w10_2

Tue, Nov 22, 2022 1:28PM 50:17

SUMMARY KEYWORDS

explanation, model, surrogate, linear model, segments, instance, feature, data, image, explain, assumptions, prediction, decision boundary, insights, method, important, case, classification, representation, counterfactual

00:00

everyone online, I'm good.

00:05

Alright, everyone online, we are back.

00:11

Right, so

00:17

let's have a look at how we can evaluate explanations. And like I said, this is purely conceptual. So this is more or less the framework within which we're operating. So you have your model that is trained on data. That's, that's capturing some aspects of a natural phenomenon.

00:40

And then this model provides predictions that then you possibly feed into your explainer, or the explainer is part of the model if you're dealing with Auntie explainability. And then these explanations end up together with predictions with the explaining who can possibly interact with both the model and the explainer.

01:05

But also the natural phenomenon, which then leads to observations that develop into a mental model. So more or less understanding of how,

01:18

how a particular phenomenon works.

01:24

If you have explanations in that context, then asking the question,

01:31

that's an explanation the work

01:34

is fairly vague. And even if you can show that the particular method works, whatever that may mean, that doesn't really tell you much.

01:46

So there are more details,

01:51

frameworks to evaluate explained ability. And this is one of them fairly popular, where you have three different

01:59

levels on which you can evaluate your method or your explanations. So let's start at the bottom. Functionally grounded evaluation here, you don't really deal with humans. So you don't do user studies, you just come up with a proxy task on which you evaluate the quality of your explanations, or a metric that you can measure their quality,

02:26

then you may step it one level up and go to human ground evaluation, where you actually deal with humans, you ask humans, to see whether explanations are helpful, useful, understandable.

02:41

But then, if you're evaluating the method itself, you may not really want to deal with a real use case. So you may want to come up with some dummy scenario, such that the users don't need any background knowledge, to be able to tell you whether an explanation works or not.

03:01

And then finally, you might actually go for the real life deployment scenario. Deal with humans that operate in that context. And with the real task. However, as you go up from a proxy task through simple tasks to real task, this becomes more and more difficult and more expensive. So for example, if you want to evaluate your method, with clinicians, their time is scars, you can't really bother them with simple tweaks, and then oh, how about if we change that? Does that work any better, so you have to be very careful

03:37

when you're doing it, and also it may be very difficult to find the resources to do so.

03:42

But this doesn't really capture the complexity, I don't think so. There are two aspects, really that come that are used to build an explanation. One is the algorithm, the method that extracts the insights. And the other one is how you present these insights to the users. Because as I showed you with a linear model, there are multiple ways to show to present the same explanation. And they probably wouldn't be perceived differently. And humans would judge their quality or usefulness differently, obviously, depending on the context, and their background knowledge. So treating both the technical and the social part of explanations as a whole

04:35

may hide some important insights.

04:39

So I think it's a bit better if you

04:43

deal with them separately before you evaluate them altogether. So for example, first of all, you have to make sure that the explainer outputs truthful explanations that they are correct, that you can repeat the same exact story process over and over again get the same

05:00

05:00

results.

05:01

And then only you can choose a communication method and an explanation type that is suitable for your use case, which may be different, depending on the operating scenario. So for example, if you have a very nice figure very nice communication method and your explanation are correct, that's great, you're explaining. But if the explanations look, if the insights look like explanations, and they're convincing, but they are incorrect, then this is potentially harmful, because you're kind of tricking people into believing in something that may not necessarily be true, then, of course, if these are neither convincing nor correct, they will just be ignored. And then if they are not, not effective, but correct, then you're just wasting your effort.

05:50

But this only deals with the quality of explanations, the insights, the thing, the

05:58

the algorithm, the process that generates them, and how you present them. There is another

06:06

way in which you can look at that. So for example, if you're explaining a prediction, that is correct, that's great. That's an explanation. But what if the prediction itself is incorrect, then this is not really, really an explanation. Unless we think of it as an insight into your model that allows you to debug it, or fix it,

06:29

you can go yet one step further. And then assess the coherence of your model with respect to natural phenomenon. So for example, the explanation may be correct, with respect to the predictive model, but that predictive model may not capture Well, the natural phenomenon, in which case, your idea of why something happened, may be correct, but the explanation that you get may not be consistent with it because the model is wrong. Based on the phenomenon, the reasons that govern the phenomenon, it's not that you are incorrect. So there are different levels at which you should look at it. And it's not as simple as just running a few user studies, and showing that something is more effective than something else.

07:21

Right, so let's wrap this part up. So first of all,

07:27

each explainability scenario is unique,

07:29

and usually requires a bespoke solution. You can't just transfer one explainability outright or method or setup into a completely different use case and hope that it will be just as effective. And as useful.

07:48

You need to remember that these are social technical constructs. So there is both the technical part that generates the explanations that extract these insights, but also a social part in which how you frame them, how you present them what assumptions you make with respect to your audience.

08:06

And then providing a single insight may be insufficient. So maybe if you rely on say, counterfactuals, they may not convey the whole story, you may need to use also feature importance or something else to get the full picture of what's happening. And maybe all of them together will be more effective than any individual method, or insight.

08:32

Right, if you want to read some more about this topics, there are two online books that you can follow.

08:41

There is a bunch of papers that summarizes the concept. So two of them I referred to, which is the critique and the human centered explainability. But there's also a general introduction paper that you can follow, and the survey of methods if you're looking for something to apply.

09:00

And then there's a bunch of software that you can use to generate simple explanations. I think we'll make the slides available after the lecture. So you can just click through these links.

09:11

Right.

09:14

The second part covers a few examples, and refers back to some aspects of the taxonomy.

09:25

And then it seems like we'll have enough time to go through one case study. Right. So what are the different kinds of excise and you've already seen that there are really three explanation families, the associations contrasts and causal mechanisms. So for associations you can for example, take something like the importance of your features, either with respect to the entire model, or with respect to a single instance classification of a single instance.

09:54

You may do feature attributions, so how much each feature contributes to

10:00

particular prediction.

10:02

And the rules also show you the association between certain features possibly ranges or features or certain categories for categorical attributes. And the prediction.

10:13

Arguably, you could put here exemplars.

10:17

If an exemplar represents a certain cohort or certain subspace, then it kind of shows the association between the expected feature values, the expected properties of an instance. And its classification results, but also are exemplars in the form of prototypes and criticisms. So prototypes are prototypical instances from across your entire data space. So we basically tried to summarize

10:48

your data space into as few instances as possible. But then this does not really tell you the full story, because they may not necessarily cover certain odd cases. And then you want to include criticisms, which are instances that tell you well, this part isn't this part of the data space is not covered as well. So again, even though these are exemplars, they kind of give you a contrast between what is well represented and what is not. And then non causal counterfactual. So contrast, if statements.

11:21

For causal mechanism, whatever you can get out of a causal model, that's great, whether it's the model as a whole, a causal path, or a causal counterfactual.

11:31

Then again, you can have different accession modalities. And I've discussed numerical summarization, visualization, visualization, and from our argumentation. So for example, you can represent

11:44

the same explanation, again, in different domains. So even though both of them are visual, you could possibly summarize the plot to the right in a paragraph of text. But again, if it's easier to communicate certain explanation, in the original domain, you may want to rely on that. So for example, on the left hand side, you can see rich regions that are important for a particular decision about the regions that either have a positive influence on classification of this image, or negative influence. But you may also present the same information in the transform domain. So here, the numbers refer to superpixel. So the image was first split into non overlapping regions. And then the importance of these regions, the influence of these regions was assessed. So you can communicate in two different ways. Here, the left one does not tell you the magnitudes, but it's easier to comprehend. Whereas the right one gives you more information gives you the magnitudes. But if you just rely on the numbers, it may not be that easy to understand. So here, both of them are complementary.

12:48

Then

12:50

the interactive aspect of explainability. And this is kind of important because

12:58

making a visualization or a system interactive, as in the user interface is interactive, does not necessarily mean that the explainability process is interactive. So for example, if the explanation that you get

13:15

is a visualization of a dataset, and you can zoom in on different parts of it, you can see more data points. That's interactive, but that doesn't allow you to investigate your your dataset, your model more, it would be made into an interactive explainer, for example, if you could move the instances around, and then see how that changes, for example, decision boundary or classification of different points. So it's not about you being able to interact with the interface, it's you being able to affect how the expressions are generated, and what content goes into the explanations.

13:56

Right, let's now go through a few examples of explanations. So here, one, one featuring partners metric is permutation partners. And that's fairly simple. So you have your data set, you predict it with a model. And then you want to see how one particular feature, the importance of one particular feature, what you do is you shuffle the values in that column for that feature.

14:27

you classify these instances, again with your model, and you see how much the predictions have changed. So the reasoning here is that if this feature is important for classification, then it should affect the predictions quite substantially.

14:46

If it's important, if doing this, this grumbling does not affect the predictions that much that it means that this feature is not reused by them all is not very important for classification.

15:00

So that allows you to assess the importance of each feature for your model.

15:07

What about the infrastructure, so here,

15:13

the idea is that, let's focus on one of this line. So let's let's use the red line,

15:19

what it tells you is that

15:23

you pick a feature. And you want to know how this feature

15:28

relates the prediction. In this case, this is a probabilistic model. So on the y axis, you have a probability of a particular class from zero to one. On the x axis, you choose a range within that feature that you are interested in. So pick a single data point. And you start changing one feature value, say from 20,

15:50

to 50,

15:52

and all other things being equal, you see how the probability of a chosen class responds to that, you can do this for the entire data set. And this more or less shows you the trend of a particular feature how, how changing a value how a particular value of a feature affects a probability of a given class.

16:16

This gray lines correspond to individual data points. So this would be an explanation of a single instance, then you can obviously plot them for the entire data set, which gives you a

16:31

better picture of what's happening. But then again, you can summarize them, you can compress them into a single line, which is this, I suppose, pink, or orange, that basically takes an average across all of the individual instances. So this tells you the response of the model, partial dependence, whereas individual conditional expectations tells you the response.

16:59

For individual instances, as you can see, it's usually worth looking at both of these, as the average may hide some interesting aspects, like for example, the by modality here, there are two groups of instances.

17:17

For counterfactuals,

17:19

this is actually a method that we developed. And it was one of the first methods that consider the data structure. So in the simplest case, say that we are generating a counterfactual for this instance here. In the simplest case, you just try to find the closest decision boundary and go across it. So good counterfactual explanation for this instance, will be this one right here.

17:45

But first of all,

17:50

it you don't really want to avoid the instances that are on a decision boundary, because the decision boundary can easily shift. So that's one thing, but also

17:59

an instance like that maybe out of distribution. So you may be prescribing a change, you may be advising somebody to do something that is impossible in the real world. Because if this instance, if this counterfactual comes from a low density region, then it means that we've never seen anything like that. So why would somebody be able to achieve that goal. But there's just one problem here. Another one is that we're actually making quite a huge jump. So even if the instance comes from a high density region, we're actually making a jump to a low density region, which means that it may not be possible to get there even though this instance is possible to achieve in the real world, it may be very improbable that you can make such a big jump. So the idea here is that instead of finding the closest counterfactual or even the closest counterfactual from a dense region, you actually want to find a path. So a sequence of changes gradually changes through high density regions that lead you to feasible and actionable counterfactuals. So here, instead of jumping here across, which is optimal with respect to the distance,

19:11

you'd rather go through this high density region and get the counterfactual that is retrieved by following a high density path.

19:22

And then the other another example, which are the some complexity is through fit. So for example, if you want to assess importance of features, beyond individual attributes,

19:35

you may want to consider a combination of features or some human intelligible rules. So one way to do that is to train some logical model on your dataset, such as a random forest, and then you use this around the forest as a feature extractor. So you extract all the possible features from your logical model. And then you use

20:00

This asks binary attributes, so you have your original feature values. And then for each instance, you see whether they satisfy a particular rule and put it as either one or zero, this expands your, your collection of attributes. And then you can train say, a linear model on the expanded range of attributes using the original target value. And this will give you the influence or the importance not only of the original features, but also of their interactions, and more intelligible.

20:32

Rules.

20:35

Right, just to briefly mention,

20:38

you may also want to explain data. But this is somewhat problematic, partly because data is already an implicit model of the world. So you have to use some tools, you were collecting the data in a certain scenario, we have a certain audience, we have certain group of people. So this already affects the but this may be one way to explain data, how it was collected, provenance of your data, etc, any cleaning that you've applied,

21:07

then you may use numerical statistical summarization. So, for example, distribution of features are cost ratio, things like that.

21:18

Using exemplars or prototypes to explain data isn't really explaining the data because here you're already making assumptions. One of these assumptions is the distance metric that you're using. In order to find for example, a

21:32

for K Means if you want to summarize it with centroids, you're making, you're using a distance. So you're here already making assumptions. This is not really explaining data, this is explained data, given a certain assumption. Same with dimensionality reduction, you make a lot of assumptions there. So you need to be careful.

21:54

Then obviously, there is transparent modeling. So the examples that you've seen role lists, linear models, decision trees, there are also more modern approaches that are that perform quite well and are transparent by design.

22:11

And finally, there's postal explainability methods very popular, such as sharp and lime, that when you don't really care about what your model is doing or how it was trained. What it is. You just care about the inputs and the outputs, and you try to infer what the model is doing try to explain it based on its behavior, its response to different inputs.

22:34

Right. So finally, let's have a look at the particular use case. And for that. I chose

22:44

surrogate explainers, and that's just because there are so many parameters there. And it's so easy to break them. So they make for good case study.

23:06

Right, so this is more or less how, on the very conceptual level how Lima Priceline is a very popular surrogate explainer. So you have your black box model. This is the blue and pink shading in the background. That's the decision boundary. And then you pick an instance, that bald Red Cross, and you're trying to explain your blackbox model in the vicinity of this instance. And here, this is achieved by fitting a locker linear models for your approximating this black box decision boundary only luckily, we have a linear model. And when I spoke about fidelity, so why surrogates may not be your best choice. That's because if this alignment, if the shape of the decision boundary of your surrogate does not correspond to the shape of the decision value, or blackbox model, then you're in trouble because you're not really explaining the right thing. So we need to be careful here.

24:08

Now why people like them? Well, first of all, they are model agnostic. So they work with any blackbox, just because they operate on inputs, outputs, you don't care. They're post hoc, so you can focus on optimizing for predictive performance. And only if somebody else Yeah, but what about explainability you throw one of these at your model, and it provides something that looks like explainability. Just amazing. And then they also work with multiple types of data. So you can have the same explain they're applied to image classifiers, text classifiers or classifiers of tabular data. And that's thanks to interpretable representations, which I'll touch upon in a moment.

24:48

But then again, I've told you in the very beginning that they don't really work because they have poor fidelity. So you may want to consider instead you

25:00

yielding

25:02

inherently interpretable models. But while we can't have a process, or we can't have a way to build explainability, for any scenario, we can have that for surrogate. So that's, again, what we studied here.

25:18

We took we said, well, we can't really solve that problem for all the explainability. But we may be can do that for surrogates. So blame is build on yourself. That's a generic framework that allows you to build bespoke bespoke surrogate explainers. So you can actually make them a bit more truthful. You can make them interactive, more flexible, more robust. But it requires, again, a lot of effort. So you don't really get anything for free here.

25:50

And then it also allows you to understand where the surrogates come from, and how to correctly interpret the results. And this is important, because even though you may get some intuition, or if you see a plot, you may think, oh, yeah, this feature is twice as important as the other one, this may not necessarily be the case, because the explainer may make some assumptions that are counterintuitive.

26:15

And again, when you build your surrogate, you should make it suitable for the use case that you're dealing with. Right, so let's have a look at surrogates for images, which is I think, super fun. So are

26:32

here at the bottom, you can see

26:38

that's annoying.

26:47

Here at the bottom, you can see a surrogate explanation. So this basically first splits an image into chunks and then it measures the importance of it chunk towards a particular prediction, in this case, golden retriever, and then it will parse it well, how does this corresponds to the information that I gave you before, there are three components in blimey first one is the interpretable representation. So this is what allows you to deal with different types of data, then you have to augment your data. And then you have to generate the explanation. So interpretable representation in case of images is super pixel segmentation. So this is what you can see to the right, it splits the split the image into non overlapping chunks,

27:35

then, you need to probe how your how your model responds to different modification of the damage. And what you do is you introduce random occlusions. So if you choose black as your occlusion method, you start removing a random number of the super pixels by just flooding them with black. And then you fit these images through your black box, and see how the how this affects the probability of a given class. And then finally, you take this probabilities, and you take your modified images, you train a linear model on that. And then you extract importance of each segment from that linear model. There's one missing piece of information here, how do you go from images into something that a linear model can handle. And this is actual how the interpretable representation for images is encoded. So once you split the image into segments, you basically encode each segment as an entry in binary vector, where one indicates that the original pixel values in that segment are preserved, whereas zero means that you remove them. Since we can't really remove the pixels, we need a proxy and this proxy is occlusion. So here we include it with block, as you can see at the bottom.

28:57

Right, so let's come back to the image, this image is classified as 99%, tennis ball. And then also we get some probability for Golden Retriever and Labrador retriever, you might want to generate an explanation for that, say we start with an explanation for a tennis ball. So here, it tells us well, this segments are colored in green, contribute positively towards predicting a tennis ball, whereas the two red ones contribute negatively? Well, you should be sort of concerned because there are some artifacts. And actually the two segments in the background are the most influential ones.

29:34

Bull is only number 14, and that's here at the very bottom.

29:39

Right, so that's not really what we expect. Let's have a look at another explanation. This next as an explanation for Golden Retriever, and that's actually even worse.

29:54

Because what it tells you is that well, if you want the golden retriever, you should just remove the ball that

30:00

is not what we would consider important for classifying this image as a golden retriever. So then this is concerning, can we study what, what this explainers actually do and what affects the explanations. So this is an interactive example. On the right, you can see a sample of various data points with segments included. And then you can see the explanation to the left. So there are two parameters more or less, that are important for surrogate explanations of images. One is the proxy that you use for information removal. So what color you use for occlusion, and the other one is how you split the image. So for example, if we change the occlusion color, so the main one is, you remove information from the segment by replacing it with the main color of pixels and close by that segment. So if you pay attention to the explanation,

31:04

when we change the occlusion color to block, the expression changes.

31:09

And again, if you change to a different color, it changed. And sometimes

31:15

it's just the change in magnitude. So the importance the influence of each segment, but other times, it actually changes the segments themselves, you can also randomize the color, and this gives you a different explanation.

31:32

Another one is how we split them up. So again, as you increase the number of segments,

31:38

the explanation changes. So you can see that just by going from large segments into slightly smaller segments,

31:48

the the explanation changes.

31:51

And then again, for even higher number of segments. So we decided to study was the influence of these two parameters on image explanations. And this is what you can see here. So on the x axis at the bottom, you see how many segments are occluded out of five. So here, run, say 100 images with randomly selected one segment occluded, here are three segments that work with four segments for all of the segments by and the last color of the line tells you what occlusion color was used. And then on the y axis,

32:32

you can see the ability of the blackbox model to recover the original class. So without any occlusions, the image had 90%, say tennis poll, if you do a random occlusion, how what how much this probability increases or decreases,

32:49

measured as a square there. So you can see that here, most of the purchase, while most of the colors behave similarly. But as you increase the number of segments, there's one that stands out. And that's the mean occlusion color. So what this tells you is that it's not very effective at removing information from the image.

33:12

And when we go back to this example,

33:16

you can actually notice something strange happening as I increase the number of segments.

33:23

Because you take the mean color of each super pixel, this doesn't really remove information just blurs the image. So with a high enough number of super pixels, you can still tell what it is, even if you occlude almost all of the segments, which basically means while all the other colors, on average, were fairly effective, this one isn't, so you shouldn't really use it. In this case, well, in any case, actually. But even though, you could see the average effects here, this doesn't really mean that using any color is safe. So if we're explaining the tennis ball here,

34:08

and if I pick

34:11

green as my occlusion color, and remember that the occlusion color is supposed to remove information, and I'm trying to explain the tennis ball, suddenly the tennis ball is not important anymore. And that's because, again, I made the information removal proxy, extremely ineffective, because what I'm replacing is a green tennis ball with a solid patch of a green color that's not removing information. This just smooths the ball, you can still tell that it's a tennis ball. So again, even though you can study the overall effects, the average effects this doesn't mean that you can then apply these conclusions to particular use cases, you still have to be very careful.

34:53

And then you can repeat a very similar procedure for tabular data. So here they'll use the iris

35:00

dataset, that's not really important. Again, how do the explanations correspond to the intuition? That

35:09

intuition about how surrogates work. So here, the interpretable data representation is you partitioning the features. Because otherwise, you'd have to communicate the importance of very precise values for each feature. So instead of that, instead of saying, Oh, if you move from three point out to 2.9, or 2.9999, that has some effect, you try to simplify the language of the explanations, and you do that by chopping the features into categorical ranges. So here, you have minus 22. Point 5.5 to one, one to plus infinity and the same on the x axis. So that's your interpretable representation. data augmentation, is you randomizing where the point is located, which hyper rectangle you place it in? And then you want to see what the how the model behaves in a particular hyper rectangle? And then again, the explanation generation is, are you fitting a linear model to this data? Again, sorry, again, there's a problem, how do you go from normal representation and the interpreter presentation. So the forward transformation is pretty straightforward, you just discretize but then you require binary concepts. So you move from discretization into binary concepts by encoding by using one for the all of the hyper rectangles aligned with the hyper rectangle in which your explained instance is located. This basically means that this instance does not necessarily

36:47

have to be here, it can be anywhere within it, and you'll still get one. Otherwise, if you're to the left, or to the right, the interpretable concept is switched off at zero. And the same applies to the other axes. However, if you do that, and then as was the case, with images, you sample in the interpretable representation, you explore the neighborhood of this instance, you end up with

37:12

binary numbers that you can't quite convert back to the original representation, which you require in order to prove the model. So that's one of the problems which you can actually overcome for the easy, you just sampling the original domain, and only then you transform, so you need to change the procedure slightly, but you can still get away with a lot of things. Right, so let's have a look at how the explanations change as we change the parameters. So here, the parameters are pretty much just the location of your instance, and how you split the feature range. So you will see that some of these will not have, if you change the split, some of them won't have any effects

37:55

at all. So these two are actually not meaningful changes, because as I explained you if the binary encoding, in the end, what happens you only care about one, one, so this is unfolding of this 110,

38:11

blue 101, the yellow one and 00 with other colors, so some of the changes are not meaningful. And that's because of how the interpretable representations designed, but then you can make meaningful changes that do not affect the explanation, so I can increase the size of the blue band. And as you can see, the explanation does not change.

38:35

But this is not the case with the other chain. So for example, if I move the split on the y axis a bit higher, and you can see that there's just one more point data point here. If I do this, then the explanation changes completely.

38:50

So again, even though these are a simple changes, that actually affect the explanation quite significantly, so that then again, you can ask a question, can we figure out what's what's happening here? And yes, you can. So again, just to remind you, this is how the data is transformed from the original representation into the binary representation. So first discretize and this is the prime notation. So the first segment is 012. And the other one is again on the y axis, same 012 And then you binarize so all of the aligned of the hyper rectangles aligned with the one that you're explaining which in this case, is the star receive one in the binary interpreted representation, so all of this will have one or the others have zero. And then the same for the y axis which results in four different regions here color coded several, one one, the explained region,

39:52

one zero,

39:54

the region aligned with the explained region on the exci

40:00

axis 01, same on the y axis and 00 off diagonal.

40:08

And then if you can actually derive a closed form solution for this sort of an explanation in the tabular data setting, this assumes that your surrogate is ordinary least squares. And pretty much once you get to this point, which describes how to calculate

40:31

the parameters in your surrogate linear model? Well, this tells you what what are the important factors in determining the magnitudes of the important features of the interpretable feature importance? So here, there are actually just two factors that are important. One is the average prediction. So it tells you, Well, you should care about the average prediction in 1110, and 1101. So this means what if your black box is a probabilistic model? What's the average prediction in yellow, and green, and blue and yellow?

41:04

And then what's the proportion of instances in this region. So again, the left part of the equation tells you how many instances are in the EXPLAIN region, in proportion to the entire horizontal band, and how many instances are in the EXPLAIN region in relation to the vertical bound. And once you know, these two properties of your, of your setup, then you can actually calculate the explanation by hand. But you can also manipulate the parameters of your explanation, such that even though it's technically valid, it will not be correct, you can generate misleading explanations that are still technically correct in terms of the explainability procedure.

41:54

Right, so just to wrap things up, our explainability audience are not really monolithic entities, they are complex systems that have a lot of parameters and assumptions need to take care of these, when deploying them. Same as with predictive models.

42:13

And they need to be configured for the problem at hand. So for example, if you know that you're explaining a tennis ball, you should not really use green as your information removal proxy.

42:25

You probably should know where they come from, what assumptions they operate within, in order to be able to appreciate these insights. So otherwise, these are just insights once you know the assumptions of the ingredients that come into them, you can only then claim once you basically have the correct background knowledge, you can only then claim that this are explainers and that you're providing explanations. And finally, since all of this is super complicated, you probably in first place should ask yourself, Do I really need to use this complex data driven model that later on needs to be explained?

43:04

Again, there are a few resources if you're interested. There's a library, there's a hands on tutorial that explores this that was running these help you add a few years ago, and some recordings from summer schools training materials, you can go through all of that. Right. So that's pretty much all that I had to say. One more advertisement, there is a project running across a few Australian universities and UNSW is part of the project. And if you're considering doing a project in either explainability fairness, or any related aspects, feel free to contact foreign or anybody you find listed here, they'll probably be happy to work with you, either on their research projects. If you're considering applying for a PhD, same talk to people, maybe there's something that's that's interesting.

43:55

Great, thanks so much.

44:05

Okay, any questions?

44:13

mentioned early on regarding data explainability and eviction lice do not account for dimensional dimensionality reduction. Can you just explain why that is? So the question was, why dimensionality reduction is not a good explainability technique for data. It's not that it's not a good technique. It's just that it again brings in a lot of assumptions. So you can't expect peacenik for example, which is one of the most popular I think, the dimensionality reduction technique to work equally well again has a lot of assumptions. And you may usually if you play with it long enough, you will get something that looks like a reasonable separation of your data.

45:00

And you have this example, for example, with amis, you can actually use this SNI anomalous, which is the handwritten digit data set. And you can see clusters of digits emerging. Here, we can actually tell that the clustering is correct, because we know how to differentiate digits. But if we don't have any insights into our data, if we don't have the labels, if we did without labels, then you will probably be highly biased. Because once you see clusters emerging, you say, Yeah, okay, that's a good one, right. But you don't really know whether it relies on the right sort of premises. So I'm not saying you shouldn't use it, I'm just saying, there are a lot of assumptions there. So it's not, it is a data explainability technique. But again, you're working with a lot of assumptions, you need to make sure that these assumptions are compatible with your data set, or the distribution of your data set.

45:56

Any other questions?

46:00

Once twice? Yes?

46:04

How do you measure the how far like the decision leaves from the session?

46:09

So good questions? How can we tell how far and instances from distant boundary? Well, ideally, you'd hope that you're using a probabilistic model, and you can get the confidence. And then if it's say, if you get a probabilistic prediction of if it's a binary classification, and you get oh, point five, then you know that it's highly uncertain. So it has to be close to the cinema wondering, if you're doing classification, then you probably can explore

46:37

the decision space. But that's that, more or less boils down to an exhaustive search in case of classification, if you have access to the structure of the model. So for example, if you have access to the structure of a decision tree, you can pick up the decision boundaries, fairly straightforward. So it very much depends on what sort of model you're using, and whether it's probabilistic classification, or classification.

47:03

But there's no one solution, unless you do an exhaustive search.

47:09

Just post hoc explanation whether it was itself like so whoever, just like researching it, try and like forcing an explanation to the model. And how do you?

47:21

So the question was whether post how explainability is just looking for insights that look right, without actually caring? Whether they're correct? And yes,

47:30

basically, because there are post hoc and unless you take a lot of care in generating them, you check all of your assumptions, you make an exhaustive neighborhood exploration, etc.

47:42

Well, yet, you still may be biasing the explanations. But in many cases, when you're generating thing explanations post hoc, so the model is not transparent itself, and you're not relying on the insights from the model,

47:56

then you're running the risk of just stopping the search, when you get something that looks to be looks like an explanation.

48:06

You know, to what extent is explainability models been adopted by industry?

48:13

The question was, what's the most popular method used in industry more or less?

48:20

That's, that's a difficult one, just because,

48:26

um, like I said, unless you actually spend the time engineering the features and training a transparent model, then you have to rely on post hoc methods. And if you are in that space, you usually go for sharp. And that's just because

48:44

there's a nice software package and the visualizations look great. So it just well, it just works out of the box. It's posthoc. It's all agnostic. You can deploy it with anything. It works because there's software and it looks great.

48:59

But then is it really explaining if it's post hoc? That that's for you to decide I suppose.

49:07

Here you're running out of time on there is number one question

49:16

please join me in thanking.

49:35

Thank you very much about making it's a really fantastic turn, please for the love of God and all that is holy. In my experience, servation has not done so already. We might be able to bring it up to 12 and a half percent of the cover if we really could actually. But seriously, thank you so much all the way to just saying wonderful things about you. And around the q&a afterwards. Of course, I do need to pack in a hurry as usual because

50:00

was

50:01

the good lecture of the next lecture is on their way in the day everyone thanks so much again Speak to you soon

50:10

thanks everyone online