Probability

CS188 - Artificial Intelligence, Spring 2014

April 1, 2014

1. Introduction

Randomness is all around you. Our reality is filled with uncertainty – the time it takes you to get to school, the number of hours you sleep, the chances you'll win the next lottery, etc. If everything in life occurred deterministically, life would be pretty boring, wouldn't it?

In the first half of the course, we've seen expectimax search and Markov Decision Processes, two techniques for optimal planning that deal with probabilities and uncertainty. Now it's time to understand how to generate these probabilities for us to use them meaningfully. In particular, we will be building up the tools needed to perform probabilistic inference: answering probabilistic queries about some factors in our model given some evidence. But, before we do that, let's go over a quick recap of probability.

2. Random Variables and Probability

A **random variable** is formally defined by: a function that maps each element in a sample space (Ω) to a real value, which is the probability of that element occuring. For the purposes of this class, you can think of a random variable as just something in the real world that can take on multiple values, and assigning every value a probability. In this class, we'll be focusing on discrete random variables, where the sample space takes on a finite number of values. Anytime we refer to a random variable, we'll assume it's discrete.

As an example, one random variable could be $Weather(W)$, and the possible values are $\{sun, rain\}$, where sun occurs 70% of the time and $rain$ occurs 30% of the time.

We can define a random variable as a table. For the $Weather(W)$ example, the corresponding table is:

W	$Pr(W)$
sun	0.7
$rain$	0.3

The probability entries in the table define the **distribution** of the random variable. A valid distribution for a random variable X must satisfy the following two criteria:

1) All entries must be non-negative; $\forall x, Pr(X = x) \geq 0$
2) The sum of all entries must be 1; $\sum_x Pr(X = x) = 1$

Even with a random variable with two possible values, there are infinitely many valid distributions. One thing to point about regarding notation – it is conventional to define random variables with a capital letter, since they are functions. Lower case variables refer to an particular outcome of the random variable. So, when we write $Pr(X = x)$, we are querying for the probability that the random variable X takes on the outcome x. A shorthand notation for $Pr(X = x)$ is $Pr(x)$, and can be used when it unambiguous which random variable x is from.

3. Joint Distributions

Sometimes we are interested in relationships between random variables. A **joint distribution** is a distribution that maps each possible joint outcome combination from a set of two or more random variables to the probability of that particular combination occurring. So, adding to the $Weather(W)$ example, we might be concerned about the random variable $StayHome(S)$ as well. The joint distribution may look something like:

W	S	$Pr(W, S)$
sun	yes	0.2
sun	no	0.5
$rain$	yes	0.2
$rain$	no	0.1

From this joint distribution, we can read off that the probability that both the events $W = sun$ and $S = yes$ occur is 0.2.

4. Marginalization

Now that we've seen joint distributions that involve a set of two or more random variables, we would also like to recover the distribution of a smaller subset of these random variables from the joint. Let's go back to our $Weather(W) - StayHome(S)$ table above. Suppose we're interested in $Pr(S = yes)$. To calculate this value, we look for the rows consistent with $S = yes$ in the joint distribution, which would be the first and third rows. Then, since the events $(W = sun \cap S = yes)$ and $(W = rain \cap S = yes)$ are disjoint (meaning that they will never occur at the same time), we can add the probabilities together to recover $Pr(S = yes)$.

$$Pr(S = yes) = Pr(W = sun, S = yes) + Pr(W = rain, S = yes) = 0.2 + 0.2 = 0.4$$

Likewise:

$$Pr(S = no) = Pr(W = sun, S = no) + Pr(W = rain, S = no) = 0.5 + 0.1 = 0.6$$

Notice that to recover the probability of $S = yes$, we had to take a sum over the entries in the joint distribution that were consistent with $S = yes$, which also amounts to summing over all possible values the unwanted variable(s) – in this case, W – while fixing $S = yes$. A similar argument holds for $S = no$. Because of this, we could have written the calculation for $Pr(S = yes)$ and $Pr(S = no)$ as:

$$Pr(S = yes) = \sum_w Pr(W = w, S = yes)$$

$$Pr(S = no) = \sum_w Pr(W = w, S = no)$$

We can compactly combine the above two equations as:

$$\forall s; Pr(S = s) = \sum_w Pr(W = w, S = s)$$

Or even more compactly:

$$Pr(S) = \sum_w Pr(W = w, S)$$

This technique is what we call **marginalization**, which involves starting from a joint distribution of several random variables, querying for the distribution of a smaller subset of these random variables, and summing out over the possible values of the unwanted random variable(s).

5. Conditional Probability and the Normalization Trick

In the previous example, we've calculated the probability that you would stay home. But what if it was currently raining? Then what is the probability that you would stay home?

Conditional probabilities help answer questions like the one above. A **conditional probability** takes the form $Pr(X = x|Y = y)$, which stands for the probability that X takes on the value x given we know that Y has the value y. In the expression above, notice that the random variable we are querying for (X) comes to the left of the conditioning bar, whereas the evidence (Y) comes to the right of the conditioning bar. This holds true as we extend conditional probabilities to more than two random variables.

To calculate the conditional probability $Pr(X = x|Y = y)$, we use the following definition:

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)}$$

This should make sense intuitively. We already know that $Y = y$, and out of the probability that $Pr(Y = y)$, we want to find how much $X = x$ is covered, which is just $Pr(X = x, Y = y)$.

Using the tools of joint distributions and marginalization discussed before, we can work out the following conditional probabilities $Pr(S|W)$ for the $Weather(W) - StayHome(S)$ example:

W	S	$Pr(S\|W)$
sun	yes	2/7
sun	no	5/7
$rain$	yes	2/3
$rain$	no	1/3

An example calculation: $Pr(S = yes|W = sun) = \frac{Pr(S=yes,W=sun)}{Pr(W=sun)} = \frac{0.2}{0.7} = \frac{2}{7}$.

Notice that the table $Pr(S|W)$ is not a distribution, since the entries do not sum to 1. However, the table holds two distributions: $Pr(S|W = sun)$ and $Pr(S|W = rain)$. There is a shortcut to calculating these distributions. Take $Pr(S|W = sun)$. We know that for a particular outcome $S = s$, $Pr(S = s|W = sun) = \frac{Pr(S=s,W=sun)}{Pr(W=sun)}$. For every s, notice that the denominator $Pr(W = sun)$ is common. Hence, we can treat the denominator as some normalization constant that will bring the sum of our entries to 1. This means, all we need to worry about is finding the numerators, and then performing a **normalization** step to ensure the probabilities sum to 1.

Using the example above to find $Pr(S|W = sun)$, we note that $Pr(S = yes, W = sun) = 0.2$ and $Pr(S = no, W = sun) = 0.5$. The sum of these two entries is 0.7, so by normalization and dividing both entries by 0.7, we get that $Pr(S = yes|W = sun) = \frac{2}{7}$ and $Pr(S = no|W = sun) = \frac{5}{7}$.

6. Product Rule

Rewriting the definition conditional probability shown above, we see that

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)}$$

becomes

$$Pr(X = x, Y = y) = Pr(X = x|Y = y)Pr(Y = y)$$

The above statement is the **product rule**. Note that we could have switched the ordering of X and Y:

$$Pr(X = x, Y = y) = Pr(Y = y|X = x)Pr(X = x)$$

7. Chain Rule

We can modify the product rule for more than two random variables. In the product rule equation, let $Y = y$ be replaced by $Y = y, Z = z$. Then we have:

$$Pr(X = x, Y = y, Z = z) = Pr(X = x|Y = y, Z = z)Pr(Y = y, Z = z)$$

Using the product rule on the last term, the equation simplifies to:

$$Pr(X = x, Y = y, Z = z) = Pr(X = x|Y = y, Z = z)Pr(Y = y|Z = z)Pr(Z = z)$$

If we have n random variables $X_1, X_2, ..., X_n$, we can write the joint distribution as:

$$Pr(x_1, x_2, ...x_n) = Pr(x_1)Pr(x_2|x_1)...Pr(x_n|x_1, ..., x_{n-1}) = \prod_{i=1}^{n} Pr(x_i|x_1, ..., x_{i-1})$$

The above is called the **chain rule**. For n random variables, there are $n!$ orderings of the chain rule.

8. Bayes' Rule

Going back to the product rule, we say two ways of writing $Pr(X = x, Y = y)$:

$$Pr(X = x, Y = y) = Pr(X = x|Y = y)Pr(Y = y) = Pr(Y = y|X = x)Pr(X = x)$$

Rewriting this equation to solve for $Pr(X = x|Y = y)$, we get:

$$Pr(X = x|Y = y) = \frac{Pr(Y = y|X = x)Pr(X = x)}{Pr(Y = y)}$$

This is **Bayes' rule**. Note the importance of this rule. On the right side, we have a $Pr(X = x)$ term in the numerator. We will refer to this as the **prior distribution** of X, prior in the sense that we have not observed the evidence $Y = y$ yet. On the left side, we have $Pr(X = x|Y = y)$, which is the **posterior distribution** of X, posterior in the sense that we have already observed $Y = y$.

9. Independence and Conditional Independence

Another way to characterize relationships between random variables is to see if they are independent or not. For two random variables X and Y to be **independent random variables**, it must hold that whether or not you observed the value of X does not change the probability distribution of Y. Symbolically put, that means that $\forall x, y; Pr(Y = y) = Pr(Y = y|X = x)$. Likewise, $\forall x, y; Pr(X = x) = Pr(X = x|Y = y)$.

Going back to the product rule:

$$Pr(X = x, Y = y) = Pr(X = x)Pr(Y = y|X = x)$$

If X and Y are independent, we know that $Pr(Y = y) = Pr(Y = y|X = x)$, so by substitution, the product rule becomes:

$$Pr(X = x, Y = y) = Pr(X = x)Pr(Y = y)$$

Conditional independence is very similar to independence; we say that two random variables X and Y are conditionally independent given a third random variable Z if, already knowing the value of Z, whether or not you have observed the value of Y does not change the distribution on $X|Z$. Symbolically put, that means that if $X \perp\!\!\!\perp Y|Z$, $\forall x, y, z; Pr(X = x|Y = y, Z = z) = Pr(X = x|Z = z)$.

Similarly, the product rule simplifies to: $Pr(X = x, Y = y|Z = z) = Pr(X = x|Z = z)Pr(Y = y|Z = z)$.