Lecture 9 - Probability

Random variable: some aspect of the world about which we may have uncertainty. Can be assigned values from a domain. We can also associate a probability value to each value of the domain.

$$\forall x, P(X=x) \geq 0$$
$$\sum_x P(X=x) = 1$$

These are just the basic rules of probability. A **joint distribution** over a set of random variables $X_1, X_2, \ldots, X_n$ assigns a probability to each outcome.

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(x_1, x_2, \ldots, x_n) = p$$

Similar to above, $P(x_1, x_2, \ldots, x_n) \geq 0$ and $\sum_{x_1, x_2, \ldots, x_n} P(x_1, x_2, \ldots, x_n) = 1$.

What is the size of a distribution if we have n variables with domain size d? It would be d^n.

Marginal Probability: This is the probability of a single event occurring without any consideration of other events. It is derived by summing or integrating over teh possible values of the other random variables. This is the law of total probability:

$$P(X=x) = \sum_j P(x, y_j)$$

Conditional Probability: What is the probability of A given that we have already observed the other variable?

$$P(a \mid b) = \frac{P(a,b)}{P(b)}$$

We can compute the conditional probability using the joint and the marginal probabilities.

$$P(x,y) = P(y)P(x|y)$$

Normalization Trick: We use this to avoid explicitly computing the marginal in the product rule.

$$P(W|T=c) = \begin{bmatrix} \frac{P(s,c)}{P(c)} \\ \frac{P(r,c)}{P(c)} \end{bmatrix} = \alpha \begin{bmatrix} P(s,c) \\ P(r,c) \end{bmatrix}$$

Where $\alpha = \frac{1}{P(c)}$. This would be good to do on paper as well, to demonstrate. Here, in this section of the course, at least, we are assuming that all of them can only have binary values assigned. We have so far looked at joint probability ($P(X,Y)$), marginal probability ($P(X)$), and conditional probability ($P(X|Y)$). We have also looked at two laws:

The law of total probability: $P(x) = \sum_y P(x,y)$, which allows us to get the marginal probability from the joint probability, and the product rule: $P(x,y) = P(x|y)P(y)$ which allows us to get the joint probability from the conditional and the marginal.

Chain Rule: What if we have more than two variables? We will apply the product rule more than once. It's recursive almost.

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | x_1, \cdots, x_{i-1})$$

Some examples of this: Assume we have three random variables now. And four too.

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)P(x_4|x_1, x_2, x_3)$$

The pattern just continues like this. If we have n random variables, we have n terms, hence the multiplication term. We can write this in multiple different ways. This way is just the simplest to see.

Probabilistic Inference: Computes a desired probability from other known probabilities.

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

We want to get $P(W)$, so we will use the law of total probability.

$$P(W) = \sum_T \sum_S P(S, T, W)$$

$$P(S) = 0.3 + 0.1 + 0.1 + 0.15 = 0.65$$

$$P(\text{rain}) = 1 - P(S) = 0.35$$

$$P(W \,|\, \text{winter}, \text{hot}) = \alpha P(W, \text{winter}, \text{hot})$$

$$= \alpha \begin{bmatrix} 0.10 \\ 0.05 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}$$

$$P(W|\text{winter}) = \frac{P(W, \text{winter})}{P(\text{winter})} = \alpha P(W, \text{winter})$$

$$= \alpha \sum_T P(W, \text{winter}, T) = \alpha \begin{bmatrix} 0.10 + 0.15 \\ 0.05 + 0.20 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

Bayes Rule: Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y)P(y) = P(y|x)P(x)$$

If we divide by $P(y)$, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)} P(x)$$

Why is this useful? We can compute the conditional without needing the joint distribution. We only use the other conditional and the marginal distributions. We can represent this with a simple Bayes network. The different terms of different names. Posterior is on the left, right is likelihood times the prior. Let's do an example.

Given $P(m) = 1/50000$. $P(S) = 1/100$. Meningitis causes stiff neck 70% of the time, i.e. $P(s|m) = 0.7$. We want to get $P(m|s)$, or the probability that the person has meningitis given that they have stiff neck.

$$\begin{aligned} P(m|s) &= \frac{P(s|m)}{P(s)} P(m) \\ &= 0.7\left(\frac{1}{50000}\right)\left(\frac{1}{100}\right) \\ &= 0.0014 \end{aligned}$$

We can also do it with the normalization trick: $P(-m|s) = P(s|-m)P(-m)$. We know that $P(-m) = 1 - \frac{1}{50000}$. Use that.