

Expert Systems

Lecture 17

Probabilistic Reasoning

Probability is important in many branches of mathematics and science, and AI is not an exception. Many tasks require us to evaluate probabilities.

This lecture explains the basic probability framework. The next couple of lectures explains how to perform actual evaluations.

Reference:

- Textbook Chapter 14

AI(0270)

AI(0270)-17.1

Difficulty

Suppose we want to have a system to diagnose dental problems.

- We would like to have a rule like this:
 $Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$
- But we **can't**. There are **many other reasons to have toothache**, and just having toothache is not enough for deducing cavity.
- So we are forced to write something like...
 $Symptom(p, Toothache) \Rightarrow Disease(p, Cavity) \vee Disease(p, GumDisease) \vee Disease(p, ImpactedWisdom) \vee \dots$
- But... it is difficult, since there are really so many reasons.
- How about 'causal rules', e.g., $Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$?
- Still incorrect (although much better), and difficult to use.

AI(0270)-17.2

AI(0270)-17.3

Using probability to model uncertainty

- Probability solves the problem by using a number to **summarizes** all the uncertainty we meet.
- So we might say that there is a 80% chance that a patient has cavity if he has a toothache. The 20% summarize all the other possibilities.
- However, First-Order Logic lacks a way for us to express **how much efforts should be spent** on each type of events.
- In fact, it doesn't have provision for expressing that **some events are more probable than the other**.
- We want to allow the agent to **commit less** on the knowledge stored.
- Instead of saying that some statement is either true or false (or unknown), we want to be able to say that a statement is **true with certain probability**.
At the same time, we will go back to use propositions to simplify discussion.

AI(0270)-17.4

What the probabilities means?

Why we say that 80%, rather than 75% or 5%?

- By the 80%, we mean that **if similar things happen many times**, 80% of the time the patient has cavity.
- Usually, such knowledge are collected **by using a large database of past** experience.
- Or better, it might be **combined** from different sources of information.
- In cases that past experience is not available or is too small, we might need to ask a domain expert to **supply a subjective estimate**.
- Note that we still consider the statement to be **either true or false** with no intermediate. We **don't** want to express that a person **has cavity to the degree of 0.8**.
That is the subject of fuzzy logic, which is not discussed in this course.

AI(0270)-17.5

- Having forward and backward chaining systems is very convenient in many cases when we need to **make decisions automatically**.
- E.g., to deduce that a patient has a particular kind of disease, one can **have a rule and check whether the patient has each symptoms**.
- And this can be done recursively.
- Such systems are called **expert system**: a system that reasons using a **lot of domain knowledge**.
I.e., most of the intelligence is "built-in" rather than being computed. These knowledge are supplied by experts themselves.
- Target of these system: match or out-perform decisions of human experts in their specific domain.
- Motivations: cost, efficiency, accuracy, etc.

Non-monotonicity of probabilistic reasoning

Probabilistic reasoning differs from FOL reasoning because the assessment **changes in both directions** as more **evidence** is collected.

- Before the patient says anything, an agent might say that the probability of a patient having cavity is 10%.
- After the patient explains that he has toothache, we might have to **modify our assessment** and say that the probability of toothache is 80%.
- If the agent use its still probe to scan through the teeth of the patient, and found something which catch the probe, it might further modify the assessment to, e.g., 97%.

So all the statements about these numbers **must specify what evidence is known**—the numbers are different if different things are known.

And our agent must be specially designed to **update these numbers**.

AI(0270)-17.6

How to use the probability

Once we have the probabilities, what we can do with them?

- Say I want to the airport for a flight.
- I know that I will miss the flight with probability 5% if I leave home 3 hours before flight. If I leave 5 hours before flight, the probability goes down to 0.5%.
- Does it automatically mean that the second plan is a **better** plan?
- **We can't say**. That depends on my preference among the possibility.
- Just like a game with probabilistic elements, we **need to know a utility function** before we can use the probabilities.
- E.g., if I have something really important to do before the flight, which can be done much better if I've got 2 extra hours, I might prefer not to leave home so early.

AI(0270)-17.7

Notation 1: Prior probability

- The **prior probability** of a proposition A is denoted as $P(A)$.
- E.g., if the proposition *Cavity* means a particular patient has cavity, then we might write $P(A) = 0.5$ to denote that **if nothing else is known**, the patient has probability of 0.5 to have cavity.
- A "proposition" here can be a **propositional symbol** like *Cavity*, or a **equality relation** like $Weather = Cloudy$, or a **logic combination** of propositions like $Cavity \wedge \neg Weather = Cloudy$.
So our language is indeed between pure propositional logic and full FOL.
- We call variables like *Weather* to be **random variables**, and each such variables have a **domain** of values which it can take.
- A **probability distribution** enumerate the probability of a variable being **each** of the possible values. E.g., $\mathbf{P}(Weather) = \{Sunny : 0.3, Cloudy : 0.5, Raining : 0.2\}$.
Note the bold P that denotes distributions rather than probability.

AI(0270)-17.8

Axioms of probabilities

For any theory, there are something that is **assumed true** without proof. If you are talking about something that doesn't have these properties, then you are not talking about probabilities.

Yet they can be sensible. It just means that we don't consider them.

- All probabilities are between 0 and 1.
So for any proposition A , $0 \leq P(A) \leq 1$
- Probability of a valid statement and a contradiction are 1 and 0.
So $P(True) = 1, P(False) = 0$
- Probability of disjunction is given by
 $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
This is just the Venn diagram.

All other properties of probabilities derive from these axioms.

AI(0270)-17.9

Notation 2: Conditional probability

Once we know some evidence, we would like to know the **posterior** or **conditional** probability, i.e., probability given some evidence.

- Given the proposition B , the posterior probability of a proposition is denoted $P(B | A)$. E.g., $P(Cavity | Toothache) = 0.8$.
- Other forms of propositions are naturally allowed. E.g., we can say something like $P(X_1 = x | X_2 = y) = 0.75$.
- We also allow probability distributions, e.g., $\mathbf{P}(X | Y = y)$, which is a set of equations like $P(X = x_1 | Y = y) = p_1, P(X = x_2 | Y = y) = p_2$, etc.
- We will further probability distributions like $\mathbf{P}(X | Y)$. This is a set of equation telling us the probability of every possible value of X given every value of Y .

AI(0270)-17.10

What is it really?

- Posterior probability **can be expressed in terms of prior probability**:
 $P(B | A) = P(B \wedge A) / P(A)$.
- E.g., if $P(B) = 0.5, P(A \wedge B) = 0.4$, then $P(A | B) = 0.4 / 0.5 = 0.8$.
- If you think about it, it is really **very reasonable**:
 - If we have 100 scenarios, then $P(B) = 0.5$ means that 50 scenarios have proposition B being true.
 - And $P(A \wedge B) = 0.4$ means that 40 scenarios have proposition A and B both true.
 - If we **already know** that B is true, then we only have 50 scenarios, so the probability to find A true is $40 / 50 = 0.8$.
- But we will treat the equation as the **definition** of conditional probabilities (because it is inconvenient to define it in another way).

AI(0270)-17.11

Independence

- We say that a proposition A is **independent** of another proposition B if we have

$$P(A | B) = P(A)$$

- Intuitively, this means that **the probability of A does not change when we acquire the knowledge B .**

- Given the definition of $P(A | B)$, we can write it as

$$P(A \wedge B) = P(A)P(B)$$

- Note that the equation is **symmetric**. If A is independent of B , then B is independent of A .

- We usually have independence relations between variables. E.g., variable X and Y are independent if, for all i and j ,

$$P(X = x_i | Y = y_j) = P(X = x_i)$$

AI(0270)-17.12

Joint probability distributions

- So far, within the P notation, we can only have **one proposition or random variable**.

- We will further overload the notation to allow **multiple random variables** within the P notation.

- It is called a **joint probability distribution**, or simply a **joint**.
As opposed to probability distribution when there is only 1 variable.

- So we might write something like $P(X, Y)$, where X and Y are random variables.

- The meaning: a set of equations like

$$P(X = x_1 \wedge Y = y_1) = p_{11},$$

$$P(X = x_1 \wedge Y = y_2) = p_{12}, \text{ etc.}$$

AI(0270)-17.13

Example: a joint is all we need

- We might have the following joint for the dentist example:

	Toothache	\neg Toothache
Cavity	0.04	0.06
\neg Toothache	0.01	0.89

I.e., $P(\text{Toothache} \wedge \text{Cavity}) = 0.04$, etc.

- Note that we either have *Toothache* or \neg *Toothache*. We **can't have both, and can't have neither**. Similar for *Cavity*.

As a consequence, adding up all number results into 1.

- E.g., To find $P(\text{Cavity})$, we use the disjunction rule:

$$P(\text{Cavity}) = P(\text{Cavity} \wedge \text{Toothache}) + P(\text{Cavity} \wedge \neg \text{Toothache}) - P(\text{Cavity} \wedge \text{Toothache} \wedge \neg \text{Toothache}) = 0.04 + 0.06 - 0 = 0.1.$$

In fact, we would directly read them off the table.

- So $P(\neg \text{Toothache} | \text{Cavity}) = 0.04 / 0.1 = 0.4$.

AI(0270)-17.14

So a big joint is enough for anything we need...

- But unluckily, we have **too many variables!**

- Assume we have n boolean variables. Then we have 2^n numbers to specify!

So we can only store at most 30 random variables, even if they are boolean!

- And for many of these numbers, **we don't have a good way to estimate them!**

Because each probability would be so tiny that it is hard to estimate, and it requires a huge set of statistic to derive anything meaningful about them.

- A fully specified joint is **too big a requirement** for our reasoning system. We need to be able to **simplify** it, even if it might make our system less accurate.

- Our strategy: **work with the conditional probabilities directly!**

Rather than going to the full joints, we will use joints that are "local".

AI(0270)-17.15

Bayes rule

- The definition of conditional probability can be written this way:

$$P(A \wedge B) = P(A | B)P(B)$$

- Note that the equation is symmetric between A and B , so we can also write:

$$P(A \wedge B) = P(B | A)P(A)$$

- This is where all the fun begins. By equating them, we have

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)}$$

- This is called the **Bayes' rule**.

- Why it can be useful to know one probability from 3? Because **usually the 3 probabilities are easier to obtain** than the last one...

AI(0270)-17.16

Conditioning of equations

- There is also a version of Bayes rule which takes **extra condition**:

$$P(B | A \wedge H) = \frac{P(A | B \wedge H)P(B|H)}{P(A|H)}$$

- This is a **conditionalized** version of the Bayes' rule.

- But there is indeed no special reason to memorize this: it just add a condition H .

- We can **treat it as a background knowledge** that rules out all worlds that H is false.

- Basically **all** properties of probabilities can be conditioned this way. So instead of remembering a conditionalized Bayes' rule, remember that we can do conditioning.

Try some! Even the definition of conditional probability can be conditioned.

AI(0270)-17.17

Example: Combining evidences

- Suppose we already know that

$$P(\text{Cavity}) = 0.1$$

$$P(\text{Toothache}) = 0.05$$

$$P(\text{Toothache} | \text{Cavity}) = 0.4$$
- Note that the probability $P(\text{Toothache} | \text{Cavity})$ is a **causal knowledge**: it describes how the cause *Cavity* affects the probability of the consequence *Toothache*. It **tends to stay unchanged**.
- Then we know that

$$P(\text{Cavity} | \text{Toothache}) = 0.4 \times 0.1 / 0.05 = 0.8$$

$$P(\text{Cavity} | \neg \text{Toothache}) = (1 - 0.4) \times 0.1 / (1 - 0.05) = 0.0632$$
- On the other hand, what we computed is **diagnostic knowledge**, which is much more tenuous.

If $P(\text{Cavity})$ changes, due to changing habits, everything changes.

AI(0270)-17.18

Relative likelihood

- Suppose another cause *GumDisease* which also causes toothache.
- In diagnose, we usually **needs to know which is more probable**, *Cavity* or *GumDisease*, and by how much.

The exact probability is less important.
- Prior probabilities are given by $P(\text{Cavity})$ and $P(\text{GumDisease})$.
- Now if we have evidence that *Toothache* is true, then it is no longer sensible to use the prior probability. What are the **posterior probabilities**?

$$P(C | T) = P(T | C) P(C) / P(T)$$

$$P(G | T) = P(T | G) P(G) / P(T)$$
- So, we **don't need to know $P(T)$ to know which is larger!**
- The values $P(T | C) P(C)$ and $P(T | G) P(G)$ are called the **relative likelihood** of *C* and *G* given *T*.

And they are much easier to remember than Bayes' law!

AI(0270)-17.19

Normalization

- In fact, given the right information, we can **recover real probabilities from relative likelihood**.
- The idea is, we **also assess the relative likelihood of the negation**. They **must add up to one**, so their ratio **gives us a probability**.
- E.g., we know $P(\text{Toothache} | \text{Cavity}) P(\text{Cavity}) = 0.4 \times 0.1 = 0.04$.
- Suppose we know $P(\text{Toothache} | \neg \text{Cavity}) = 1/90$.
- Then the relative likelihood of $\neg \text{Cavity}$ given *Toothache* is

$$P(\text{Toothache} | \neg \text{Cavity}) P(\neg \text{Cavity}) = 0.9/90.$$
- So $P(\text{Cavity} | \text{Toothache}) = 0.04 / (0.04 + 0.9/90) = 0.8$.
- We call the **sum of the two likelihood** to be a **normalizing constant**, and the process **normalization**.

We write $P(A|B) = \alpha P(B|A)P(A)$, where α is the normalizing constant.

AI(0270)-17.20

Combining evidences

Suppose now we have one **more evidence**. The agent put a probe around the patient's teeth, and **something catches the probe** ("*Catch*").

- Again, we have assessment of **causal knowledge** $P(\text{Catch} | \text{Cavity})$ (e.g., 0.95) and $P(\text{Catch} | \neg \text{Cavity})$ (e.g., 1/180).
- So we want to know how much we should believe *Cavity* now?
- Again, since the **evidence changed**, we can't use the old value 0.8.
- We want $P(\text{Cavity} | \text{Toothache} \wedge \text{Catch})$. But how to get its value?
- The conditioned Bayes' rule (treating *T* background) tells us that

$$P(\text{Cav} | T \wedge \text{Catch}) = \alpha' P(\text{Cav} | T) P(\text{Catch} | T \wedge \text{Cav}).$$
- Bayes' rule tells us that

$$P(\text{Cav} | T) = \alpha P(\text{Cav}) P(T | \text{Cav})$$

AI(0270)-17.21

Our difficulty

So now we have this...

$$P(\text{Cav} | T \wedge \text{Catch}) = \alpha'' P(\text{Cav}) P(T | \text{Cav}) P(\text{Catch} | T \wedge \text{Cav}).$$

Note the **power of normalizing constants**... we don't need to evaluate them until the last step. Multiple normalizing constants combines to a "larger" constant.

This works as long as the values are really "constants", i.e., do not involve *Cav*. As long as this is the case, we won't get another "constant" when we change *Cav* to $\neg \text{Cav}$, so everything works out.

So how difficult is it to evaluate each term?

- We got $P(\text{Cavity})$ and thus $P(\neg \text{Cavity})$ directly.
- We also have $P(\text{Toothache} | \text{Cavity})$. $P(\text{Toothache} | \neg \text{Cavity})$ might be somewhat more difficult, but experts are still willing to guess it.

More difficult because it needs to summarize all other causes of toothache.
- But what about $P(\text{Catch} | \text{Toothache} \wedge \text{Cavity})$??

AI(0270)-17.22

Conditional independence

- It is very strange to estimate the value of that probability: we are asking for the probability of a consequence given the cause **and another consequence**.
- But **does it really matter that much?**
- If we believe that *Cavity* is the **direct cause** of both *Toothache* and *Catch*, then it really shouldn't.
- In particular, we **assume** that the only relation between *Toothache* and *Catch* is that they **both are caused by Cavity**.
- Once we know whether *Cavity* is true or false, they are **independent**:

$$P(T | \text{Cav} \wedge \text{Catch}) = P(T | \text{Cav})$$

$$P(\text{Catch} | \text{Cav} \wedge T) = P(\text{Catch} | \text{Cav})$$
- We call these **conditional independence assumptions**.

AI(0270)-17.23

How it helps?

- With the conditional independence assumption, our formula becomes
$$P(Cav|T \wedge Catch) = \alpha'' P(Cav)P(T|Cav)P(Catch|Cav).$$
- Now everything becomes a probability expressing **causal knowledge**, which is **easy to estimate**.
- So we have basically solved our “easiest” problem to combine evidences—given our conditional independence assumptions.
- Unluckily, the real world contains **much more variables...**
- and we have to make **much more independence assumptions**.
- It just becomes clumsy... how to express all of them? And how to **perform computations of posterior probabilities** in such complex situations?
- Luckily, there is really a way out...