Johner Institute

Interpretability of machine learning models

With Christoph Molnar, Prof. Dr. Christian Johner

Transcript

00:00:05 Speaker 1

Medical Device Insights, a podcast by the Jona Institute for medical device manufacturers, authorities and notified bodies.

00:00:17 Speaker 1

More and more medical devices are using machine learning technologies, i.e. a subgroup of artificial intelligence, and at the same time regulators are also busy on the move.

00:00:29 Speaker 1

for example, the Chinese N.

00:00:31 Speaker 1

N.

00:00:31 Speaker 1

P.

00:00:31 Speaker 1

A.

00:00:32 Speaker 1

has just published a draft in which it also sets further requirements for the topic of machine learning.

00:00:39 Speaker 1

I myself am on the road in a W.

00:00:41 Speaker 1

H.

00:00:41 Speaker 1

O.

00:00:42 Speaker 1

working committee, which has also set itself the goal of helping to develop such a guideline.

00:00:47 Speaker 1

And with these requirements that we find in these guidelines, the topic of interpretability is also emerging more and more.

00:00:56 Speaker 1

and so I had invited Christoph Molnar here, who has been involved in this area for many, many years.

00:01:04 Speaker 1

Hello Christoph, would you like to briefly introduce yourself to what you do and how you have been involved in this field of interpretability and machine learning in the past?

00:01:14 Speaker 2

Yes, hello Christian, thank you for the invitation.

00:01:16 Speaker 2

Yes, I've been dealing with the topic for three and a half years now.

00:01:21 Speaker 2

I started out a bit unusually, that I started writing about it with a book.

00:01:26 Speaker 2

So, my background is also in statistics.

00:01:27 Speaker 2

So, I already had a lot to build on, so to speak, and at the moment I'm doing a P.H.D.

00:01:35 Speaker 2

especially on the topic.

00:01:37 Speaker 2

Well, I also do research myself and yes, I'm more or less finished right now.

00:01:42 Speaker 2

You never know exactly how long it will take.

00:01:45 Speaker 2

Exactly, but I am also very intensively involved with the topic of interpretability in machine learning.

00:01:50 Speaker 2

and have already written software products, for example, a software package in R for it, with which you can then interpret machine learning models.

00:02:00 Speaker 1

I also came across you through this book you just mentioned.

00:02:04 Speaker 1

So for all listeners, we have linked this below in the accompanying materials.

00:02:09 Speaker 1

Before we get into that, you could briefly explain to us what interpretability means exactly.

00:02:15 Speaker 2

Yes, you always get different

00:02:18 Speaker 2

different answers, depending on who you ask.

00:02:21 Speaker 2

Such a very general answer would perhaps be like this, this is the degree to which a human being can understand a machine, i.e. that can understand how a decision was made.

00:02:33 Speaker 2

But of course that's very, a very soft explanation, so it's a very soft definition, and that's also a, I would say, of the bigger war points still on the field or of the difficulties,

00:02:45 Speaker 2

that it is not always easy to show that an explanation or an interpretation is also the correct one of a model.

00:02:52 Speaker 1

Mhm, so you've already arrived at the pitfalls.

00:02:55 Speaker 1

You had recently given me a wonderful paper, they had looked at the interpretability again from 2 directions or dissected its 2 parts, namely once in the area of explainability, so that you can use it as a

00:03:11 Speaker 1

user somehow gets explained what is happening, so to speak, what the black box is doing there and the other is transparency, where you really get a more detailed understanding inside the model.

00:03:24 Speaker 1

But this definition has not been accepted everywhere and that's probably why we still have a softness in these definitions.

00:03:33 Speaker 1

Yes, now you've already talked a bit about the fact that there are still a few pitfalls, even beyond the definition.

00:03:42 Speaker 1

Yes, where do you stand in machine learning and its interpretability and what pitfalls do you see in particular?

00:03:48 Speaker 2

Exactly, yes, we recently wrote our own paper on the subject, a very short one, where we list a bit of what possibilities there are to misinterpret a model.

00:04:00 Speaker 2

In this case, we thought a bit, so if a scientist uses this model, a machine model of machine learning, to analyze his own data,

00:04:10 Speaker 2

But of course it also applies to this, if you use a model in a product and there are many things that you can do wrong in the interpretation, then among other things.

00:04:21 Speaker 2

So it starts with very simple things, such as that a model, if you train it badly, i.e. overfitting, that you adapt too much to the training data, so to speak, but then in the end not at all for new data, which the model can then predict very badly.

00:04:40 Speaker 2

If you interpret such a model now, then of course you get what you then interpret out of the model, so to speak, i.e. what the most important features were and how they influence the prediction.

00:04:50 Speaker 2

Of course, this can then be misleading, as it actually is, so to speak, the facts.

00:04:56 Speaker 1

Do I understand you correctly?

00:04:57 Speaker 1

Then we would actually have a double problem in the area of overfitting.

00:04:59 Speaker 1

We would have a model that would not be so accurate in the real world, with the real data in the field.

00:05:07 Speaker 1

or does not make the predictions so precisely and misleads us in the belief of how it works inside.

00:05:15 Speaker 1

So that would be a double problem, so to speak.

00:05:17 Speaker 2

Not quite, because that's the interpretation of it, so to speak, or if we now have a ranking output to us, for example, which were the most important features,

00:05:26 Speaker 2

Then what's at the top is what was the most important thing for the model.

00:05:30 Speaker 2

But what would be the wrong way, so to speak, is that we then, because we now mistakenly think that

the model is correct and has also used the correct features.

00:05:39 Speaker 2

But because we have overfed, the most important thing, so to speak, not necessarily the most important, would then be, so to speak, in the underlying phenomenon that we are investigating.

00:05:47 Speaker 2

So in the sense, there is also the one, it depends on the focus, whether you only look at the model as itself, then so to speak.

00:05:55 Speaker 2

take this feature ranking again, the most important feature was actually the most important, maybe for the model, but not the most important in reality.

00:06:03 Speaker 1

Which of the models, maybe if we take a step back, you would consider particularly easy and which are particularly difficult in terms of interpretability.

00:06:13 Speaker 2

Of course, there are also differences, so in general it is often said that the simple linear regression models are used as

00:06:22 Speaker 2

easy to interpret or decision trees.

00:06:25 Speaker 2

So linear regression models are actually always, then you actually multiply the input data and add it up and then get a prediction.

00:06:36 Speaker 2

Or in the decision trees you can draw the tree and then explain to yourself, so to speak, or even visually look at how the decision is made by following the path in the tree, so to speak.

00:06:47 Speaker 2

to the decision, for example to the classification.

00:06:50 Speaker 2

And in the linear model, you can interpret these weights that the model estimates and then also see how they influence the prediction.

00:07:01 Speaker 2

What then and the more complex this internal structure becomes, of course, of a model, so we think of neural networks, for example, we actually have such sums, but there we have very, very many, so to speak, mathematical multiplications as well.

00:07:15 Speaker 2

and sums, which are then connected one after the other in the so-called layers.

00:07:19 Speaker 2

Sure, the individual operations are easy to understand.

00:07:21 Speaker 2

So it's also very transparent, if you, if I were to give you all the weights now and the architecture of that of the neural network.

00:07:28 Speaker 2

However, it is no longer comprehensible how the decision is made in the end by the model.

00:07:34 Speaker 1

Mhm, and that's why we need.

00:07:37 Speaker 1

Procedures that help us to understand it anyway, even if we are not now able to somehow understand 100000 or even millions of weights or their effect afterwards on the decision of the model.

00:07:49 Speaker 1

Now we will briefly return to the question of what can go wrong with the interpretation and you just mentioned, you have also published a paper on this, which we have also linked below.

00:08:02 Speaker 1

Could you give us a few more vivid examples, such as

00:08:06 Speaker 1

you think you have understood something, but perhaps you have not understood it correctly.

00:08:11 Speaker 2

A big issue is, so to speak, the causal interpretation.

00:08:15 Speaker 2

So, if you have a model now, it makes certain predictions, then you would also like to look at it, O.

00:08:24 Speaker 2

K., how, how, how does a certain input feature influence my prediction, one would perhaps also like to interpret it causally, whether the fact that it also had a causal connection in the real world,

00:08:36 Speaker 2

with what should come out.

00:08:38 Speaker 2

But you can't always do that, automatically.

00:08:41 Speaker 2

So certain assumptions have to be met.

00:08:43 Speaker 2

So a very simple example would be if I make or build a model now that predicts whether it will rain tomorrow.

00:08:50 Speaker 2

And a good feature could actually be that I look at it, is the floor wet today?

00:08:57 Speaker 2

Because if the ground is wet today, then I know it will probably rain tomorrow.

00:09:01 Speaker 2

So that's now also one.

00:09:03 Speaker 2

Example, where a conflict can arise, so to speak, between finding a feature, a a feature is useful for a prediction, so to speak, but as soon as we include it in the model, we are no longer allowed to interpret it causally.

00:09:15 Speaker 2

So the model would learn that the ground is wet today, tomorrow it will rain.

00:09:19 Speaker 2

But we can't interpret that causally, because if we did, it would mean we could dump water on the ground and into the rain shows for tomorrow.

00:09:26 Speaker 2

In that case, the problem is that the actual causal feature is that it's raining today.

00:09:32 Speaker 2

And that's to blame, so to speak, that the ground is wet and that it could rain sooner tomorrow.

00:09:36 Speaker 1

So that means what could happen is that someone thinks they have discovered a causality or that after interpreting the model, they come to that conclusion, yes.

00:09:47 Speaker 1

I have a feature that is crucial for the prediction, but in itself you had caught the wrong feature.

00:09:54 Speaker 1

Yes, if I understand you correctly, they have already interpreted that this feature is decisive, but the conclusion from this feature is unfortunately a wrong one.

00:10:02 Speaker 1

Would you have an example from medicine, because I think many of our listeners are in the field of medicine, medical devices.

00:10:09 Speaker 2

Yes, so I also have an example from my own research, so to speak, or rather, where I

00:10:16 Speaker 2

worked more as a statistician, we also looked at observational data to see whether a certain drug, so-called TNF-alpha inhibitors, are effective, and in particular we looked at the progression of ossification in the spine, because you have very good drugs against the inflammation itself.

00:10:34 Speaker 2

These are these TNF-alpha inhibitors, but it is still unclear, or what we have investigated, is whether they also have an influence and to stop this progression.

00:10:45 Speaker 2

if we also have a, then rather these statistical models, i.e. a linear regression model used for this.

00:10:51 Speaker 2

But there is also the connection that these drugs help very well to reduce this inflammation.

00:10:59 Speaker 2

But the inflammation in turn has an influence on the progression, of course, i.e. from this ossification in the spine.

00:11:06 Speaker 2

Now, if you look at inflammation, these inflammation values in the model,

00:11:11 Speaker 2

then the model learns itself, which is a very, very good prediction feature for how strong the progression will take place in a certain spine.

00:11:21 Speaker 2

And the effect of the drug, swallowed by it, so to speak, is completely almost over this inhibition of inflammation.

00:11:30 Speaker 2

This means that in the analysis, where we include both and thus have both in our model, it turns out in the end that the drug is not

00:11:37 Speaker 2

was relevant for that, so to speak, does not help in the against this progression of of ossification.

00:11:43 Speaker 2

But that's just a mistake, because it helps by lowering the inflammation.

00:11:49 Speaker 2

So that's another case where if you include the wrong features, so to speak, you don't have the correct ones, so you have to think about what kind of statements you want to have in the end, so not not to have, but what you want to investigate, how do you have to set up your model in order to understand the

00:12:06 Speaker 2

question that interests you.

00:12:07 Speaker 1

If you look at the solutions again very briefly, i.e. in the first example you mentioned, with the weather forecast, I understood you to have taken out the wet ground as a feature, because the more important feature, namely the rain, would have been the rain on the current day, as a factor that has a particularly strong effect on the forecast for tomorrow's weather.

00:12:27 Speaker 1

what should have been taken in the example of your model, in which the drug and these inflammatory parameters have now been included as a feature and which predicted how the disease would then develop.

00:12:41 Speaker 1

That means that this ossification is taking place, how should one have reacted to correct it?

00:12:46 Speaker 2

So in this case, the answer was that we take the anti-inflammatory out of the model.

00:12:52 Speaker 2

And we also did other analyses, because these are so-called mediation analyses, where you look at, so to speak, how much of the effect goes beyond the effect of the reduction of inflammation and how much is the direct effect of the drug, so to speak.

00:13:07 Speaker 2

But if you look at it a bit more generally, so to speak, what do you do in general or how can you approach such problems in general, it helps and that's what we did in this case, a so-called

00:13:19 Speaker 2

to draw a graph like that.

00:13:20 Speaker 2

That means you actually paint the features, so you paint yourself causal arrows, so to speak.

00:13:26 Speaker 2

This is also something where you really need domain experts for this and there are assumptions in it and that's not what you can necessarily always learn automatically from the data.

00:13:34 Speaker 2

So there, that's just this, you can see the correlation, but the causality, for that you often have to put assumptions in addition.

00:13:41 Speaker 1

If I understand you correctly, your recommendation is not to just try to make the models

00:13:48 Speaker 1

to understand that you can find out which features, i.e. input values, are now decisive for a prediction,

but that you take a close look at how these features depend on each other, in order to then take out any intermediate values if necessary, so that you don't come to false causalities.

00:14:07 Speaker 2

So now, if you also think about products, a combination of non-causal features makes a machine learning model more vulnerable, for example

00:14:17 Speaker 2

against attacks, for example, if you now think something like a credit bureau score, if they would now take, for example, how many credit cards you have, that may not be causal at all, if if now it's just an example, so it doesn't have to be true, but if now, for example, more credit cards, but they are all in the positive, would be good for a good credit bureau score, then someone could go here now and just open a lot of credit card accounts.

00:14:38 Speaker 2

But of course it is not causal, his.

00:14:41 Speaker 2

so to speak, the probability of repaying a loan or something has not improved.

00:14:45 Speaker 2

Therefore, this would not be a causal feature and that of course makes the model more susceptible to abuse if it does not use causal features.

00:14:53 Speaker 1

If you've just talked about interpretability itself, so to speak, how will you find out how good a model really is?

00:15:04 Speaker 1

So, if I understand you correctly, one possibility would be to simply use the

00:15:08 Speaker 1

Box, see what's going on inside, help develop a deeper understanding.

00:15:14 Speaker 1

Is it possible to tell how powerful a model is purely on the basis of quality parameters, i.e. I might say sensitivity or accuracy, or are there any pitfalls that you should watch out for?

00:15:27 Speaker 2

So the important thing is of course always, so the interpretability is of course one thing.

00:15:32 Speaker 2

The other thing is,

00:15:34 Speaker 2

Of course, that you evaluate it properly.

00:15:35 Speaker 2

This means that you evaluate it on test data and also look at the performance number, which is of course also good if you also look at the uncertainty of it, so to speak.

00:15:46 Speaker 2

So, how much does it spread?

00:15:48 Speaker 2

Because often you see each other, so this is also a very general problem, often you only look at one number at a time, but these numbers are always estimated with data.

00:15:55 Speaker 2

So that can be something like the sensitivity that we, that we measure, but of course it can also be if we now calculate the feature importance, because

00:16:03 Speaker 2

we estimate all this with data and that is fraught with uncertainty.

00:16:07 Speaker 2

That's why the recommendation is always to always look at this range or the variance of such values.

00:16:14 Speaker 2

Because especially if they spread a lot, it can of course be that you get a higher value or then misinterpret it, so to speak.

00:16:21 Speaker 1

Even if I summarize it again, the recommendations that you implement and thus also give to medical device manufacturers, I think I have now heard 3.

00:16:31 Speaker 1

The first is

00:16:32 Speaker 1

to try to really understand these models as to take on the topic of interpretability, so to speak.

00:16:39 Speaker 1

And I think your book is a great help.

00:16:43 Speaker 1

By the way, the book can be viewed free of charge on the web at GitHub.

00:16:47 Speaker 1

The second tip I have heard is, if you have now taken a closer look at the model, not to succumb to any mistakes, that you think you have understood it, but have not done it in the sense that you

00:17:01 Speaker 1

Confused correlation and causalities or failed to discover real causalities, such as dependencies on features.

00:17:10 Speaker 1

And the third tip I've heard is that if you have determined quality parameters, you just mentioned sensitivity again, that you don't rely on this measure alone,

00:17:23 Speaker 1

but to find out what uncertainties these parameters are actually associated with.

00:17:29 Speaker 1

So, what is the confidence interval and probably you don't just have to specify it as a value with a confidence interval, but make it dependent again, for example on the patient population.

00:17:43 Speaker 1

In other words, that you may be able to use this value with your uncertainty over the entire age range or for different age intervals

00:17:52 Speaker 1

.

00:17:53 Speaker 1

Did I understand that correctly or were those the things you actually said?

00:17:56 Speaker 2

Yes, exactly, of course there is much more.

00:17:58 Speaker 2

Of course, you have only looked at a few minor things, but yes, I think they are at least important topics.

00:18:05 Speaker 1

Also for all those who want to know more about it now, my tip would be to study the two sources just mentioned.

00:18:13 Speaker 1

One is this paper, which highlights the most common pitfalls.

00:18:18 Speaker 1

in the interpretability of machine learning methods.

00:18:22 Speaker 1

And the second would be Christoph Molnar's book: ‚Interpretability of Machine Learning'.

00:18:27 Speaker 1

This is published in English and free of charge on GitHub.

00:18:31 Speaker 1

There are other articles that we also link in that we have here in the blog on our specialist article.

00:18:37 Speaker 1

These include, for example, articles on the validation of machine learning libraries and an article on regulatory requirements.

00:18:45 Speaker 1

We also have

00:18:47 Speaker 1

video trainings in the audit, on exactly these topics and a guideline that Christoph and I have created together and which is now currently being used at the W.H.O.

00:18:57 Speaker 1

is also developed.

00:18:59 Speaker 1

And I think they have a whole lot of sources and ideas about how they can test their machine learning methods, their models in their medical devices, so that we can achieve what we all want, namely medical devices that are safe and that are

00:19:17 Speaker 1

have the highest possible clinical efficacy.

00:19:19 Speaker 1

Christoph, thank you very much for joining us.

00:19:22 Speaker 2

Gladly, thank you for the invitation.