The final two lectures on probability will cover some basic methods for answering questions about probability spaces. We will apply them to the problem of choosing safer squares in Minesweeper.

Cartesian-product sample spaces and independence

Many sample spaces are constructed by taking the Cartesian product of the domains of a set of random variables $X = \{X_1, \ldots, X_n\}$. This means that each sample point (or atomic event) corresponds to a complete assignment of values to all the variables. Sample points, then, are very much like models in propositional logic.

JOINT DISTRIBUTION

Given a Cartesian-product sample space, the corresponding probability space is defined by a **joint distribution** on the variables, which assigns a probability to each possible complete assignment of values. For example, if the variables are $Heads_1$ and $Heads_2$, the Boolean outcomes of two independent unbiased coin tosses, then the joint distribution is

$$P(Heads_1 = true \cap Heads_2 = true) = 0.25 \quad P(Heads_1 = true \cap Heads_2 = false) = 0.25$$
$$P(Heads_1 = false \cap Heads_2 = true) = 0.25 \quad P(Heads_1 = false \cap Heads_2 = false) = 0.25$$

In dealing with problems containing several variables, we will often want to write equations manipulating entire joint distributions in one go, rather than having to deal with all the separate probabilities as above. We will therefore introduce some useful notation. We will write $P(X_i)$ to represent the distribution of a variable X_i; you can think of this as a vector. For example,

$$P(Heads_1) = \langle P(Heads_1 = true), P(Heads_1 = false)\rangle = \langle 0.5, 0.5\rangle$$

The joint distribution of $X_1, \ldots, X_n$ is written $P(X_1, \ldots, X_n)$. For example, the joint distribution spelled out above is denoted by $P(Heads_1, Heads_2)$. Using distribution notation like this allows us to simplfy many equations. For example, to say that $Heads_1$ and $Heads_2$ are independent, we can write

$$P(Heads_1, Heads_2) = P(Heads_1)P(Heads_2)$$

which should be viewed as a shorthand for the four equations produced by instantiating the variables in every possible way consistently across the equation.

$$P(Heads_1 = true \cap Heads_2 = true) = P(Heads_1 = true)P(Heads_2 = true)$$
$$P(Heads_1 = true \cap Heads_2 = false) = P(Heads_1 = true)P(Heads_2 = false)$$
$$P(Heads_1 = false \cap Heads_2 = true) = P(Heads_1 = false)P(Heads_2 = true)$$
$$P(Heads_1 = false \cap Heads_2 = false) = P(Heads_1 = false)P(Heads_2 = false)$$

Thus, you can think of the joint distribution $P(X_1, \ldots, X_n)$ as an n-dimensional table.

One further useful shorthand that is specific to Boolean variables: instead of $P(Heads_1 = true)$ and $P(Heads_1 = false)$, it is common to use the lower-case variable name with logical notation: $P(heads_1)$ and $P(\neg heads_1)$. Then, instead of taking intersections and unions of events (sets of sample points), we can use logical operators.

Thus, $P(Heads_1 = false \cap Heads_2 = false)$ becomes $P(\neg heads_1 \wedge \neg heads_2)$. A comma is often used in place of the $\wedge$ symbol.

Let's look at another example: n balls in m bins. We can define the sample space as a Cartesian product of the variables $Ball_1, \ldots, Ball_n$, each of which has possible values $\langle 1, \ldots, m \rangle$. (Note: we write the domain as an ordered tuple rather than a set because we want to be able to have elements of the distribution correspond to elements of the domain.) Thus, the joint distribution $P(Ball_1, \ldots, Ball_n)$ is an n-dimensional table with m^n entries. Obviously, for all but tiny values of m and n, we cannot hope to specify each entry individually. As with coin-tossing, we may assume independence:

$$P(Ball_1, Ball_2, \ldots, Ball_n) = P(Ball_1)P(Ball_2)\ldots P(Ball_n)$$

Lecture 17 assumed that each ball has an equal probability of going into any bin: for all i, $P(Ball_i) = \langle 1/m, \ldots, 1/m \rangle$. This assumption is *separate* from the assumption of independence. We could have a different tossing mechanism, such as a pin table, that gave a highly nonuniform distribution across the bins; we can even have a different tossing mechanism for each ball. As long as the balls don't interfere with each other, the independence equation will be valid.

Even with a different, nonuniform distribution for each ball, the product $P(Ball_1)P(Ball_2)\ldots P(Ball_n)$ requires only $O(mn)$ numbers to specify it, compared to $O(m^n)$ for the whole joint distribution without independence. Therefore, when it holds, independence can give an exponential reduction in the complexity of specifying (and, we shall see, reasoning with) a probability space. Unfortunately, independence is quite rare in the real world.

Consider the random variables $Weather$ (with values $\langle sunny, rainy, cloudy, snow \rangle$), $Toothache$, $Cavity$, and $Catch$ (all Boolean), where $Catch$ is true iff the dentist's nasty steel probe catches in my tooth. Together, these variables define a sample space with $4 \times 2 \times 2 \times 2 = 32$ entries. Can we use independence to simplify the joint distribution? Certainly, it's reasonable to say that the weather is indifferent to my dental problems, and vice versa. So we have

$$P(Weather, Toothache, Cavity, Catch) = P(Weather)P(Toothache, Cavity, Catch)$$

That is, the 32-element joint distribution can be *factored* into a 4-element distribution $P(Weather)$ and an 8-element distribution $P(Toothache, Cavity, Catch)$. (Given that the distributions must all sum to 1, we're really using $3+7$ numbers instead of 31 numbers in the full joint.)

So far, so good. But there are no independence relationships among the three dentistry variables. And if we want a real dentistry model, we may need hundreds of variables, all of which are dependent. That is, the dentistry part of the joint distribution may need $2^{hundreds}$ of numbers to specify it! We will see that we can do better.

Answering questions

A question, in the probabilistic context, requests the value of a conditional distribution, typically for a single variable X_i given evidence e (some specific values for evidence variables E, where $E \subseteq X - \{X_i\}$). That is, we want $P(X_i|e)$. For example, we might want to know $P(Cavity|toothache)$.

Method 1: summing joint distribution entries

Given a full joint distribution $P(X_1, \ldots, X_n)$, compute the probability $P(X_i|e)$ as follows. First, use the definition to rewrite the conditional probability as the ratio of two event probabilities; then, express the

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Table 1: A full joint distribution for the world consisting of the variables *Toothache*, *Cavity*, and *Catch*.

event probabilities as sums of probabilities of sample points—that is, probabilities of complete assignments to all the variables. Letting $Y = X - E - \{X_i\}$ (i.e., the variables other than the evidence and the query variable), we have

$$P(X_i|e) = \frac{P(X_i,e)}{P(e)} = \frac{\sum_y P(X_i,e,y)}{\sum_{y,x_i} P(x_i,e,y)}$$

where the summations are taken over all possible values of the corresponding sets of random variables. Notice that, from our definition of Y, the probabilities inside the summation are taken directly from the full joint distribution.

Table 1 shows a full joint distribution for the sample space defined by the random variables *Toothache*, *Cavity*, and *Catch*. Using the above formula, we can answer questions such as

$$\begin{aligned} P(Cavity|toothache) &= \frac{P(Cavity,toothache)}{P(toothache)} = \frac{\sum_{catch} P(Cavity,toothache,catch)}{\sum_{cavity,catch} P(cavity,toothache,catch)} \\ &= \frac{\langle 0.108+0.012, 0.016+0.064\rangle}{0.108+0.012+0.016+0.064} = \langle 0.6, 0.4\rangle \end{aligned}$$

That is, the probability of having a cavity given a toothache is 0.6.

Another way to look at the computation is to see that the *toothache* evidence simply restricts the computation to the left-hand side of Table 1. *Within* this restricted universe, which has just the two variables *Cavity* and *Catch*, we just need to compute $P(Cavity)$, which is $\langle 0.108+0.012, 0.016+0.064\rangle = \langle 0.12, 0.08\rangle$. Of course, this doesn't add up to 1, but we can scale it by $1/(0.12+0.08)$ so that it does. We obtain the same answer: $\langle 0.6, 0.4\rangle$. The scaling factor $1/(0.12+0.08)$ is exactly $1/P(toothache)$, i.e., $1/P(e)$. This works because $P(e)$ is a constant with respect to the query variable *Cavity*, i.e., $P(e)$ has the same value whether
NORMALIZATION
Cavity is true or false. It is common, therefore, to write the term $1/P(e)$ as a scaling or **normalization** constant α, which is chosen to make the final answer sum to 1:

$$P(X_i|e) = \alpha P(X_i,e) = \alpha\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

This way of doing things greatly simplifies the equations that we write out.

Method 2: Bayes' rule

Theorem 24.1: *For any random variables A, B,*

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

Proof: This is a trivial application (both ways) of the chain rule:

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

followed by division by $P(B)$. $\square$

Why is such a trivial theorem so important? Its main use is to reverse the direction of conditioning. Typically, one wants to compute $P(Cause|Effect)$, but one's knowledge of the domain is more commonly available in the causal direction: $P(Effect|Cause)$. Bayes' rule provides the needed transformation.

BAYESIAN UPDATING PRIOR LIKELIHOOD POSTERIOR

The form $P(A|B) = \alpha P(B|A)P(A)$ also illustrates the idea of **Bayesian updating**. $P(A)$ is the **prior** distribution on A. To accommodate evidence about B, this is multiplied by the **likelihood** of that evidence based on A, i.e., $P(B|A)$. After normalization, we obtain the **posterior** probability $P(A|B)$. The process can then be repeated. (But see below for what this repetition really means.)

Consider the disease-testing problem in Lecture 18. The information defining the problem is as follows:

$$P(Testpositive|disease) = \langle 0.9, 0.1\rangle$$
$$P(Testpositive|\neg disease) = \langle 0.2, 0.8\rangle$$
$$P(Disease) = \langle 0.05, 0.95\rangle$$

Given that one has tested positive, the probability of disease can be computed using Bayes' rule as follows:

$$\begin{aligned} P(Disease|testpositive) &= \alpha P(testpositive|Disease)P(Disease) \\ &= \alpha\langle 0.9, 0.2\rangle\langle 0.05, 0.95\rangle \\ &= \alpha\langle 0.045, 0.19\rangle \approx \langle 0.19, 0.81\rangle \end{aligned}$$

Conditional independence

Suppose one has a toothache and the dentist's nasty steel probe catches in one's tooth. Using Bayes' rule, we have

$$P(Cavity|toothache, catch) = \alpha P(toothache, catch|Cavity)P(Cavity)$$

To complete this calculation, we need to know $P(Cavity)$—simple enough, in effect just one number—and $P(toothache, catch|Cavity)$. The latter term is one representative of a table that is $2\times 2\times 2$, i.e., just as big as the joint distribution itself. And as the set of possible symptoms gets larger, this becomes exponentially larger and hence impractical.

One is tempted (justifiably, it turns out) to write

$$P(toothache, catch|Cavity) = P(toothache|Cavity)P(catch|Cavity)$$

and hence obtain

$$P(Cavity|toothache, catch) = \alpha P(toothache|Cavity)P(catch|Cavity)P(Cavity)$$

CONDITIONALLY INDEPENDENT

This factorization assumes that *Toothache* and *Catch* are **conditionally independent** given *Cavity*.

Definition 24.1 (Conditional Independence): Variables X and Y are conditionally independent given variable Z iff

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

Equivalent definitions are:

$$P(X|Y,Z) = P(X|Z) \quad \text{and} \quad P(Y|X,Z) = P(Y|Z)$$

This assumption is reasonable *because of our knowledge of the variables in question and their relationships*. (In the same way, we model n coin tosses as absolutely independent because we believe the coins are tossed so as not to interfere with each other.) *Given knowledge of the existence (or not) of the cavity*, the probability that the dentist's probe catches in one's tooth is independent of whether or not I happen to have a toothache. Note that this independence holds only given information about the cavity; if we don't know about the cavity, then $Toothache$ and $Catch$ are clearly dependent, since each suggests a cavity that, in turn, makes the other more likely.

Conditional independence occurs very frequently in the real world, because causal processes are "local." This usually allows the joint distribution on a set of variables to be factored into a set of small conditional distributions. We can see this by applying the chain rule:

$$\begin{aligned}
&P(Toothache, Catch, Cavity)\\
&\quad= P(Toothache|Catch, Cavity)P(Catch|Cavity)P(Cavity) \quad \text{(Chain Rule)}\\
&\quad= P(Toothache|Cavity)P(Catch|Cavity)P(Cavity) \quad \text{(conditional independence)}
\end{aligned}$$

If we count the numbers carefully, taking into account all the sum-to-1 constraints, we see that the joint (7 numbers) is reduced to $2+2+1=5$ numbers through the conditional independence assumption. This may not sound like much; but in realistic applications, we often see reductions from 2^{1000} numbers down to a few thousand.

Application: Bayesian updating with Gaussian distributions

This section looks at what happens if we have a Gaussian prior distribution describing our uncertainty about some quantity X, and then we obtain several measurements $y_1, \ldots y_n$ of that quantity, each of which has a Gaussian error associated with it. (For example, we might measure the temperature using 5 different thermometers, or we might count votes 5 times using different teams of counters.) We assume that the measurements are *conditionally* independent given the true value of X; this is entirely reasonable if the measurement processes are independent. Note that the measurements are not *absolutely* independent—indeed, they are highly correlated because they are all measurements of the same thing.

The prior distribution associated with X is a Gaussian with mean μ_0 and variance σ_0^2:

$$f(x) = \alpha \exp(-(x-\mu_0)^2/2\sigma_0^2)$$

If each measurement has Gaussian error, then *if the true value of* X *is* x, the distribution for Y_i is a Gaussian centered on x:

$$f(y_i|x) = \alpha \exp(-(y_i - x)^2/2\sigma_e^2)$$

where σ_y^2 is the variance of the error distribution.

Now we are interested in computing the posterior density of X. We want the true value given the measurements, but our model of the measurement process describes the probability of a measurement given the true value. So we need to apply Bayes' rule:

$$\begin{aligned}
f(x|y_1, \ldots y_n) &= \alpha' f(y_1, \ldots y_n|x) f(x) \quad \text{(Bayes' rule)}\\
&= \alpha' \prod_{i=1}^{n} f(y_i|x) f(x) \quad \text{(conditional independence)}
\end{aligned}$$

Thus, we start with the prior distribution and multiply it by successive distributions for each measurement.

Let us look at the first step of this process. We have

$$
\begin{aligned}
f(x|y_1) &= \alpha' f(y_1|x) f(x) \\
&= \alpha'' \exp(-(y_1-x)^2/2\sigma_y^2)\exp(-(x-\mu_0)^2/2\sigma_0^2) \\
&= \alpha'' \exp - \left(\frac{\sigma_0^2(y_1-x)^2 + \sigma_y^2(x-\mu_0)^2}{2\sigma_y^2\sigma_0^2} \right) \\
&= \alpha'' \exp - \left(\frac{x^2 - 2\frac{\sigma_0^2 y_1 + \sigma_y^2\mu_0}{\sigma_y^2+\sigma_0^2} x + \frac{\sigma_0^2 y_1^2 + \sigma_y^2\mu_0^2}{\sigma_y^2+\sigma_0^2}}{2\sigma_y^2\sigma_0^2/(\sigma_y^2+\sigma_0^2)} \right) \\
&= \alpha''' \exp - \left(\frac{(x - \frac{\sigma_0^2 y_1 + \sigma_y^2\mu_0}{\sigma_y^2+\sigma_0^2})^2}{2\sigma_y^2\sigma_0^2/(\sigma_y^2+\sigma_0^2)} \right)
\end{aligned}
$$

That is, after Bayesian updating, the posterior density $f(x|y_1)$ is a Gaussian with mean and variance given by

$$
\begin{aligned}
\mu_1 &= (\sigma_0^2 y_1 + \sigma_y^2\mu_0)/(\sigma_y^2+\sigma_0^2) \\
\sigma_1^2 &= \sigma_y^2\sigma_0^2/(\sigma_y^2+\sigma_0^2)
\end{aligned}
$$

The maths looks complicated but actually what's going on is very simple. We are adding two exponents, each of which is a quadratic in x, so we get a quadratic in x. We "complete the square", writing ax^2+bx+c as $a(x-b/2a)^2$ plus some constants which disappear into the new normalizing constant α'''. The new variance is extracted from the coefficient of x^2 and the new mean from the coefficient of x. In short, the product of two Gaussians is a Gaussian.

The process of updating the mean and variance is easier to understand if we write it using the *precision* τ, which is the inverse of the variance. The update equations become

$$
\begin{aligned}
\mu_1 &= (\tau_y y_1 + \tau_0\mu_0)/(\tau_y+\tau_0) \\
\tau_1 &= \tau_y + \tau_0
\end{aligned}
$$

That is, the new mean is a precision-weighted average of the old mean and the new measurement; and the new precision is the sum of the old precision and the measurement precision.

These results are quite intuitive. The update for the mean reflects the fact that the new mean always lies between the old mean and the measured value, but will move closer to the measured value as the measurement becomes more accurate. Conversely, if we are already very confident in the value (high precision) then the new measurement makes little difference. The precision always increases, representing the fact that new information should increase certainty.

After n measurements, an easy induction proof shows us that

$$
\begin{aligned}
\mu_n &= (\tau_y \sum_{i=1}^{n} y_i + \tau_0\mu_0)/(n\tau_y+\tau_0) \\
\tau_n &= n\tau_y + \tau_0
\end{aligned}
$$

Thus, as $n \to \infty$, μ_n tends to the mean value of the observations and σ_n tends to $\sigma_y/\sqrt{n}$. For the special case of samples drawn from a Gaussian distribution, then, we obtain exactly the result suggested by the Central Limit Theorem.