Inference

So far we have set the foundation for how we can represent probabilistic models with multiple random variables. These models are especially useful because they let us perform a task called "inference" where we update our belief about one random variable in the model, conditioned on new information about another. Inference in general is hard! In fact, it has been proven that in the worst case, the inference task, can be NP-Hard where n is the number of random variables [1].

First we are going to practice it with two random variables (in this section). Then, later in this unit we are going to talk about inference in the general case, with many random variables.

Earlier we looked at conditional probabilities for events. The first task in inference is to understand how to combine conditional probabilities and random variables. The equations for both the discrete and continuous case are intuitive extensions of our understanding of conditional probability:

The Discrete Conditional

The discrete case, where every random variable in your model is discrete, is a straightforward combination of what you know about conditional probability (which you learned in the context of events). Recall that every relational operator applied to a random variable defines an event. As such the rules for conditional probability directly apply: The conditional probability mass function (PMF) for the discrete case:

Let X and Y be discrete random variables.
Def: Conditional definition with discrete random variables.

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

Def: Bayes' Theorem with discrete random variables.

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

In the presence of multiple random variables, it becomes increasingly useful to use shorthand! The above definition is identical to this notation where a lowercase symbol such as x is short hand for the event $X = x$:

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

The conditional definition works for any event and as such we can also write conditionals using cumulative density functions (CDFs) for the discrete case:

$$\begin{aligned} \text{P}(X \leq a|Y = y) &= \frac{\text{P}(X \leq a, Y = y)}{\text{P}(Y = y)} \\ &= \frac{\sum_{x \leq a} \text{P}(X = x, Y = y)}{\text{P}(Y = y)} \end{aligned}$$

Here is a neat result: this last term can be rewritten, by a clever manipulation. We can make the sum extend over the whole fraction:

$$\begin{aligned} \mathrm{P}(X \leq a|Y = y) &= \frac{\sum_{x \leq a} \mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)} \\ &= \sum_{x \leq a} \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)} \\ &= \sum_{x \leq a} \mathrm{P}(X = x|Y = y) \end{aligned}$$

In fact it becomes straight forward to translate the rules of probability (such as Bayes' Theorem, law of total probability, etc) to the language of discrete random variables: we simply need to recall that every relational operator applied to a random variable defines an event.

Mixing Discrete and Continuous

What happens when we want to reason about *continuous* random variables using our rules of probability (such as Bayes' Theorem, law of total probability, chain rule, etc)? There is a simple practical answer: the rules still apply, but we have to replace probability terminology with probability *density* functions. As a concrete example let's look at Bayes' Theorem with one continuous random variable.

Def: Bayes' Theorem with mixed discrete and continuous.

Let X be a continuous random variable and let N be a discrete random variable. The conditional probabilities of X given N and N given X respectively are:

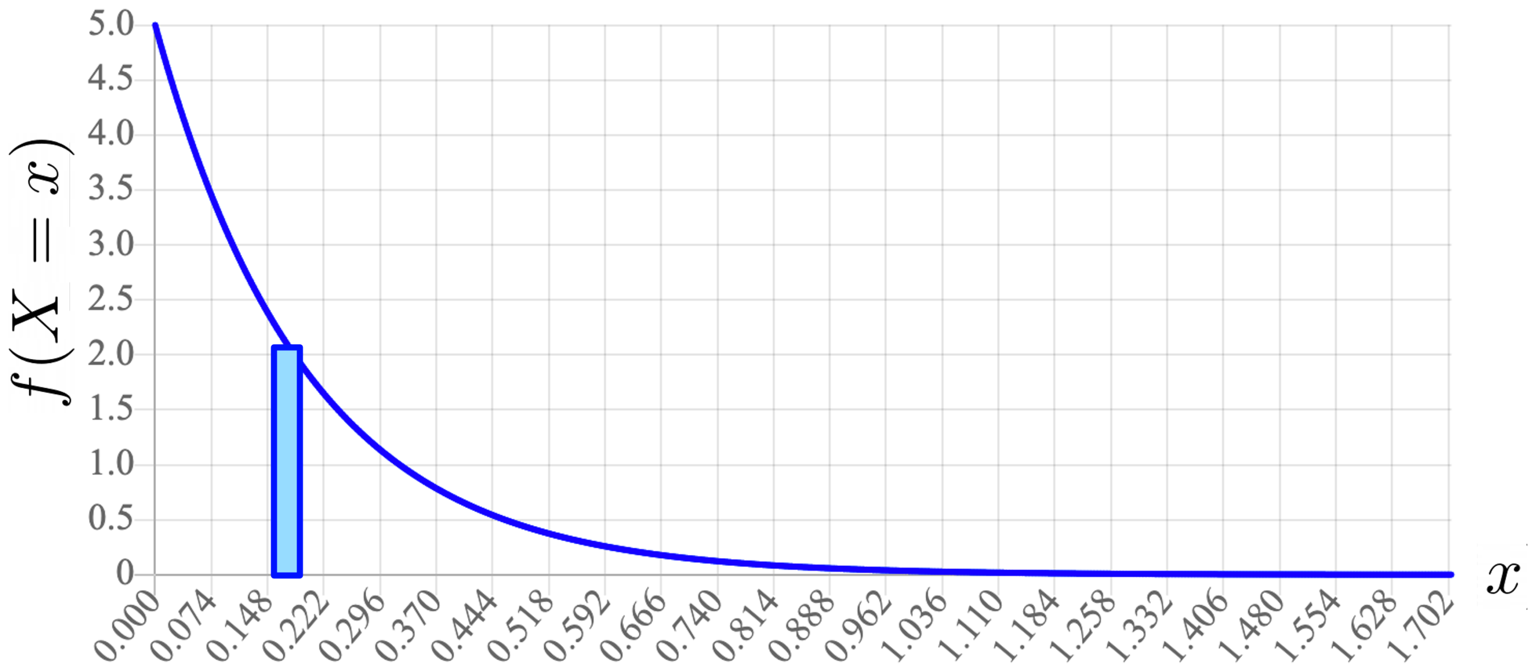
$$f(X = x|N = n) = \frac{\mathrm{P}(N = n|X = x) f(X = x)}{\mathrm{P}(N = n)}$$

$$\mathrm{P}(N = n|X = x) = \frac{f(X = x|N = n)\,\mathrm{P}(N = n)}{f(X = x)}$$

These equations might seem complicated since they mix probability densities and probabilities. Why should we believe that they are correct? First, observe that anytime the random variable on the left hand side of the conditional is continuous, we use a density, whenever it is discrete, we use a probability. This result can be derived by making the observation:

$$\mathrm{P}(X = x) = f(X = x) \cdot \epsilon_x$$

In the limit as $\epsilon_x \to 0$. In order to obtain a probability from a density function is to integrate under the function. If you wanted to approximate the probability that $X = x$ you could consider the area created by a rectangle which has height $f(X = x)$ and some very small width. As that width gets smaller, your answer becomes more accurate:



A value of ϵ_x is problematic if it is left in a formula. However, if we can get them to cancel, we can arrive at a working equation. This is the key insight used to derive the rules of probability in the context of one or more continuous random variables. Again, let X be continuous random variable and let N be a discrete random variable:

$$\begin{aligned}
\mathrm{P}(N=n|X=x) &= \frac{P(X=x|N=n)\,\mathrm{P}(N=n)}{P(X=x)} && \text{Bayes' Theorem} \\
&= \frac{f(X=x|N=n)\cdot\epsilon_x\cdot\mathrm{P}(N=n)}{f(X=x)\cdot\epsilon_x} && \mathrm{P}(X=x)=f(X=x)\cdot\epsilon_x \\
&= \frac{f(X=x|N=n)\cdot\mathrm{P}(N=n)}{f(X=x)} && \text{Cancel } \epsilon_x
\end{aligned}$$

This strategy applies beyond Bayes' Theorem. For example here is a version of the Law of Total Probability when X is continuous and N is discrete:

$$f(X=x)=\sum_{n\in N} f(X=x|N=n)\,\mathrm{P}(N=n)$$

Probability Rules with Continuous Random Variables

The strategy used in the above section can be used to derive the rules of probability in the presence of continuous random variables. The strategy also works when there are multiple continuous random variables. For example here is Bayes' Theorem with two continuous random variables.

Def: Bayes' Theorem with continuous random variables.

Let X and Y be continuous random variables.

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

Example: Inference with a Continuous Variable

Consider the following question:

Question: At birth, girl elephant weights are distributed as a Gaussian with mean 160kg, and standard deviation 7kg. At birth, boy elephant weights are distributed as a Gaussian with mean 165kg, and standard deviation of 3kg. All you know about a newborn elephant is that it is 163kg. What is the probability that it is a girl?

Answer: Let G be an indicator that the elephant is a girl. G is Bern(p = 0.5) Let X be the distribution of weight of the elephant.

$X|G = 1$ is $N(\mu = 160, \sigma^2 = 7^2)$
$X|G = 0$ is $N(\mu = 165, \sigma^2 = 3^2)$

$$\mathrm{P}(G = 1|X = 163) = \frac{f(X = 163|G = 1)\,\mathrm{P}(G = 1)}{f(X = 163)} \qquad \text{Bayes}$$

If we can solve this equation we will have our answer. What is $f(X = 163|G = 1)$? It is the probability density function of a gaussian X which has $\mu = 160, \sigma^2 = 7^2$ at the point x is 163:

$$\begin{aligned} f(X = 163|G = 1) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} && \text{PDF Gauss} \\ &= \frac{1}{7\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{163-160}{7}\right)^2} && \text{PDF } X \text{ at } 163 \end{aligned}$$

Next we note that $\mathrm{P}(G = 0) = \mathrm{P}(G = 1) = \frac{1}{2}$. Putting this all together, and using the law of total probability to compute the denominator we get:

$$\begin{aligned} &\mathrm{P}(G = 1|X = 163) \\ &= \frac{f(X = 163|G = 1)\,\mathrm{P}(G = 1)}{f(X = 163)} \\ &= \frac{f(X = 163|G = 1)\,\mathrm{P}(G = 1)}{f(X = 163|G = 1)\,\mathrm{P}(G = 1) + f(X = 163|G = 0)\,\mathrm{P}(G = 0)} \\ &= \frac{\frac{1}{7\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{163-160}{7}\right)^2} \cdot \frac{1}{2}}{\frac{1}{7\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{163-160}{7}\right)^2} \cdot \frac{1}{2} + \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{163-165}{3}\right)^2} \cdot \frac{1}{2}} \\ &= \frac{\frac{1}{7} e^{-\frac{1}{2}\left(\frac{9}{49}\right)}}{\frac{1}{7} e^{-\frac{1}{2}\left(\frac{9}{49}\right)} + \frac{1}{3} e^{-\frac{1}{2}\left(\frac{4}{9}\right)^2}} \\ &\approx 0.328 \end{aligned}$$