CSC 665: Artificial Intelligence

Lecture: PGMs II

1 Probability: inference

In a probability model, the inference question always takes the following form: given that we have observed evidence E, what is the probability of X? In other words, doing inference means computing $P(X \mid E)$, where each of X and E may denote either a single random variable or event, or a vector (tuple) of random variables or events.

Example: Returning to the setting of dental diagnosis, the inference question your dentist may be interested in is $P(\text{Cavity} = 1 \mid \text{Toothache} = 1)$, for binary random variables Cavity and Toothache. Suppose we have the following joint distribution over the three random variables Toothache, Cavity, and Catch. (Here, Catch is the random variable indicating whether the dentist's metal pokey tool catches on the tooth in question.)

Toothache	Cavity	Catch	$P(\text{Toothache}, \text{Cavity}, \text{Catch})$
0	0	0	0.576
0	0	1	0.144
0	1	0	0.008
0	1	1	0.072
1	0	0	0.064
1	0	1	0.016
1	1	0	0.012
1	1	1	0.108

Note that because there are three random variables, each of which can take on one of two possible values, specifying the joint distribution involves writing down $2^3 = 8$ probabilities.

Given the full joint distribution, we can compute any probability involving any subset of these three random variables. For example, suppose we want to compute $P(\text{Cavity} = 1)$. This requires us to sum up the probabilities in the rows where Cavity = 1, ignoring the values of the other two random variables:

$$\begin{aligned} P(\text{Cavity} = 1) &= 0.108 + 0.012 + 0.072 + 0.008 \\ &= 0.2. \end{aligned}$$

Similarly, the unconditional probability that you have a cavity or a toothache (or both) is

$$\begin{aligned} P(\text{Cavity} = 1 \vee \text{Toothache} = 1) &= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 \\ &= 0.28. \end{aligned}$$

These examples illustrate the operation of *marginalization*: using a joint probability distribution over several random variables to compute the probability of some subset of those variables. The general rule for computing the marginal probability of X given a joint distribution over X and Y is

$$P(X) = \sum_y P(X, Y = y).$$

That is, we let Y range over all possible values y, and add the up the corresponding joint probabilities.

In addition to marginal probabilities, we can also use the joint distribution to compute conditional probabilities. Using the definition from the previous lecture, you should verify that

$$P(\text{Cavity} = 1 \mid \text{Toothache} = 1) = 0.6.$$

We can use the definition again to show that

$$P(\text{Cavity} = 0 \mid \text{Toothache} = 1) = 0.4.$$

The fact that $0.6 + 0.4 = 1$ illustrates the general principle that a conditional probability distribution $P(X \mid E)$ is a valid probability distribution *over the random variable* X.

Together, the operations of marginalization and conditioning allow us to answer any probability query we like from the joint distribution. Given a joint distribution over random variables X, Y, E, the conditional probability that $X = x$ given $E = e$ is

$$\begin{aligned} P(X = x \mid E = e) &= \sum_y P(X = x, Y = y \mid E = e) \\ &= \frac{\sum_y P(X = x, Y = y, E = e)}{P(E = e)} \\ &= \frac{\sum_y P(X = x, Y = y, E = e)}{\sum_x \sum_y P(X = x, Y = y, E = e)}. \end{aligned}$$

Note that in the last equation we have expressed $P(X \mid E)$ purely in terms of joint probabilities of the form $P(X, Y, E)$, which by assumption we have access to (e.g. in the form of a table, as in the dentist example).

Given that the joint distribution lets us answer any possible inference question, it's worth asking how much work it is to fully specify a joint distribution.

Example: Returning to the dice rolling example, how many probabilities are needed to specify the joint distribution of the random variables Die_1 and Die_2. Well each of these random variables can take on six possible values, so it would seem that I need to specify 36 distinct probabilities. But consider the following probabilities:

$$\begin{gathered} P(\text{Die}_2 = 6 \mid \text{Die}_1 = 1) \\ P(\text{Die}_2 = 6 \mid \text{Die}_1 = 2) \\ \vdots \\ P(\text{Die}_2 = 6 \mid \text{Die}_1 = 6) \end{gathered}$$

How are these probabilites related? Unless the dice are communicating with each other, these probabilities must all be the same! In other words,

$$P(\text{Die}_2 = 6 \mid \text{Die}_1 = 1) = \cdots = P(\text{Die}_2 = 6 \mid \text{Die}_1 = 6) = P(\text{Die}_2 = 6).$$

The unconditional probability is equal to the conditional probability because in this case, the thing that we are conditioning on (i.e. Die_1 having a certain value) has no bearing on the value of Die_2.

How does this affect the amount of information we must store to specify the joint distribution? Recall that

$$P(\text{Die}_2 \mid \text{Die}_1) = \frac{P(\text{Die}_1, \text{Die}_2)}{P(\text{Die}_1)}.$$

But because $P(\text{Die}_2 \mid \text{Die}_1) = P(\text{Die}_2)$, we have

$$P(\text{Die}_2) = \frac{P(\text{Die}_1, \text{Die}_2)}{P(\text{Die}_1)},$$

or equivalently

$$P(\text{Die}_1, \text{Die}_2) = P(\text{Die}_1)P(\text{Die}_2).$$

Thus, the joint distribution factors into the product of two marginal distributions. Since each marginal distribution only requires 6 numbers to specify, the full joint distribution only requires that we store 12 numbers, not 36.[1] Indepedence is one of the most important concepts in all of probability theory.

Definition: Two random variables X and Y are said to be independent if any of the following equivalent equations hold:

$$P(X \mid Y) = P(X)$$
$$P(Y \mid X) = P(Y)$$
$$P(X, Y) = P(X)P(Y)$$

As the last of these equations shows, independence allows us to factor a joint distribution into a product of marginal distributions. From a computational perspective, this reduces the size of the representation. Given n random variables that each can take on k possible values, specifying the joint distribution in general requires $O(k^n)$ space. But if all n random variables are independent, the the space complexity is just $O(kn)$. In other words, independence reduces the size of the representation from exponential to linear in the number of variables — a huge reduction.

The final result we present is an immediate consequence of the definitions we've given already. But it's important enough to have its own name.

Theorem: (Bayes' rule) Given random variables X, Y, we have

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}.$$

The result can be proved by using the chain rule on $P(X, Y)$ twice, equating both possible expansions. From there, Bayes' rule follows from a simple algebraic rearrangement, as you should verify.

Bayes' rule is most useful when you have data in the "causal" direction (e.g. the probability of symptoms given a disease), and would like to make an inference in the "diagnostic" direction (e.g. the probability of a disease given symptoms).

Example: Let D be a binary random variable indicating whether you have some disease, and let T be a binary random variable indicating whether you test positive for that disease. Suppose the test is "95% reliable," i.e.

$$P(T = 1 \mid D = 1) = 0.95$$
$$P(T = 0 \mid D = 0) = 0.95.$$

Suppose further that 1% of the population has the disease. If you test positive, what is the probability you have the disease? This is an inference question in which we are conditioning on the evidence $T = 1$. The

[1] Technically this should read, "10 numbers, not 35," because given $n - 1$ probabilities for a random variable that can take on n possible values, the nth probability is constrained to be whatever value ensures that all n probabilities sum to 1.

relevant probability can be computed with Bayes' rule as follows:

$$\begin{aligned} P(D = 1 \mid T = 1) &= \frac{P(T = 1 \mid D = 1)P(D = 1)}{P(T = 1)} \\ &= \frac{P(T = 1 \mid D = 1)P(D = 1)}{P(T = 1 \mid D = 1)P(D = 1) + P(T = 1 \mid D = 0)P(D = 0)} \\ &= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} \\ &\approx 0.16. \end{aligned}$$