

Lesson 8: Dealing With Uncertainty

Introduction

Up until now, the fundamental mathematical basis for all of our expert system discussion has been *logic*. Every fact has been true or false, every rule either correct or a bug. In this lesson, I introduce a new basis for developing expert systems: probability. I show how uncertain facts and rules can be represented, and how they can be used to update a system's assessment of conclusions' likelihoods.

sidebar begins

Required Readings:

Chapter 9 (Representing Uncertainty), S 3.3.2 (MYCIN's knowledge base)

sidebar ends

By the end of this lesson, you will have had your memory of probability refreshed and renewed. You will have seen how to apply probability theory to evidentiary reasoning using probabilistic rules. You should understand how this can be implemented using certainty factors. You will also have a basic idea of what fuzzy logic is and how it might be used in an expert system.

Probabilistic Reasoning

You already know that the power of rule-based languages is greater than that of first-order logic. Logic is *monotonic*: the total number of true statements can only increase. In logic, when a new statement is inferred, that of course means it is true. It can *never* be falsified; logic contains no notion of time, and so if we conclude P and later conclude $\neg P$, we have a contradiction.

On the other hand, expert systems *do* allow for falsification of previously asserted facts: they can be retracted. This allows us to tentatively propose a conclusion and later rescind it based on new information. That also allows us to build systems which use high-level rules propose hypotheses that are later tested (and perhaps retracted) by more specialized rules. This should be familiar to you by now. But, the fact of the matter is that asserting a statement indicates that it is true and retracting it says that it is false.

What if we want to say something like, “the probability of the patient having tuberculosis is 5%”? We could write something like (disease (name tuberculosis) (probability 0.05)). If pretty much all of our facts can admit probabilities, then we'll need a probability slot for every one! Now, let's say we have a rule that states, “If someone has tuberculosis then they have an 80% chance of coughing”. How do we combine rules, facts, and probabilities? That is the major topic of this lesson. Let's start off with a review of the basics of probability theory, then I'll show you how to extend rules to become probabilistic.

Representation

A probabilistic representation associates with each sentence (either declarative fact or rule) a number between zero (certainly false; an event that can never occur) and one (certainly true; an event

that always happens). This number is called the *probability* of the sentence being true. We can indicate this by using a logic-like notation or a functional notation (the latter being most likely what you've seen before when you learned about probability). So, we can write:

Uncertain fact $Rain(\text{Tomorrow}), 0.3$ or $P(Rain(\text{Tomorrow})) = 0.3$.

Certain fact $bird(\text{Tweety}), 1$ or $P(bird(\text{Tweety})) = 1$.

Uncertain conjunction $tuberculosis \wedge coughing, 0.04$ or $P(tuberculosis, coughing) = 0.04$.

Certain conjunction $executive(\text{Bill}) \wedge Writer(\text{Jack}), 1$ or $P(executive(\text{Bill}), Writer(\text{Jack})) = 1$.

Uncertain rule $P(flies(x)|bird(x)) = 0.98$.

Certain rule $P(bird(x)|ostrich(x)) = 1$.

You'll notice in the last two examples that probabilistic rules are usually expressed as conditional probabilities. The *condition probability* $P(p|q)$ can be read "the probability of p given q "; in other words, the chance that p is true *if we assume* (or know) that q is true ($P(q) = 1$). This is *not* the same as $P(p \rightarrow q)$, as you can see from the following example.

Example: Conditional probability versus implication

Let's say that a person can be either old or young, but not both (these are non-overlapping categories; we'll revisit the issue of someone being a bit of both, in the section on fuzzy logic). So, if we know someone is young, then we know they're not old: $P(old(x)|young(x)) = 0$. The structurally similar logic-based expression would be $P(old(x) \rightarrow young(x)) = P(\neg old(x) \vee young(x))$, the probability that someone is either not old or is young. If "not old" implies "young" (which is what we've assumed), then this is just the probability that someone is young. Depending on our cutoff age for "young", we might say that this probability is something like 0.7.

Alternatively, we might write a rule that tries to capture the gist of the conditional expression: $young(x) \rightarrow \neg old(x)$. The probability of this is just $P(\neg young(x) \vee \neg old(x)) = P(\neg(young(x) \wedge old(x)))$ — the chance that someone is not *both* old and young — which is clearly 1. Like the conditional probability, this represents certain knowledge, but it is still *not* the same expression (even if it pretty much captures the same knowledge).

Inference Rules

Hopefully, the above example is sufficient to motivate you to understand how to reason using conditional probabilities. The inference rule for conditional probability that is the counterpart of our familiar one for logic would be:

$$\frac{P(A) \quad P(B|A)}{P(B) = P(A)P(B|A)} \quad (8-1)$$

So, if we know that half of all people with colds have a cough and that, at any one time, 10% of people have a cold, then half of that 10%, or 5% of all people, will have a cough at any time: $P(\text{cough}) = P(\text{cold})P(\text{cough}|\text{cold})$.

Sometimes, instead of the conditional probability, we have information about the *joint probability* of an event: the chance of *both* A and B . If we think of A and B as sets, then $P(A)$ is the fraction of the universe that set A covers (the “size” of set A), $P(B)$ is the “size” of set B , and $P(A, B)$ is the fraction of the universe that $A \cap B$ covers. We can compute the conditional probability from a joint probability as:

$$P(B|A) = \frac{P(A, B)}{P(A)} \quad (8-2)$$

So, if 10% of all people have a cold at any time, and 5% of all people have a cough *and* a cold at any time, then the chance that someone with a cold will have a cough is 0.5, $P(\text{cough}|\text{cold}) = P(\text{cold, cough})P(\text{cold})$.

Reasoning From Evidence

Another way that we might want to use knowledge of conditional probabilities is inferring the cause of some sort of symptom. For example, given some problems that a car is exhibiting (hesitation, strange noises), we might want to know the probability of it needing an engine overhaul. This is a *diagnosis* problem, which is one of the major applications of expert systems. A similar situation might arise when you have knowledge of the actions of a person (the external *evidence*) and want to infer something about their internal motivation (one or more *hypotheses*). This would be the situation when an expert system is part of a web server, trying to infer what a person wants to buy from her clicks.

Well, we might have some statistical information that we’ve gathered from the population at large, concerning the probabilities of an hypothesis being true and some evidence being true, *without regard to any other information*. We might also know something about the causal nature of the situation — the chance that, if the hypothesis is true, the evidence will be observed. This information is often called *a priori* probabilities, because it is know *before* we start any problem solving (before we run the expert system to perform a diagnosis). With this *a priori* information, what we want to do is solve the inverse problem: given that we observe some evidence, what is the chance that the hypothesis is the underlying cause. This is *Bayes’ rule*:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \quad (8-3)$$

Example: Probabilistic Medical Diagnosis

Let's say we want to know the chance of someone having tuberculosis given that they're coughing. Our hypothesis H is "tuberculosis" and our evidence E is "coughing". We could get information about the incidence of H in the general population ($P(H)$) from, say, the National Institutes of Health, and for E ($P(E)$) from some similar source. We could also get the causal information: the fraction of people with tuberculosis who have a cough, $P(E|H)$. Let's put some numbers behind this *a priori* information:

- $P(\text{Tuberculosis}) = 0.05$
- $P(\text{Coughing}) = 0.2$
- $P(\text{Coughing}|\text{Tuberculosis}) = 0.8$

Using Bayes' rule, we can then conclude:

$$\begin{aligned} & P(\text{Tuberculosis}|\text{coughing}) \\ &= \frac{P(\text{Coughing}|\text{Tuberculosis})P(\text{Tuberculosis})}{P(\text{Coughing})} \\ &= \frac{0.8 \times 0.05}{0.2} \\ &= 0.2 \end{aligned}$$

In other words, if 5% of the population has tuberculosis and 20% are coughing at any particular time, and 80% of tuberculosis patients cough, then if we see someone coughing, there's a 20% chance they have tuberculosis. This may seem high, but of course these are bogus numbers. You can see from the probabilities above that 4% of the general population will be coughing tuberculosis patients (0.8×0.05), leaving 16% coughing from other causes. So, one-fifth of folks who cough have tuberculosis (4% out of 20%), which is our result.

By the way, this conclusion — that, if our patient is coughing, there's a 20% chance he has tuberculosis — is called an *a posteriori* probability. We call it that because we arrive at it *after* we have the particular case's information that the patient *is* coughing.

Multiple Causes

What if there are multiple possible causes of a particular piece of evidence? For example, what if a person can be coughing from either a cold or tuberculosis? This is no particular problem, we can use Bayes' rule above to compute the conditional probability for each hypothesis. Of course, we are no longer free to start with whatever bogus numbers we want for the *a priori* information (the *priors*) — it has to be consistent. Assuming that is the case, Bayes' rule will give us the chance of our patient having a cold and the chance that he has tuberculosis.

Sometimes, evidence isn't a symptom, but rather the result of some test. This should be familiar to you, as doctors (to continue our medical theme) often are entertaining multiple hypothesized causes for some ailment. How to decide among multiple possible tests? There are of course many factors, including cost, pain/discomfort, etc. One important factor is what we might call the *power*

of the test: how good the test will be at pinpointing the cause. Knowing nothing about medicine, we can still say something about the power of tests based solely on the probability of a particular test outcome if a particular ailment is the cause.

Let's say we're contemplating a test which, if the ailment is H_1 then there's a $P(E|H_1)$ chance that the result will be E . Similarly, if the ailment is H_2 , then there's a $P(E|H_2)$ probability that the test result will be E . How much will the test tell us? In other words, how well will the test discriminate between H_1 and H_2 ? From Bayes' rule, we can identify one bad situation: the case when $P(E|H_1) = P(E|H_2)$. In that case,

$$\begin{aligned}
 P(H_1|E) &= \frac{P(E|H_1)P(H_1)}{P(E)} \\
 P(H_2|E) &= \frac{P(E|H_2)P(H_2)}{P(E)} \\
 \frac{P(H_1|E)}{P(H_2|E)} &= \frac{P(E|H_1)P(H_1)}{P(E)} \frac{P(E)}{P(E|H_2)P(H_2)} \\
 &= \frac{P(H_1)}{P(H_2)} \tag{8-4}
 \end{aligned}$$

In words, if E gives us the same information about H_1 and H_2 , then knowing E will not help us at all in distinguishing between H_1 and H_2 .

Example: A test that provides no information

If $P(\text{Coughing}|\text{Tuberculosis}) = P(\text{Coughing}|\text{Cold})$, then knowing someone is coughing does not help in deciding whether the person has a cold or tuberculosis. The *a posteriori* probabilities we would assign to the two diseases would be the same as their *a priori* ones.

The Law of Total Probability

Lacking additional information (which we will come to when we talk about conditional independence), one way to figure out the probabilities for different possible “states of the world” is to compute these for each state. Let's say that there are n different possible states the world can be in, called s_1, \dots, s_n , such that $P(s_1 \vee \dots \vee s_n) = 1$ (so at least one of these must be true). For simplicity's sake, let's also assume that, for all $i \neq j$, $P(s_i \wedge s_j) = 0$. In other words, s_1, \dots, s_n form a *partition*, much in the same way that slicing a pizza is a partition of the pie — we can be in one state or another, but not both. For example, at a ‘T’ intersection, we can turn left or we can turn right, but we cannot do both, so (left, right) form a partition.

The following rule applies for the partition (s_1, \dots, s_n) :

$$\begin{array}{r}
 P(s_1) \\
 \vdots \\
 P(s_n) \\
 P(H|s_1) \\
 \vdots \\
 P(H|s_n)
 \end{array}$$

$$P(H) = \sum_{i=1}^n P(H|s_i)P(s_i) \tag{8-5}$$

So, if we know the priors for each possible way that something might happen, and it can only happen one way, then we can compute the probability of some outcome.

Example: Traffic accident

Let’s say there’s a particular ‘T’ intersection where we know that there’s a 1% chance of a driver going left having an accident and 2% of a driver turning right having one. If we assume an equal rate of drivers turning left and right at the intersection, then the law of total probability tells us that there is a 1.5% chance of a driver arriving at the intersection having an accident.

Example: Cheaters never prosper

Imagine you’re taking a proctored exam. One of the proctor’s jobs is to make sure nobody cheats. Let’s say there’s a 60% chance of getting caught cheating while the proctor is actively scanning the room looking for cheats. However, this particular proctor isn’t very conscientious — half the time, he’s reading a novel. When he’s reading, there’s only a 10% chance that a cheater will get caught. What’s the overall probability that a cheater will get caught? (*Popup answer:* $P(\text{proctor-reading}) = P(\text{proctor-not-reading}) = 0.5$, $P(\text{get-caught}|\text{proctor-not-reading}) = 0.6$, $P(\text{get-caught}|\text{proctor-reading}) = 0.1$. So,

$$\begin{aligned}
 P(\text{get-caught}) &= \\
 &P(\text{get-caught}|\text{proctor-reading})P(\text{proctor-reading}) \\
 &+ P(\text{get-caught}|\text{proctor-not-reading})P(\text{proctor-not-reading}) \\
 &= 0.1 \times 0.5 + 0.6 \times 0.5 = 0.35
 \end{aligned}$$

This general approach, in which we consider all possibilities and compute the overall chances, is called *case analysis*. As you might expect, case analysis is really only feasible when there are a small number of cases. Luckily, the structure of our world (as embodied in our rules) can allow us to eliminate some cases.

Independence

The key to getting away from case analysis is realizing that not all cases tell us something about our hypothesis. We say that H is *independent* of E if $P(H|E) = P(H)$ — knowing about E doesn't change our assessment of H . Clearly, this is a case we needn't consider.

A little trickier situation arises when there is a causal link between two forms of evidence. We'll get to the causal part in a moment; let's stick to the probabilistic part for now. We say that H is *independent* of E_1 *given* E_2 if $P(H|E_1 \wedge E_2) = P(H|E_2)$. This is *conditional independence*. This is the situation where, once we know about E_2 , we don't need to know about E_1 (because it won't tell us anything).

Example: Independence

A doctor would probably consider the chance of someone having tuberculosis to be independent of them exhibiting the symptom of having itchy skin. Because tuberculosis doesn't make you itch, $P(\text{tuberculosis}|\text{skin-itching}) = P(\text{tuberculosis})$.

Example: Conditional Independence

Imagine you're flying from Seattle to Philadelphia, with a stop in Chicago. Clearly, your arrival time in Philadelphia depends on your departure time from Seattle. However, once you depart from Chicago, arrival time in Philadelphia can be estimated solely from *that* departure time — there's no need to consider the Seattle departure time anymore. This is conditional independence: $P(\text{Philly-11PM}|\text{Seattle-1PM} \wedge \text{Chicago-8PM}) = P(\text{Philly-11PM}|\text{Chicago-8PM})$.

It is important to remember that these definitions of independence and conditional independence are *purely statistical* — they do *not* depend on any causal relationships among the events. The reason I say this is that we're most used to *causal independence*, such as the two examples above. We gain this knowledge from our understanding of how the world works; this is knowledge we could encode within an expert system. We shouldn't confuse statistical correlation — *statistical independence* — with causal independence. All causally independent events will be statistically independent, but the converse is not true.

Example: Causal independence vs. statistical independence

Suppose that a security guard watching an appliance store suffers from some mysterious disease that causes him to sleep between 8PM and 9PM. Suppose further that this is also the same time when "Friends" is on TV. No *causal link* exists between the show and the security guard's sleeping, but a *statistical link* exists: If a thief know that "Friends" is on, then he knows the guard is sleeping. The thief can . . .

Simplifying the Control Problem

Let's return to the situation where we are interested in a number of variables. What knowledge is needed so that we can reason among them? What would certainly be sufficient would be the joint probability distribution over *all* of the variables: the probability of all possible combination

of outcomes. If we have n variables that partition our situation of interest and each variable has two possible outcomes (they are binary variables), how many different numbers (probabilities) will we need? (*Popup answer: 2^n , the number of different combinations of n binary values.*)

Example: Joint probability distribution

Consider a situation in which there may or may not be a fire at your work. There is a fire alarm there, which might or might not be sounding. Someone might have been tampering with the alarm (or maybe not). When you arrive, you might or might not see people leaving the building, or smoke coming from the building. So, we have 5 binary variables: fire, tampering, alarm, leaving, smoke. We would need $2^5 = 32$ different probabilities:

$$\begin{aligned}
 &P(\text{fire, tampering, alarm, leaving, smoke}) \\
 &P(\text{fire, tampering, alarm, leaving, } \overline{\text{smoke}}) \\
 &\quad \vdots \\
 &P(\overline{\text{fire}}, \overline{\text{tampering}}, \overline{\text{alarm}}, \overline{\text{leaving}}, \overline{\text{smoke}})
 \end{aligned}$$

Here, I've used **variable** to indicate the variable is true and $\overline{\text{variable}}$ to indicate that it is false.

If we have this complete joint probability information, then can reason among the variables freely. But, what if we don't? Say we have instead:

$$\begin{aligned}
 &P(\text{alarm}|\text{fire}) \\
 &P(\text{alarm}|\text{tampering}) \\
 &P(\text{fire}) \\
 &P(\text{tampering}) \\
 &P(\text{smoke, leaving, alarm})
 \end{aligned}$$

Its difficult to tell if we can compute $P(\text{fire}|\text{leaving})$ from this information. Conditional independence can simplify the matter.

Making Use of Independence

As you've already seen, joint probabilities need a lot of numbers (information). Even our simple "burning building" example would need 2^5 probabilities. Clearly, this is an infeasible approach for large numbers of variables (and/or variables that can take on more than the two values of "true" and "false"). Luckily, conditional independence can simplify the necessary knowledge acquisition process.

In the real world, things cause other things to happen, which in turn cause other things to happen. Thus, there is a *chain of events* that naturally occurs. The pitcher throws the ball a bit high, you swing the bat and get under the ball, the ball is a pop fly behind you and to the left, and your neighbors window gets broken. Because things are causally related and because the cause

happens before the effect (we're talking everyday experience here, not quantum mechanics), we can sort events according to time and/or causal relationship.

Suppose we have n variables, which can be ordered x_1, x_2, \dots, x_n , so that no x_{i+j} is a cause for any x_i (any event listed before it). For each x_i , we can come up with a list of other variables which can be its *direct* causes. Let's call this list π_i . All of the variables in π_i come before x_i ; in other words, $\forall x_k \in \pi_i, k < i$. Other variables in $\{x_1, \dots, x_{i-1}\}$ may cause x_i indirectly, but x_i is conditionally independent of $\{x_1, \dots, x_{i-1}\} - \pi_i$ given π_i .

We can now obtain the probability of any combination of events from a relatively small set of conditional probabilities (relatively small, compared to the number of joint probabilities). To obtain some $P(x_1, \dots, x_n)$, it suffices to have $P(x_1), P(x_2|\pi_2), \dots, P(x_n|\pi_n)$. This is merely an application of a chain rule to conditional independence:

$$\begin{aligned} P(x_1, \dots, x_n) &= P(x_1)P(x_2|x_1) \dots P(x_n|x_1, \dots, x_{n-1}) \\ &= P(x_1)P(x_2|\pi_2) \dots P(x_n|\pi_n) \end{aligned} \tag{8-6}$$

Since π_i might be only a small portion of $\{x_1, \dots, x_i\}$, the amount of numbers needed is drastically reduced.

Example: Using independence to compute joint probability

Let's reason among our by-now familiar binary variables, alarm, people-leaving-building, fire, tampering, and smoke. As described above, we first need to order these. Well, alarms usually don't cause fires, nor do people leaving the building or smoke. A similar argument can be made for what causes tampering with an alarm. Assuming that the alarm responds to the fire itself and not to the smoke, we might get the ordering: fire, tampering, alarm, leaving, smoke.

Next, we need to figure out the interdependencies. A reasonable analysis might be:

$$\begin{aligned} \pi_{\text{fire}} &= \emptyset \\ \pi_{\text{tampering}} &= \emptyset \\ \pi_{\text{alarm}} &= \{\text{fire, tampering}\} \\ \pi_{\text{leaving}} &= \{\text{alarm}\} \\ \pi_{\text{smoke}} &= \{\text{fire}\} \end{aligned}$$

At this point, we know the probabilities we require:

1. $P(\text{fire})$
2. $P(\text{tampering})$
3. $P(\text{alarm}|\text{fire})$
4. $P(\text{alarm}|\text{tampering})$
5. $P(\text{leaving}|\text{alarm})$

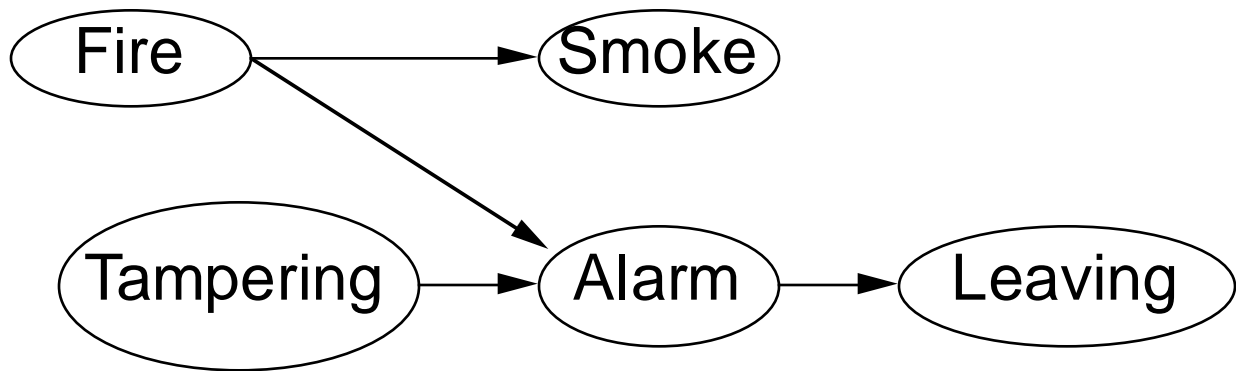


Figure 8.1: An influence diagram showing the causal relationship among different events.

6. $P(\text{smoke}|\text{fire})$

We can now compute the probability of, for example, there being an actual fire, with the alarm going off, people leaving the building, and smoke coming out:

$$\begin{aligned}
 &P(\text{fire, alarm, leaving, smoke}) \\
 &= P(\text{fire})P(\text{alarm}|\text{fire})P(\text{leaving}|\text{alarm})P(\text{smoke}|\text{fire})
 \end{aligned}$$

The power of this approach is manifold. Not only does it simplify knowledge acquisition, but also knowledge storage and reasoning. And, as we will see next, it implies a causal structure which has many features in common with our familiar rule bases.

Bayesian Networks (Influence Diagrams)

We can produce a graphical representation of the abovementioned causal relationships using the following simple algorithm:

1. Draw a node for each variable x_i . If you want to be neat about it, draw them so they're sorted topologically according to their order i .
2. Draw an arc from node x_j to node x_i if and only if $x_j \in \pi_i$.

The result of this algorithm for our running example is given in figure

Certainty Factors

Certainty factors were originally developed *ad hoc* for MYCIN. The idea was to minimize the amount of computation required to *update* the probability (actually, “certainty”) ascribed to facts as evidence is accumulated *incrementally*. Since this was developed for a rule-based system, the update computations were limited to the facts and their certainties on the rule LHS and RHS *before* the rule fires and certainty information attached to the rule itself (as in the uncertain rule formalization mentioned at the beginning of this lesson).

The approach taken was to attempt to model human expert *belief*, rather than probabilities. From a philosophical point of view, this is not as unreasonable or unscientific as it sounds. Fundamentally, probabilities describe the frequency of occurrence of reproducible events. What we are trying to do, however, is represent the amount of support for a particular fact, based on some other facts.

The definition of certainty factor of some hypothesis H due to evidence E is:

$$CF(H, E) = \frac{MB(H, E) - MD(H, E)}{1 - \min [MB(H, E), MD(H, E)]} \quad (8-7)$$

where MB is the measure of (increased) belief in H because of E and MD is the measure of (increased) disbelief in H due to E . The denominator here serves the purpose of, for example, preventing a small amount of evidence that increases MD from overwhelming a larger amount of evidence supporting a larger MB. Thus, our certainty is the *relative* difference between the two measures, rather than the *absolute* difference.

MB and MD are defined in terms of probabilities as:

$$MB(H, E) = \begin{cases} 1 & P(H) = 1 \\ \frac{\max[P(H|E), P(H)] - P(H)}{1 - P(H)} & P(H) < 1 \end{cases} \quad (8-8)$$

$$MD(H, E) = \begin{cases} 1 & P(H) = 0 \\ \frac{\min[P(H|E), P(H)] - P(H)}{-P(H)} & P(H) > 0 \end{cases} \quad (8-9)$$

So, MB and MD range between zero and one, while CF lies between -1 and 1. If a hypothesis is certainly true ($P(H|E) = 1$), what are their values? (*Popup answer:* MB = 1, MD = 0, and so CF = 1) If a hypothesis is certainly false ($P(\bar{H}|E) = 1$? (*Popup answer:* MB = 0, MD = 1, CF = -1) What if H is independent of E ? (*Popup answer:* If they're independent, then $P(H|E) = P(H)$, so MB = 0, MD = 0, and CF = 0)

To apply certainty factors to rules, we need to be able to combine CFs on rule LHS and compute resultant CFs for assertions on rule RHS. If n facts are ANDed together on the LHS, the composite certainty is

$$CF(F_1 \wedge F_2 \wedge \dots \wedge F_n) = \min [CF(F_1), CF(F_2), \dots, CF(F_n)] \quad (8-10)$$

Similarly, the certainty for a disjunction is

$$CF(F_1 \vee F_2 \vee \dots \vee F_n) = \max [CF(F_1), CF(F_2), \dots, CF(F_n)] \quad (8-11)$$

and for a negation is

$$CF(\neg F) = -CF(F_1) \quad (8-12)$$

Assuming we have computed the certainty factor $CF(L)$ for a rule LHS, and that the certainty factor associated with the rule itself is $CF(R)$, we can determine the certainty factor associated with a fact F asserted on the RHS as

$$CF(F) = CF(L) CF(R) \quad (8-13)$$

Note that the $CF(R)$ is a statement of what our certainty in the conclusion F should be if the LHS is known to be certainly true. So, it makes sense then that we should multiply this by the actual belief we have in the LHS.

The final issue we need to deal with is *updating* certainties. Let's say rule R_1 asserts a fact F with certainty factor $CF_1(F)$ and that another rule R_2 asserts the same fact with a certainty factor of $CF_2(F)$. Clearly, we want to have the ability to increase or decrease our certainty in F based on multiple sources of evidence. So, we need a way to combine certainty factors:

$$CF(F) = \begin{cases} CF_1(F) + CF_2(F) - CF_1(F) CF_2(F) & CF_1, CF_2 > 0 \\ \frac{CF_1(F) + CF_2(F)}{1 - \min[|CF_1(F)|, |CF_2(F)|]} & \text{only one } > 0 \\ CF_1(F) + CF_2(F) + CF_1(F) CF_2(F) & CF_1, CF_2 < 0 \end{cases} \quad (8-14)$$

For example, if $CF_1(F) = 0.5$ and $CF_2(F) = 0.5$, then the combined certainty factor would be $CF(F) = 0.75$. If another rule concludes F with a certainty factor of -0.5 , the new certainty factor would be what? (*Popup answer: $CF_{\text{new}} = (0.75 - 0.5)/(1 - 0.5) = 0.5$, which makes sense: the new information canceled out one of the previous pieces of evidence.*)

Fuzzy Logic

I'm going to dispense with making any distinction between fuzzy sets and fuzzy logic and focus instead on their basic concepts and simple uses. The logic and set theory that you are used to is binary in nature: an object is either in a set or not, a variable is either true or not. So, for example, you might say that a person is tall, meaning that they are a member of the set of tall people or that the property "tall" is "true" for them. This is similar to the act of assigning a value to some variable: you say that a person is 170cm tall.

You've already seen one way to deal with uncertainty: probability. Fuzzy logic presents an alternative way of expressing imprecision or uncertainty.

In set theoretic terms, we replace the binary member function with a real-valued one which ranges from zero (definitely not in the set) to one (definitely in the set). This is a intuitively more attractive approach. Ask yourself this: is a 170cm man tall? 180cm? 190cm? 200cm? We could define a cutoff value for binary set membership, and say that any man taller than 190cm is tall, but this probably wouldn't satisfy everyone. On the other hand, if we surveyed a number of people, we might find that few felt a 170cm man is tall, more for 180cm, still more for 190cm, and pretty much everyone for 200cm. This implies a gradually increasing *membership function*, $f(\text{height}) : \text{height} \rightarrow [0, 1]$, which maps height to the closed interval $[0, 1]$.

Similarly, instead of assigning binary truth values to variables, we can assign fuzzy truth values, ranging from zero to one. These indicate the *degree* of membership for that variable in the set of true things (rather than just being a member or not).

So, given the basic definition of fuzzy membership or fuzzy truth values, we "just" need to redefine set and logic operators to have "working" systems (of course, the trick is to do it right so that the systems we produce are consistent and reach conclusions that make sense).

Fuzzy Set Operations

We can define the fuzzy set operations for the fuzzy sets A and B with membership functions $f_A(x)$ and $f_B(x)$:

Equality $A = B$ if and only if (iff) $\forall x, f_A(x) = f_B(x)$.

Complement The complement A' of A is described by the set membership function $\forall x, f_{A'}(x) = 1 - f_A(x)$.

Subset $A \subseteq B$ iff $\forall x, f_A(x) \leq f_B(x)$. To be a proper subset, $A \subset B$ if the above is satisfied and $\exists x, f_A(x) < f_B(x)$.

Union If $C = A \cup B$ then $\forall x, f_C(x) = \max[f_A(x), f_B(x)]$.

Intersection If $D = A \cap B$ then $\forall x, f_D(x) = \min[f_A(x), f_B(x)]$.

Difference Set difference is expressed with the *bounded difference* operator, $|-|$. If $G = A|-|B$ then $\forall x, f_G(x) = \max[0, f_A(x) - f_B(x)]$.

Fuzzy Logic Operations

Similarly, we can define the following fuzzy logic operations for the fuzzy predicates $M(x)$ and $N(x)$ (which mean “ x is an M ” and “ x is an N ”, respectively):

Negation $\neg M(x) = 1 - M(x)$.

Conjunction $M(x) \wedge N(x) = \min[M(x), N(x)]$.

Disjunction $M(x) \vee N(x) = \max[M(x), N(x)]$.

How would you compute fuzzy implication? (*Popup answer: From the definition of implication, $M(x) \rightarrow N(x) = \max[1 - M(x), N(x)]$.)*)

Note that the conjunction and disjunction operations are consistent with those developed for combining MYCIN certainty factors on a rule LHS. However, there are two differences between these operations and those for crisp sets/logic:

1. Fuzzy logic allows us to entertain a contradiction, $M(x) \wedge \neg M(x) = \min[M(x), 1 - M(x)]$. So, if the truth value for an object \mathbf{O} is $M(\mathbf{O}) = 0.9$, then $M(\mathbf{O}) \wedge \neg M(\mathbf{O}) = 0.1$.
2. $M(x) \vee \neg M(x) = \max[M(x), 1 - M(x)]$ is not necessarily a tautology. So, if the truth value for an object \mathbf{O} is $M(\mathbf{O}) = 0.9$, then $M(\mathbf{O}) \vee \neg M(\mathbf{O}) = 0.9$.

There is a third-party contributed extension to JESS for fuzzy logic (actually, it is a set of Java classes which are compatible with JESS). See the JESS web site for the link.