6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Recitation: Inferring a Parameter of Uniform Part 1

Hi. In this problem, Romeo and Juliet are back and they're still looking to meet up for a date. Remember, the last time we met up with them, it was back in the beginning of the course and they were trying to meet up for a date but they weren't always punctual. So we modeled their delay as uniformly distributed between 0 and 1 hour.

So now in this problem, we're actually going to look at variation. And we're going to ask the question, how do we actually know that the distribution is uniformly distributed between 0 and 1 hour? Or it could also be the case that it is uniformly distributed between 0 and half an hour, or zero and two hours. How do we actually know what this parameter of the uniform distribution is?

OK, so let's put ourselves in the shoes of Romeo who's tired of being stood up by Juliet on all these dates. And fortunately, he's learned some probability since the beginning of course, and so have we. And in particular we've learned Bayesian inference. And so in this problem, we're actually going to use basically all the concepts and tools of Bayesian inference that we learned chapter eight and apply them. So it's a nice review problem, and so let's get started.

The set of the problem is similar to the first Romeo and Juliet problem that we dealt with. They are meeting up for a date, and they're not always punctual and they have a delay. But instead of the delay being uniformly distributed between 0 and 1 hour, now we have an extra layer of uncertainty. So if we know sum theta, then we know that the delay, which we'll call x is uniformly distributed between 0 and the theta.

So here's one possible theta, theta 1. But we don't actually know what this theta is. So in the original problem we knew that theta was exactly one hour. But in this problem we don't know what theta is. So theta could also be like this, some other theta 2. And we don't know what this theta is.

And we choose to model it as being uniformly distributed between 0 and 1. So like I said, we have two layers now. We have uncertainty about theta, which is the parameters of the uniform distribution. And then we have uncertainty in regards to the actual delay, x.

OK, so let's actually write out what these distributions are. So theta, the unknown parameter, we're told in the problem that we're going to assume that is uniformly distributed between 0 and 1. And so the PDF is just 1, when theta is between 0 and 1, and 0 otherwise. And we're told that, given what theta is, given what this parameter is, the delay is uniformly distributed between 0 and this theta. So what that means is that we know this conditional PDF, the conditional PDF of x given theta is going to be 1 over theta if x is between 0 and theta, and 0 otherwise.

All right, because we know that given a theta, x is uniformly distributed between 0 and theta. So in order to make this uniform distribution, it's the normalization or the heights, you can think of it, has to be exactly 1 over theta. So just imagine for a concrete case, if theta were 1, 1 hour in the

original problem, then this would just be a PDF of 1 or a standard uniform distribution between 0 and 1. OK, so now this is, we have the necessary fundamentals for this problem.

And what do we do in inference? Well the objective is to try to infer some unknown parameter. And what we have is we have a prior which is our initial belief for what this parameter might be. And then we have some data. So in this case, the data that we collect is the actual observed delayed for Juliet, x.

And this model tells us how this data is essentially generated. And now what we do is, we want to use the data and our prior belief, combined them somehow, and use it to update our belief into what we call our posterior. In order to do that, we use Bayes' rule, which is why this is called Bayesian inference.

So when we use Bayes' rule, remember the Bayes' rule is just, we want to now find the posterior which is the conditional PDF of theta, the unknown parameter, given x. So essentially just flip this condition. And remember Bayes' rule is given as the following. It's just the prior times this conditional PDF of x given theta divided by the PDF of x.

All right, and we know what most of these things are. The prior or just the PDF of theta is 1. The condition PDF of x given theta is 1 over theta. And then of course we have this PDF of x.

But we always have to be careful because these two values are only valid for certain ranges of theta and x. So in order for this to be valid we need theta to be between 0 and 1 because otherwise it would be 0. So we need theta to be between 0 and 1. And we need x to be between 0 and theta. And otherwise this would be 0. So now we're almost done.

One last thing we need to do is just calculate what this denominator is, f x of x. Well the denominator, remember, is just a normalization. And it's actually relatively less important because what we'll find out is that this has no dependence on theta. It will only depend on x. So the importance, the dependence on theta, will be captured just by the numerator.

But for completeness let's calculate out what this is. So it's just a normalization. So it's actually just the integral of the numerator. You can think of it as an application of kind of total probability.

So we have this that we integrate over and what do we integrate this over? Well we know that we're integrating over theta. And we know that theta has to be between x-- has to be greater than x and it has to be less than 1. So we integrate from theta equals x to 1.

And this is just the integral from x to 1 of the numerator, right? This is just 1 and this is 1 over theta. So it's the integral of 1 over theta, d theta from x to 1. Which when you do it out, this is the integral, this is log of theta. So it's log of 1 minus log of x. Log of 1 is 0.

X, remember x is between 0 and theta. Theta is less than 1. So x has to be between 0 and 1.

The log of something between 0 and 1 is negative. So this is a negative number. This is 0. And then we have a negative sign.

So really what we can write this as is the absolute value of log of x. This is just so that it would actually be negative log of x. But because log of x is negative we can just-- we know that this is actually going to be a positive number. So this is just to make it look more intuitive.

OK so now to complete this we can just plug that back in and the final answer is-- this is going to be the absolute value log of x or you could also rewrite this as 1 over theta times absolute value log of x. And of course, remember that the actual limits for where this is valid are very important. OK, so what does this actually mean?

Let's try to interpret what this answer is. So what we have is this is the posterior distribution. And now what have we done? Well we started out with the prior, which was that theta is uniform between 0 and between 0 and 1. This is our prior belief.

Now we observed some data. And this allows us to update our belief. And this is the update that we get.

So let's just assume that we observe that Juliet is late by half an hour. Well if she's late by half an hour, what does that tell us about what theta can be? Well what we know from that at least is that theta cannot be anything less than half an hour because if theta were less than half an hour there's no way that her delay-- remember her delay we know has to be distributed between 0 and theta. There's no way that her delay could be half an hour if theta were less than half an hour. So automatically we know that now theta has to be somewhere between x and one which is where this limit comes in.

So we know that theta have to be between x and 1 now instead of just 0 and 1. So by observing an x that cuts down and eliminates part of the range of theta, the range that theta can take on. Now what else do we know? Well this, we can actually plot this. This is a function of theta.

The log x, we can just think of it as some sort of scaling factor. So it's something like 1 over theta scaled. And so that's going to look something like this. And so what we've done is we've transformed the prior, which looks like flat and uniform into something that looks like this, the posterior.

So we've eliminated small values of x because we know that those can't be possible. And now what's left is everything between x and 1. So now why is it also that it becomes not uniform between x and 1? Well it's because, if you think about it, when theta is close to x, so say x is half an hour. If theta is half an hour, that means that there's higher probability that you will actually observe something, a delay of half an hour because there's only a range between 0 and half an hour that x can be drawn from.

Now if theta was actually 1 then x could be drawn anywhere from 0 to 1 which is a wider range. And so it's less likely that you'll get a value of x equal to half an hour. And so because of that values of theta closer to x are more likely. That's why you get this decreasing function.

OK, so now let's continue and now what we have is this is the case for if you observe one data point. So you arrange a date with Juliet, you observe how late she is, and you get one value of x. And now suppose you want to get collect more data so you arrange say 10 dates with Juliet. And for each one you observe how late she was. So now we can collect multiple samples, say n samples of delays.

So x1 is her delay on the first date. Xn is her delay on the nth date. And x we can now just call a variable that's a collection of all of these.

And now the question is, how do you incorporate in all this information into updating your belief about theta? And it's actually pretty analogous to what we've done here. The important assumption that we make in this problem is that conditional on theta, all of these delays are in fact conditionally independent. And that's going to help us solve this problem.

So the set up is essentially the same. What we still need is a-- we still need the prior. And the prior hasn't changed. The prior is still uniform between 0 and 1.

The way the actual delays are generated is we still assume to be the same given conditional on theta, each one of these is conditionally independent, and each one is uniformly distributed between 0 and theta. And so what we get is that this is going to be equal to-- you can also imagine this as a big joint PDF, joint conditional PDF of all the x's. And because we said that they are conditionally independent given theta, then we can actually split this joint PDF into the product of a lot of individual conditional PDFs. So this we can actually rewrite as PDF of x1 given theta times all the way through the condition PDF of xn given theta.

And because we assume that each one of these is-- for each one of these it's uniformly distributed between 0 and theta, they're all the same. So in fact what we get is 1 over theta for each one of these. And there's n of them. So it's 1 over theta to the n.

But what values of x is this valid for? What values of x and theta? Well what we need is that for each one of these, we need that theta has to be at least equal to whatever x you get. Whatever x you observe, theta has to at least that. So we know that theta has to at least equal to x1 and all the way through xn. And so theta has to be at least greater than or equal to all these x's and otherwise this would be 0.

So let's define something that's going to help us. Let's define x bar to be the maximum of all the observed x's. And so what we can do is rewrite this condition as theta has to be at least equal to the maximum, equal to x bar. All right, and now we can again apply Bayes' rule. Bayes' rule will tell us what this posterior distribution is.

So again the numerator will be the prior times this conditional PDF over PDF of x. OK, so the numerator again, the prior is just one. This distribution we calculated over here. It's 1 over theta to the n. And then we have this denominator. And again, we need to be careful to write down when this is actually valid. So it's actually valid when x bar is greater than theta-- I'm sorry, x bar is less than or equal to theta, and otherwise it's zero.

So this is actually more or less complete. Again we need to calculate out what exactly this denominator is but just like before it's actually just a scaling factor which is independent of what theta is. So if we wanted to, we could actually calculate this out. It would be just like before. It would be the integral of the numerator, which is 1 over theta to the n d theta. And we integrate theta from before, it was from x to 1.

But now we need to integrate from x bar to 1. And if we wanted to, we can actually do others. It's fairly simple calculus to calculate what this normalization factor would be. But the main point is that the shape of it will be dictated by this 1 over theta to the n term.

And so now we know that with n pieces of data, it's actually going to be 1-- the shape will be 1 over theta to the n, where theta has to be at least greater than or equal to x bar. Before it was actually just 1 over theta and has to be between x and 1. So you can kind of see how the problem generalizes when you collect more data.

So now imagine that this is the new-- when you collect n pieces of data, the maximum of all the x's is here. Well, it turns out that it's the posterior now is going to look something like this. So it becomes steeper because it's 1 over theta to the n as opposed to 1 over theta. And it's limited to be between x bar and 1. And so with more data you're more sure of the range that theta can take on because each data points eliminates parts of theta, the range of theta that theta can't be.

And so you're left with just x bar to 1. And you're also more certain. So you have this kind of distribution.

OK, so this is kind of the posterior distribution which tells you the entire distribution of what the unknown parameter-- the entire distribution of the unknown parameter given all the data that you have plus the prior distribution that you have. But if someone were to come to ask you, your manager asks you, well what is your best guess of what theta is? It's less informative or less clear when you tell them, here's the distribution. Because you still have a big range of what theta could be, it could be anything between x and 1 or x bar and 1. So if you wanted to actually come up with a point estimate which is just one single value, there's different ways you can do it.

The first way that we'll talk about is the map rule. What the map rule does is it takes the posterior distribution and just finds the value of the parameter that gives the maximum posterior distribution, the maximum point in the posterior distribution. So if you look at this posture distribution, the map will just take the highest value.

And in this case, because the posterior looks like this, the highest value is in fact x. And so theta hat map is actually just x. And if you think about it, this kind of an optimistic estimate because you always assume that it's whatever, if Juliet were 30 minutes late then you assume that her delay is uniformly distributed between 0 and 30 minutes. Well in fact, even though she arrived 30 minutes late, that could have been because she's actually distributed between 0 and 1 hour and you just happened to get 30 minutes. But what you do is you always take kind of the optimistic, and just give her the benefit of the doubt, and say that was actually kind of the worst case scenario given her distribution.

So another way to take this entire posterior distribution and come up with just a single number, a point estimate, is to take the conditional expectation. So you have an entire distribution. So there's two obvious ways of getting a number out of this. One is to take the maximum and the other is to take the expectation.

So take everything in the distribution, combine it and come up with a estimate. So if you think about it, it will probably be something like here, would be the conditional distribution. So this is called the LMS estimator. And the way to calculate it is just like we said, you take the conditional expectation.

So how do we take the conditional expectation? Remember it is just the value and you weight it by the correct distribution, in this case it's the conditional PDF of theta given x which is the posterior distribution. And what do we integrate theta from? Well we integrate it from x to 1. Now if we plug this in, we integrate from x to 1, theta times the posterior.

The posterior we calculated earlier, it was 1 over theta times the absolute value of log x. So the thetas just cancel out, and you just have 1 over absolute value of log x. Well that doesn't depend on theta. So what you get is just 1 minus x over absolute value of log x.

All right, so we can actually plot this, so we have two estimates now. One is that the estimate is just theta-- the estimate is just x. The other one is that it's 1 minus x over absolute value of log x. So we can plot this and compare the two.

So here's x, and here is theta hat, theta hat of x for the two different estimates. So here's you the estimate from the map rule which is whatever x is, we estimate that theta is equal to x. So it just looks like this. Now if we plot this, turns out that it looks something like this.

And so whatever x is, this will tell you what the estimate, the LMS estimate of theta would be. And it turns out that it's always higher than the map estimate. So it's less optimistic. And it kind of factors in the entire distribution.

So because there are several parts to this problem, we're going to take a pause for a quick break and we'll come back and finish the problem in a little bit.

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013