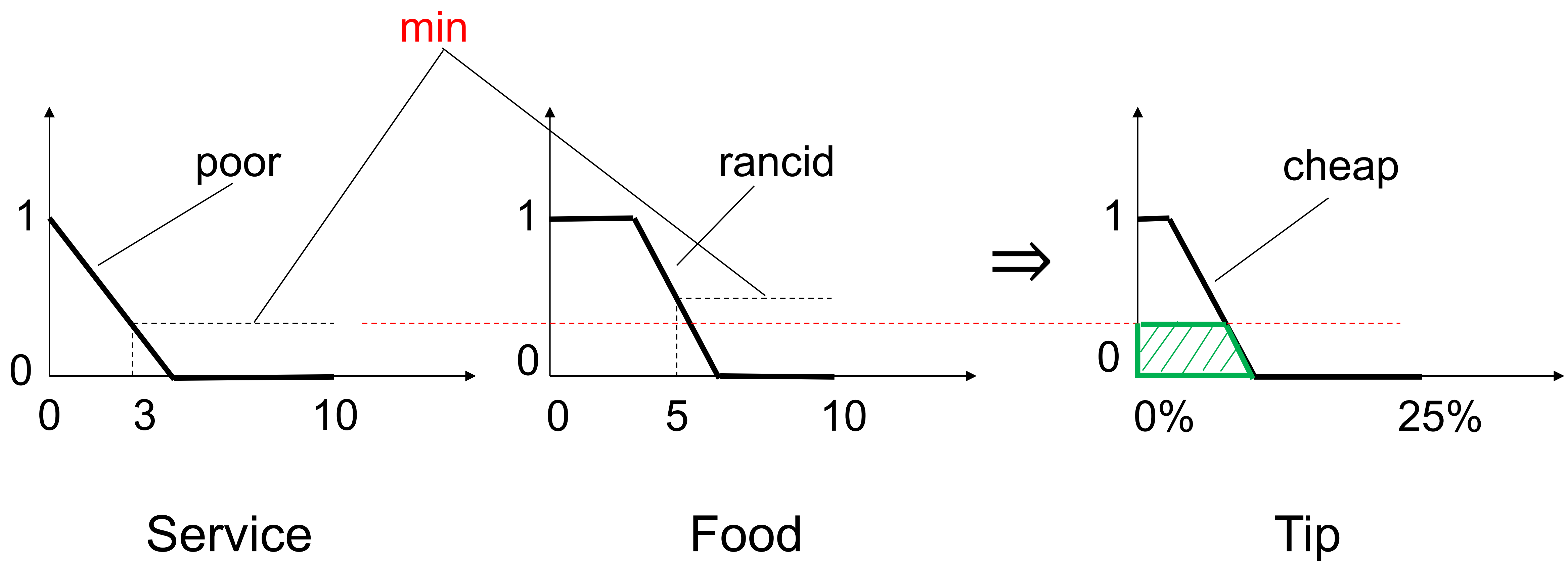
Algorithm

1. Transform the inputs into the degrees to which each of the vague predicates used in the antecedents hold
2. Evaluate the antecedents – combine the degrees of applicability of all the predicates in the antecedent of a rule
3. Evaluate the consequents by determining the degrees to which the predicates of each consequent side should be satisfied. An intuitive way is that the consequent should hold only to the degree that the rule is applicable
4. Aggregate the consequents – obtain a single degree curve for the “tip” base function
5. Defuzzify the output – generate a value for the tip from the aggregated degree curve at step 4. One way to do that is to take the center of the area under the curve.

Vagueness

Example of reasoning

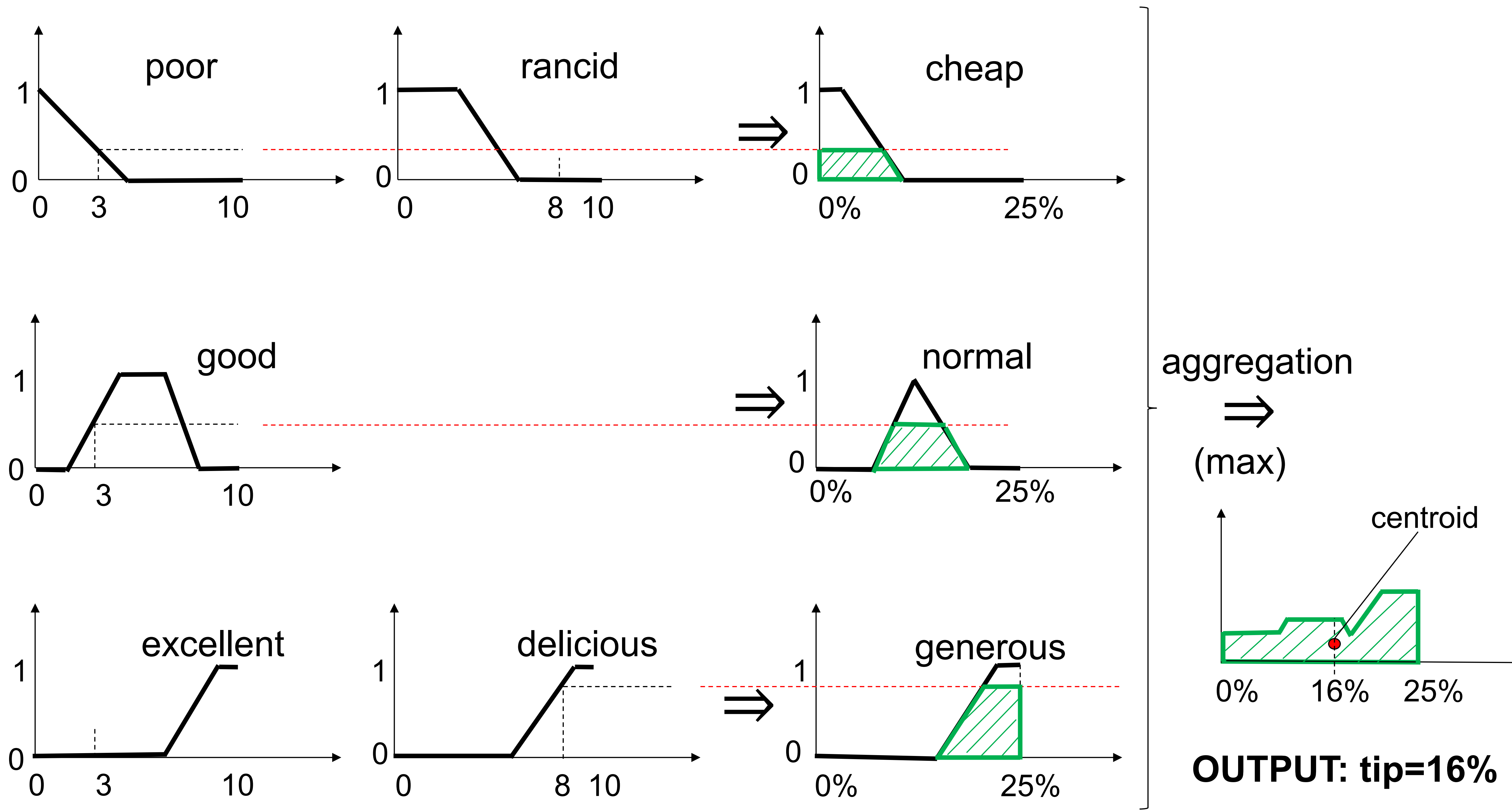
If the service is poor (3) and the food is rancid (5) then the tip is cheap.



Vagueness

The example from slide 24

1. If the service is poor or the food is rancid then the tip is cheap.
2. If the service is good then the tip is normal.
3. If the service is excellent or the food is delicious then the tip is generous.



OUTPUT: tip=16%

INPUT: service=3 **food=8**

Vagueness

To do:

We are given the following rules:

1. If learning effort is high or subjects are easy then grade is high.
2. If learning effort is moderate then grade is good.
3. If learning effort is not high and subjects are difficult then grade is low.

The learning effort, subjects and the grades are represented by numbers on a 10-point scale.

Which are the base functions in this example?
Choose/draw a degree curve for each predicate in the rules (high, moderate, …)

Given the values for learning effort and subjects, compute the grade (based on the degree curves that you choose for the vague predicates and using the three rules above).
Example: learning effort =4, subjects (difficulty) =6, grade=?