Computer Science Department - Rutgers University **Fall 2017**

CS 520: Probabilistic Knowledge Bases and Queries

16:198:520
Instructor: Wes Cowan

On some level, *probability* is an assessment of belief in the presence of uncertainty. Events that we are certain of are given full belief, 100%, or $P(\text{Event}) = 1$, events that we are certain *against* are given no belief, 0%, or $P(\text{Event}) = 0$. Notice the symmetry this interpretation implies, that

$$P(\neg\text{Event}) = 1 - P(\text{Event}).$$

What are events? Events are assessments of possible outcomes - outcomes related to any objects of interest. We can talk about the event that the roll of a die comes up 4, or the compound event that the roll of a die comes up event (i.e., roll $= 2 \vee$ roll $= 4 \vee$ roll $= 6$). We frequently talk in terms of 'random' variables, variables that *have some value*, but we have some uncertainty about what that value is. We frequently will talk in terms of toy examples such as flipping coins or rolling dice, but we can use the same language to model many things of interest, for instance assessing belief about whether or not people will contract a disease, whether or not a stock price will go up or down (and to what value), how members of populations will vote, the belief that an opponent will or will not make certain moves in a game, or risks associated with various possible actions. The mechanics of probability direct how assessments of belief in some events should accurately and appropriately be used to assess belief in other events; a common example of this will be to ask how knowledge of certain or partial outcomes influences belief about other outcomes (how does the knowledge that the result of a roll was even effect the belief that the result of the roll was 4?). This frequently has the flavor of diagnosis, asking how knowledge of certain symptoms effects belief in underlying causes. In general, given your current state of knowledge, it is frequently of interest to determine how that knowledge effects your belief in other unknowns or future outcomes.

1 The Basics

Kolmogorov formalized the following axioms or rules for probability:

Definition 1 (Kolmogorov's Axioms)

1) *For any event E,*

$$P(E) \geq 0.$$

There is no such thing as ***negative belief.***

2) *For Ω as the event of all possible outcomes (of anything of interest),*

$$P(\Omega) = 1.$$

There is total belief in ***something****.*

3) *For disjoint or exclusive events, i.e., for $i \neq j$, if E_i happens then E_j cannot, and vice versa, we have*

$$P\left(E_1 \vee E_2 \vee E_3 \vee \ldots\right) = P(E_1) + P(E_2) + P(E_3) + \ldots.$$

Belief sums over ***exclusive outcomes.***

Note, the 3rd axiom gives an immediate answer to the question what is the probability that the roll is even: $P(\text{roll is even}) = P(\text{roll} = 2, 4, \text{ or } 6) = P(\text{roll} = 2) + P(\text{roll} = 4) + P(\text{roll} = 6) = 1/6 + 1/6 + 1/6 = 1/2$. Hence we see how this basic mechanics allow us to answer probabilistic queries in terms of more basic, 'atomic' probabilistic knowledge.

These are not necessarily the most useful properties in and of themselves, formal and basic to the extreme, but they do lead to the following important consequences:

Definition 2 (Kolmogorov's Consequences)

- *For $\emptyset$ as the 'empty' event, or set of 'no outcomes', we have $P(\emptyset) = 0$. To see this, note that $\Omega = \Omega \vee \emptyset$ is a disjoint partition of Ω.*
- *For any event E,*
$$0 \leq P(E) \leq 1.$$
To see this, note that we have $\Omega = E \vee \neg E$, hence
$$1 = P(\Omega) = P(E \vee \neg E) = P(E) + P(\neg E) \geq P(E) + 0 = P(E).$$
Related to this, we get the previous intuition, that $1 = P(E) + P(\neg E)$.
- *If $Event_A$ occurring implies that $Event_B$ must occur, then we have*
$$P(Event_A) \leq P(Event_B).$$
To see this, consider partitioning $Event_B$ into outcomes where A has happened, and outcomes where A has not happened.

1.1 Conditioning, Independence, and Conditional Independence

To answer the question of how certain knowledge influences belief, we introduce the idea of *conditional probability*. We may be unlikely to believe a certain event generally, but additional information may change what we are willing to believe. We may generally be unwilling to believe that we will get a new car at some point over the next year, but if we knew that we had or will win the lottery, that belief will change. Given the event B, the conditional probability of A is

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}, \tag{1}$$

i.e., the relative belief assigned to A and B out of all the belief assigned to B.

Conditional probability can be particularly useful when the events of B effect or influence what we believe about A in a known or understandable way. For instance, if we know that we are given a fair coin, the probability of getting heads may be 0.5 - but if we do not know that we are given a fair coin, it may be harder to assess what the probability of getting heads is: $P(\text{heads}|\text{fair}) = 0.5$ but $P(\text{heads}) =???$. (We will return to this example shortly.)

Conditioning is frequently of use when analyzing events that are related by cause and effect. Based on the definition of conditional probability, we get the fact that

$$P(A \wedge B) = P(B)P(A|B), \tag{2}$$

which can be interpreted as saying that the probability of A and B happening may be analyzed as first the probability of B happening, then *given that B has happened* the probability of A happening. Note, by symmetry, $P(A \wedge B) = P(A)P(B|A)$ as well. The classic example of this is that if a bag has two red balls and three black balls, the probability of pulling out two black balls is:

$$P(\text{black}_1 \wedge \text{black}_2) = P(\text{black}_1)P(\text{black}_2|\text{black}_1) = (3/5)(2/4) = 0.3. \tag{3}$$

Note the utility of this - allowing us to break a more complex probabilistic query down in terms factors more easily analyzed. This can also be applied in multiple steps:

$$P(A \wedge B \wedge C) = P(A)P(B|A)P(C|A \wedge B). \tag{4}$$

Notice the following important relation for conditional probabilities:

$$P(A|B) + P(\neg A|B) = \frac{P(A \wedge B) + P(\neg A \wedge B)}{P(B)} = \frac{P((A \wedge B) \vee (\neg A \wedge B))}{P(B)} = \frac{P(B)}{P(B)} = 1. \tag{5}$$

However, in general, no useful relation can be derived for $P(A|B) + P(A|\neg B)$.

It is not uncommon, however, that A and B might be events such that B tells us nothing about A. If B is that we are given a fair coin, that certainly influences A as getting heads, but if B is that Jupiter is aligned with Mars, it is much harder to argue that this has any influence on the outcome of A. In that case, we say that A and B are **independent**, or

$$P(A|B) = P(A), \tag{6}$$

that is that B tells us nothing that influences our belief in A. The prototypical example in this case is coin flips - the outcome of the first flip of a coin (generally) tells us nothing about the outcome of the second flip of that coin. (This will be revisited.)

If A is independent of B as above, we get a number of important relations. Note that $P(A \wedge B) = P(B)P(A|B) = P(B)P(A)$, but this additionally gives us that

$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(B)P(A)}{P(A)} = P(B), \tag{7}$$

hence we see that independence is a symmetric relationship - if A is independent of B, then B is independent of A, and vice versa. It is worth noting that this product formulation is frequently given as the definition of independence, if A and B are independent then $P(A \wedge B) = P(A)P(B)$. But this conditional probability definition is somewhat more natural, in my opinion.

We can extend this in a somewhat technical or abstract way to *conditional independence*. The idea here is that events A and B may not be independent themselves, but some other event C might give us information that allows us to separate A and B:

$$\begin{aligned} P(A \wedge B) &\neq P(A)P(B) \\ P(A \wedge B|C) &= P(A|C)P(B|C), \end{aligned} \tag{8}$$

i.e., given knowledge C, we may address A and B independently.

2 A Coin Based Example

Consider the following example. Someone shows you CoinA and CoinB; Coin A has a probability of giving heads of 0.5, i.e., is fair, and Coin B has a probability of giving heads of 0.7, i.e., is slightly biased in favor of heads. The person then gives you one of the coins, with know indication as to which one it is. What is your initial 'probabilistic knowledge base'? This can be summarized in two tables: First, because you have know reason to suspect you are given one coin over the other, you assess equal belief / probability between the two:

$$P(\text{Coin A}) = 0.5 \text{ and } P(\text{Coin B}) = 0.5. \tag{9}$$

We additionally have the following *Conditional Probability Table*, assessing the conditional beliefs about various outcomes:

$$\begin{aligned} P(H| \text{ Coin A}) = 0.5 \text{ and } P(T| \text{ Coin A}) = 0.5 \\ P(H| \text{ Coin B}) = 0.7 \text{ and } P(T| \text{ Coin B}) = 0.3. \end{aligned} \tag{10}$$

This captures everything we currently know or believe about the situation we are in. And we can begin to put queries to it. For instance, consider the following equation: will the first flip be heads? How much belief or probability are we willing to assign to this outcome?

To address this question, we utilize the first of our two primary tools: **marginalization**. This allows us to split an event into its component events, which may be easier to asses:

$$P(H) = P((H \wedge \text{ Coin A}) \vee (H \wedge \text{ Coin B})) = P(H \wedge \text{ Coin A}) + P(H \wedge \text{ Coin B}). \tag{11}$$

Note the use of Kolmogorov here, splitting the probability based on disjoint events. The general theme of marginalization is: *for a certain event to have happened, what else must have happened?*

We can then utilize the second primary tool, **conditioning**. Note that the event of getting heads *and* Coin A is somewhat difficult to assess, but getting heads *given* that we have Coin A is much easier.

$$P(H) = P(\text{Coin A})P(H| \text{ Coin A}) + P(\text{Coin B})P(H| \text{ Coin B}) = 0.5 * 0.5 + 0.5 * 0.7 = 0.6. \tag{12}$$

So we see that, given the initial state of our knowledge, the probability of the initial flip being heads is 0.6 - we have a slight bias in belief towards heads, because we know one of the coins is biased towards heads. But the mechanics of probability, in this case marginalization and conditioning, allow us to accurately assess based on our initial knowledge what our current belief should be.

More interesting is the following question: how does collected data inform our current belief? For instance, if we were to flip the coin and get a heads, how should we use this data to **update** our **prior** beliefs about whether or not we have Coin A or Coin B? Using the definition of conditional probability, we have

$$P(\text{Coin A}|H) = \frac{P(\text{Coin A} \wedge H)}{P(H)} = \frac{P(\text{Coin A})P(H| \text{ Coin A})}{P(H)}. \tag{13}$$

The above is in fact an application of **Bayes' Theorem**, allowing us to express our **posterior** or updated belief given data in terms of the **prior** belief, $P(\text{Coin A})$, the **likelihood** of getting that data in that case, $P(H|\text{Coin A})$, normalized by the total probability of getting that data. But as you can see, this is simply a repeated application of the notion of conditional probability. We can plug in what we know from our probabilistic knowledge base:

$$P(\text{Coin A}|H) = \frac{0.5 * 0.5}{0.6} \approx 0.4167. \tag{14}$$

Hence, given our initial data point of a heads, we should readjust our belief that we have Coin A down, from 50% to about 41.7%.

Here's an interesting question: Given that the first flip is a heads, what is the probability that the second flip is also a heads? If we were dealing with a single coin, flips are generally taken to be independent. However in this case, the first flip tells us something about how likely the coin is to be A or B (as evidenced in the above computation), and that in turn can tell us something about what the second flip is likely to be. In particular, we have

$$\begin{aligned} P(H_2|H_1) &= \frac{P(H_2 \wedge H_1)}{P(H_1)} \\ &= \frac{P(H_2 \wedge H_1 \wedge \text{Coin A}) + P(H_2 \wedge H_1 \wedge \text{Coin B})}{P(H_1)} \\ &= \frac{P(\text{Coin A})P(H_2 \wedge H_1|\text{Coin A}) + P(\text{Coin B})P(H_2 \wedge H_1|\text{Coin B})}{P(H_1)} \end{aligned} \tag{15}$$

At this point, we can argue that *given the identity of the coin*, the two flips are independent of each other. This yields

$$\begin{aligned} P(H_2|H_1) &= \frac{P(\text{Coin A})P(H_1|\text{Coin A})P(H_2|\text{Coin A}) + P(\text{Coin B})P(H_1|\text{Coin B})P(H_2|\text{Coin B})}{P(H_1)} \\ &= \frac{P(\text{Coin A})P(H|\text{Coin A})^2 + P(\text{Coin B})P(H|\text{Coin B})^2}{P(H_1)} \\ &= \frac{0.5(0.5)^2 + 0.5(0.7)^2}{0.6} \\ &\approx 0.61667. \end{aligned} \tag{16}$$

So we see that while the probability of the first flip being heads is 0.6, the probability of the second flip being heads *given that the first flip is heads* increases to approximately 0.617. This is because of the increased belief assigned to Coin B after the first heads - but it numerically verifies the intuition that the flips are *not* independent in this case, i.e., $P(H_1) \neq P(H_2|H_1)$.

We can also consider the effect of collecting even more data on our beliefs. Imagine that we flip the coin twice, and get two heads in a row - how should we update our belief that we have Coin A or Coin B? Marginalization and Conditioning. However, there are at least two or three ways we could approach this computationally. The first is a naive, direct application of the rules in the following way:

$$P(\text{Coin A}|HH) = \frac{P(\text{Coin A})P(HH|\text{Coin A})}{P(HH)} = \frac{P(\text{Coin A})P(H|\text{Coin A})^2}{P(H_1)P(H_2|H_1)} \approx \frac{0.5(0.5)^2}{0.6 * 0.61667} \approx 0.337836. \tag{17}$$

The second way is to consider not updating the prior from scratch, but update our *previously updated beliefs.* This yields the expression

$$\begin{aligned} P(\text{Coin A}|H_1H_2) &= \frac{P(\text{Coin A} \wedge H_1 \wedge H_2)}{P(H_1 \wedge H_2)} \\ &= \frac{P(H_1)P(\text{Coin A}|H_1)P(H_2|\text{Coin A} \wedge H_1)}{P(H_1)P(H_2|H_1)} \\ &= \frac{P(\text{Coin A}|H_1)P(H_2|\text{Coin A})}{P(H_2|H_1)}, \end{aligned} \tag{18}$$

where the last step has not only cancelled out $P(H_1)$ (i.e., the probability of data that you have already computed / dealt with does not matter), but also used the conditional independence property to give $P(H_2|\text{Coin A} \wedge H_1) = P(H_2|\text{Coin A})$. *(Why is this true?)* Note that this expression is in terms of the likelihood of the data given the hypothesis, $P(H_2|\text{Coin A})$, and **the previously updated priors** $P(\text{Coin A}|H_1)$. This formulation allows you to consider essentially a sequential updating of belief - at any time, your current beliefs should represent an accurate assessment of all prior belief, knowledge, and data, hence you can update from your current state of knowledge based on new data.

A Note On Implementation: The previous equation essentially expressed the new updated posterior in terms of the prior updated posterior (at any point, a past posterior may become the current prior). As such, $P(\text{Coin A}|H_1)$

was a value available based on prior updates and computations, and $P(H_2|\text{Coin A})$ was available in our knowledge base. But what of $P(H_2|H_1)$? For this simple example computation is not difficult, but it is easy to imagine things spiraling long out of control with more data.

So consider the following formulation: Let $F_1, F_2, \ldots, F_n$ be the observed sequence of flips so far, and hence we have computed priors

$$P(\text{Coin A}|F_1, F_2, \ldots, F_n) \text{ and } P(\text{Coin B}|F_1, F_2, \ldots, F_n). \tag{19}$$

A similar argument to the previous suggests the following update formula, given a new flip observation F_{n+1}:

$$\begin{aligned} P(\text{Coin A}|F_1, F_2, \ldots, F_n, F_{n+1}) &= \frac{P(\text{Coin A}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin A})}{P(F_{n+1}|F_1, F_2, \ldots, F_n)}, \\ P(\text{Coin B}|F_1, F_2, \ldots, F_n, F_{n+1}) &= \frac{P(\text{Coin B}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin B})}{P(F_{n+1}|F_1, F_2, \ldots, F_n)}. \end{aligned} \tag{20}$$

Again, much of these update formulas are available either due to the probabilistic knowledge base, or prior updates we have already done. The denominator is problematic, however.

But we may observe that the denominator in each case is the same. Imagine substituting in some unknown factor, α, so that

$$\begin{aligned} P(\text{Coin A}|F_1, F_2, \ldots, F_n, F_{n+1}) &= \alpha P(\text{Coin A}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin A}), \\ P(\text{Coin B}|F_1, F_2, \ldots, F_n, F_{n+1}) &= \alpha P(\text{Coin B}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin B}). \end{aligned} \tag{21}$$

This is only moderately useful, as we have replaced one unknown with another unknown. But notice that we have the additional relationship, that

$$P(\text{Coin A}|F_1, F_2, \ldots, F_n, F_{n+1}) + P(\text{Coin B}|F_1, F_2, \ldots, F_n, F_{n+1}) = 1. \textit{Why?} \tag{22}$$

We can utilize this in the following way, plugging in our expressions in terms of the unknown α and solving:

$$\alpha = \frac{1}{P(\text{Coin A}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin A}) + P(\text{Coin B}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin B})}. \tag{23}$$

This allows us to express our update equations completely in terms of knowns:

$$\begin{aligned} &P(\text{Coin A}|F_1, F_2, \ldots, F_n, F_{n+1}) \\ &= \frac{P(\text{Coin A}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin A})}{P(\text{Coin A}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin A}) + P(\text{Coin B}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin B})}, \\ &P(\text{Coin B}|F_1, F_2, \ldots, F_n, F_{n+1}) \\ &= \frac{P(\text{Coin B}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin B})}{P(\text{Coin A}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin A}) + P(\text{Coin B}|F_1, F_2, \ldots, F_n)P(F_{n+1}|\text{Coin B})}. \end{aligned} \tag{24}$$

This suggests the following general principle for updates: *For each possible outcome or hypothesis, scale the prior by the likelihood of the new data, then normalize by the sum of the scaled priors.* This is a general recipe for Bayesian Updating.