EPISODE 100

[INTRODUCTION]

[0:00:10.4] SC: Hello and welcome to another episode of TWiML Talk, the podcast where I interview interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

A quick thanks to everyone who participated in last week's TWiML online meet up. It was another great one. If you missed it, the recording will be posted to the meet up page at twimlai.com/meetup. Definitely check it out.

I never seized to be amazed by the generosity and creativity of the TWiML community, and I'd like to send a special shout out to listener Shirin Glander for her exceptional sketch notes. Shirin has been creating beautiful hand-sketched notes of her favorite TWiML episodes and sharing them with the community. Shirin, we truly love and appreciate what you're doing with those, so please keep up the great work. We'll link to her sketch notes in the show notes for this episode and you should definitely follow her on Twitter, @shiringlander for more.

This is your last chance to register for the Rework Deep Learning and AI Assistance Summit in San Francisco, which are these Thursday and Friday, January 25th and 26th. These events features leading researchers and technologists like the ones you heard in our Deep Learning Summit Series last week. The San Francisco event is headlined by Ian Goodfellow of Google Brain, Daphne Koller of Calico Labs and more. Definitely check it out and use the code TWIMLAI for 20% off of registration.

In this episode of the show, I host the largest group of guest I've ever had on a single podcast. I speak with Arthur Gretton, Wittawat Jitkrittum, Zoltán Szabó and Kenji Fukumizu who, alongside with Wenkai Xu, authored the 2017 NIPS Best Paper Award Winner; A Linear-time Kernel Goodness-of-fit Test.

In our discussion, we cover what exactly a goodness-of-fit test is and how it can be used to determine how well a statistical model applies to a given real-world scenario. The group and I then discussed this particular test, the applications of this work, as well as how this work fits in with other research the group has recently published. Enjoy.

[INTERVIEW]

[0:02:30.9] SC: All right everyone, I am in Long Beach, California at NIPS and I've got distinct pleasure of being seated with a group who has authored one of the best paper award winners here at NIPS. It is by far the largest group that I've ever interviewed at one time on the podcast, and so I'm going to allow them to introduce themselves and please name role, title, affiliation, and maybe while you're at it, a little bit about your research and areas of interest.

[0:03:02.1] AG: Thank you. I'm Arthur Gretton. I'm a professor at the Gatsby Computational Neuroscience unit at University College London. My research interest, generally in how to represent and compare probabilities. On one hand, if you've got two groups of objects, you're trying to tell whether they're similar or different. That's one of the problems that we might address. Another is in reasoning. So if you're observing something that might have an effect on something else, you want to know how strong that effect is and how confident you are in whether that effect exists.

[0:03:37.0] SC: Great. Thanks, Arthur.

[0:03:38.9] WJ: Hello. My name is Wittawat Jitkrittum, I'm from the Gatsby University College London. I'm a Ph.D. student studying computer science. My interests are statistical tests, hypothesis testing and a bit on Bayesian inference.

[0:03:54.5] SC: Okay, great. Kenji?

[0:03:56.3] KF: Hello. I'm Kenji Fukumizu. I'm a professor at the Institute of Statistical Mathematics in Tokyo and I'm basically interested in statistics in the machine learning and my interest is also ranging from the mathematical aspects of machine learning to the application of machine learning.

[0:04:14.2] SC: Okay, great. Zoltán?

[0:04:16.8] ZS: Hello. I'm Zoltán Szabó from CMAT, Center of Applied Mathematics, École Polytechnique France. I'm a research associate professor. Generally, I'm interested in information theory algorithm, how to defined similarities between many different objects; graphs, time series, dynamical systems, vectors and how to apply it in different information [inaudible 0:04:38.6] context including hypothesis testing.

[0:04:42.5] SC: Okay. Who wants to tell us about the paper that you guys submitted?

[0:04:46.9] AG: I'm happy to do that. The paper is a goodness-of-fit test. What this means is that we have some model of the world and we want to tell whether that model reflects the world well or not, and if not, why not? An example that we used in the talk was that if you have a model of where crime occurs in Chicago, you might want to know whether that model accurately predicts where the crime will occur, and if there is some shortcoming to the model, you might want to know where that is.

As an example, we used to rather crude crime model, which was mistakenly predicting crimes in the lake next to the city, where clearly there weren't any crimes. This would automatically flag lows. There's no piracy at the moment in Chicago. This is a specific instance.

The difficulty we encountered with this is that if the model that you have, the world is very complicated, then it can be very difficult to compare this with data. In our crime example, I imagine trying to figure out whether it's likely that somebody's going to commit a crime. Well, this is going to depend on their social network, on the state of the economy, on the political atmosphere and so on. All of these things are very complicated and interact. You can state what the interactions are, but then to figure out how these all combine to create a probability is impossible. It's just not mathematically feasible.

Nonetheless, we might want to know when our test has difficulties, and the way that we've formulated it on our test is able to locate the mismatch between data and model without having to figure out these probabilities and do this impossible computation.

[0:06:27.8] SC: What was the title of the paper?

[0:06:29.8] WJ: The title is A Linear-time Kernel Goodness-of-fit Test.

[0:06:34.6] SC: Linear-time Kernel Goodness-of-fit Test. Are there a standard set of goodness-of-fit test? That folks would use otherwise?

[0:06:44.9] AG: In certain simple cases, there are goodness of-fit-test. If you have a very simple model that you're comparing with, like a simple Gaussian, then there are other goodness-of-fit test. In the case of the Gaussian, we don't have this complexity that I mentioned, where you have many interacting effects that make computing probabilities, say close from impossible. So what we have tried to address is the case when these models are very complex.

[0:07:11.9] SC: Out of curiosity, what brought you all together from disparate parts of the world?

[0:07:17.9] WJ: Yes, we are like frequent collaborators. Zoltán used to be at Gatsby unit. We worked together before. I and Arthur also frequently visit Kenji at Institute of Statistical Mathematics, so we're kind of frequent collaborators.

[0:07:33.4] SC: When I think about a best paper, there are lots of reasons why a paper might — When best paper brought applicability, elegance of the solution and others. Why do you think this paper was selected as one of the best papers for NIPS?

[0:07:52.9] AG: I think it's a combination of broad applicability and the analysis that comes with it. One, I guess, key thread in machine learning is like the formulating generative models for complicated data. One of the key research directions of deep mind for example is reinforcement learning. They want an agent that is in a world and that is able to learn from that world.

What this requires in practice to train it is a simulation of the world, which is very realistic. One important, I guess, component of that is to know when your simulation of the world is not accurate. What is it missing? I think for this reason, a paper which is able to like troubleshoot these models of the world that people might use in practice is a very much interest across the community.

I think sort of coupled with this sort of very applied, I guess, benefit is that we have some nice theory results. One of the phrases in the title was linear time, so the time that it takes is no more than the time that it takes to sort of look at each item. [inaudible 0:09:00.2] time mean you sort of have to look at all pairs of items. So that becomes very expensive.

What another result that we showed in the paper is of the various ways that one could derive a linear time goodness-of-fit test with the properties that we've talked about. The way that we've proposed is
provably better. So it's going to give you that test which has the number of true positives, I should say, is going to improve faster for this test and for the other alternatives.

[0:09:28.6] SC: When you say linear time, is it linear in the number of examples or the number of distributions that you're trying to fit to?

[0:09:36.9] AG: Linear in the number of examples. We have a single model, crime map of Chicago and then we're looking — Is linear in the number of crimes.

[0:09:44.5] SC: Okay. How does it work? What's the process for applying it?

[0:09:50.4] WJ: These goodness and this new goodness-of-fit test takes in two inputs. The first one is a model. For example, in the Chicago crime example, let's say we want to model spatial density of robbery events in Chicago, let's say. That's going to be the model, just some density function. The second input will be a collection of points. In this case, it's going to be observed robbery events. Each point is just like a latitude-longitude coordinate of one event, for example. On the way that this works, is that the test constructs some function. It's essentially a scoring function that tells where the mismatch is between the model and the data, and once we have the scoring function, we can then use it to find sort of a region where the mismatch is the largest and then we can pinpoint, "Okay. That is where there is a mismatch," then this is how we criticize the model.

In the end, we would get a region indicating, "Okay. This is where the model doesn't fit quite well to the data." Then as a modeler, they can just improve the model based on the hints on the evidence given by the test.

[0:10:58.9] SC: When you say region, region of what? We're not talking about a region of the map of Chicago. We're talking about a region of the map.

[0:11:02.9] WJ: It is actually region of map. Yeah, a really physical region actually.

[0:11:08.0] SC: Does that correspond to — If we're not talking about a map example, but it sounds like this applies more broadly. What's the generalization of a region?

[0:11:20.0] WJ: Right. The technical term would be — It's going to be the domain where your data live in, so whatever that domain is. I don't know. Typically, I guess, if you have data represented as a table, then you have many columns, then I guess columns correspond to the number of features or the number of attributes. Then let's say your data live in D, where D is the number of columns, then that is going to be like the equivalent of the region that I mentioned.

[0:11:50.1] SC: It's going to identify regions in your feature space?

[0:11:54.1] WJ: In the feature space, yeah.

[0:11:55.8] SC: Okay.

[0:11:57.1] ZS: Let me give you maybe two other examples, because this paper is in fact part of a larger, let's say, package of hypothesis test, time hypothesis tests, and we use this technology in, for example, computer region on natural language processing. In the first case you have, let's say, images of, let's say, happy or sad people, and then the question is whether you can detect the difference between the two emotions, or if there's difference which are, let's say, muscles that are responsible for this difference. That's the domain in this case, let's say muscle activities or parts of the face.

In the other application natural language processing, we have, let's say, documents from two different topics, categories, like neuroscience Bayesian inference, and then the questions again, whether these two possibly different topics can be discriminated. If this is the case, what type of keywords you should look at? In this case, that’s the domain. What are the most distinguishing keywords?

[0:13:04.7] SC: In the case of the first example, in applying this result, you would identify, for example, regions of the face that distinguish between the emotions or —

[0:13:17.4] AG: Yeah, that’s right. For instance, if you had sort of emotions that required you to frown or to put the sides of your mouth downwards, like you’re unhappy, which causes the sort of lines on each side to standout, these would be the regions, the spatial regions in the face which distinguish the sort of contended emotions from the angry or upset emotions. Yeah, the domain is sort of all of the possible regions in the face, just the spatial domain, and then the salient regions that matter in distinguishing them are locations around the eyebrows and to each side of the nose and mouth.

[0:13:51.1] SC: One of the areas that it sounds like this plays into is domain of explainability of AI models. Is that one of the primary motivators as well for this research?

[0:14:03.8] WJ: I think so, yes, because — Yeah, it’s basically a way of showing the shortcoming of your model, like what it is that your model fails to explain about your data. I think, yeah, explainability is really one of the reasons I think that the paper was given the award. Yeah.

[0:14:20.8] KF: We are using the more complex model recently, and so it’s very important to say that’s always our — That model correctly reflects a data [inaudible 0:14:32.9].

[0:14:34.1] SC: What are the requirements of the model that would allow us to apply this? Is it applicable to like box set of models or do we need to know something about either the models or the distributions of our data or other things in order to apply this technique?

[0:14:50.9] AG: It does need to be a probabilistic model. By contrast, one might think of these adversarial networks where there you might not have actually a model of the probabilities of the images that it’s generating, if it’s generating images. So what we do require is that we’re able to write the model as something that if you are able to normalize it, to take the sum overall possible states, would be a valid probability.

Some ways of generating data from randomness, just take the data and apply a bunch of transforms to it, but it's not how or if you could turn that into a probability of the outputs given the random noise that you've fed in. We are only able to deal at this stage where the case where you have this thing that you could write as a probabilistic model where you're able to normalize it.

[0:15:40.2] SC: Okay. Maybe taking a step back, what was the fundamental realization or innovation that kind of lead you down the path that lead to this paper? It sounds like, Zoltán, you mentioned this paper is one in a series of works. But what inspired this particular result?

[0:15:58.0] AG: Zoltán mentioned is was one of sequence of works. So the first work was just in comparing sets of objects. In Zoltán example, sets of faces with positive and negative emotions. Notice that model network appeared in that description. It's just I have set A, set B and I want to know why and where these two sets are different.

The second test in the series was a test of whether two things are dependent. So as an example, if you show to a human some movie, then their brain will be active in a way that depends on the movie that they're watching. This dependence —

[0:16:29.9] SC: Can we take a step back as you talk through each of these papers in this sequence, kind of what the key results were as well?

[0:16:38.1]AG: Sure. For the first paper, we were able to show, again, a linear time test of where these two sets of objects were different. Again, it only costly time to compute, so that means linear in the number of objects that we're comparing, and we were also given a diagnosis of where it was that the two sets were different.

[0:16:59.5] SC: My first kind of flash intuition on this would be to try to do some kind of difference or something like that and maybe convolve the images together or something is, are you thinking about in the same domain or more like from pure probabilistic perspective or —

[0:17:17.2] AR: We need to think in terms of sets of objects rather than objects. For example, if one looks at a pair of faces, you can take their difference and you can see where that difference occurs. If you have two sets, it can be that the mean is the same, but once it has a higher variance then the other. So we care about any difference that might occur between two sets of objects.

Yeah, the second paper was about dependence testing, and in this case, again, we have a linear time test. We also care about dependence that might be quite complicated. So the dependence that you might learn in statistics 101, you notice that when one thing increases, like I press own my accelerator, my car goes faster, so linear dependence. What we might care about is dependence that is much more complicated than that. For example, like if you're adjusting a parameter on a robot arm, it might, on average, hit the same target, but it might be that if you under-damped the robot arm, it's half the time above the target, half the time below, so the variance is going up depending on this parameter even though if you just looked at the means, you would get the same means. This is a very trivial example, but illustrating this sort of complexity of dependence. That's, first of all, testing pairs of samples; second, testing dependence, and then this led us to say, "Well, what if we don't have two sets of samples, but a sample and a model? What's the best way to approach that?" and that led to our third paper.

[0:18:43.6] SC: Okay. Right. Does the sequence continue? What's next for the group of you?

[0:18:50.8] WJ: One possible future work is that continuing from the third work, which is — So now that we have a model and we have one data set, right? But in practice, in reality, most of the time your model is probably wrong and you know [inaudible 0:19:07.6] that it is probably wrong, but now a more relevant question is; given two models that are wrong, two competing models that are wrong, which one fits better? This is now a model comparison.

The current version is model criticism. There's only one model. We try to criticize where it goes wrong, but now we can also extend to two models that we know are wrong, but we want to ask, "Okay, which one fits better?" I think this is one possible direction.

[0:19:37.4] SC: Then maybe, "Which fits better, where?" As opposed to a binary [inaudible 0:19:40.2] this sequence could continue.

[0:19:41.0] WJ: Exactly. That's what we are —

[0:19:43.3] SC: Nice. For the rest of you, any kind of parting words as we wrap up?

[0:19:51.4] ZS: There's another possibility of extension. Here are all these lineal time tests. The underlying assumption is that the probabilistic model lies on, essentially on vectors, like [inaudible 0:20:02.1] Euclidean vectors, which is possibly the simplest concept to understand. One might argue maybe there are some dependencies between the coordinates somehow. So like if you, for example, think of graphs or like images, not all the pixels can — were like completely independent. So often, there's some underlying low dimension structure in the background and hidden under the hood. That's another possible direction to extend this framework.

[0:20:32.6] SC: Awesome. Any other thoughts? How's NIPS been for everyone?

[0:20:37.7] AG: Great.

[0:20:38.5] ZS: It's a bit hectic.

[0:20:41.9] KF: I'm very impressed by NIPS giving award to this work, because this is the age of artificial intelligence and there are many people looking at applications with deep learning, but our research are very basic ones, but we are using in any method. We are using many complex models, and the modern grid system is a very important and basic research and I'm very happy that NIPS gives respect, well such type of basic work.

[0:21:12.7] SC: There was a call here at NIPS, somewhat a controversial call for kind of more basic work and more rigor in the way we look at machine learning. Any thoughts or comments on that? It sounds like you would all agree with the general idea?

[0:21:29.0] AG: There was an inspiring call to arms, which has caused a lot of discussion. In a way, I think it's also a call to arms that is already being acted on in the sense that this year there has been a lot more thought about fundamentals of algorithms that are being used very successfully in the past two years. One example that was raised in this call to arms is understanding when you're optimizing with a model with a huge number of parameters are deep learning methods are. What the pathologies of these optimization methods are when correlations cause you to converge very slowly or converge poorly.

Many of the, I guess, presentations by the optimization community, this NIPS, were actually addressing that. I think there was a standing ovation as you know when this call to arms are made, and I think that's because the spirit of this [inaudible 0:22:21.7] was very much in the air that people were saying, "Okay. We've got this amazing success on applications. Now let's understand what's still holding us back," and then that will help us to progress even further.

[0:22:36.2] SC: Great. Arthur, Wittawat, Kenji, Zoltán, thanks so much for taking a few minutes to chat with us about your paper, and congratulations on the award.

[0:22:46.5] WJ: Thank you very much.

[0:22:46.8] AG: Thank you.

[0:22:47.6] AG: Thank you.

[0:22:47.7] KF: It was our pleasure.

[END OF INTERVIEW]

[0:22:51.7] SC: All right everyone, that's our show for today. Thanks so much for listening and for your continued feedback and support. For more information on Arthur, Wittawat, Zoltán, Kenji or any of the topics covered in this episode, head on over to twimlai.com/talk/100.

Of course, we'd be delighted to hear from you either via a comment on the show notes page or via Twitter directly to me at @samcharrington or the show at @twimlai.

Thanks once again for listening, and catch you next time.

[END]