P-4

As well the fact that, you can measure this accuracy in different ways. And for particular problems you will have different matrix to measures accuracy.

P-4

yes. There are some NFRs which are straight forward to measure, like accuracy. You can measures accuracy by simple accuracy matrix, f1 scores, root mean square error.

P-4

The usability of machine learning system is a bit tricky to measure, and sometimes you have to come up with this ad hoc matrix to know about it how usable the system.

P-4

The user usually good way to do it like you have a control deployment of your model. You get yourself a few users that are going to interact with your machine learning model. Then you start collecting numbers out of this control deployment and try to measure how useful is this model for this particular set of individuals.

P-4

I can give you an example maybe, we developed this model for one of our customers. This was a bettering chain. Basically what this people was doing that, without using the machine learning system, we have this super long list of services. They have to review manually. This list is like thousands of items. And machine learning system was suggesting which of these items in the list they should focus on. Instead of reviewing thousands of items on average each user was looking at the list of five or very less items those are very important for them. And this was a way to measure the usability of our model. At the end of the project, we could show the management, instead of looking of thousands of items list you have to focus on maybe five, maybe ten items if you use machine learning system. It will take to process two or three days without system, and with system it will take maybe one day. You have a gain of reduction of time, like only half of the time you need to spend your business or even less. That’s how you do bring usability aspect of machine learning system.

P-2

We will need to measure non-functional requirements.

P-2

I guess you run it and then you measure for the response time for example you can measure that right!! That’s the system running. You calculated the time simply.

P-2

Like I said the example different different algorithm running then you compare and then you choose the best one. This is like more higher level testing that I needed than I guess to check if it’s really delivering or reliable deliveres the best algorithm. Then you are needs to check on the higher level.

P-7

Oh yes, then I do.

P-7

So the best way for me to measure something besides the correctness and whatever, for example, the maintainability, is coming, requiring to come back to the same project a few months later and so on. And then trying to move on with the few changes. If you really run the whole project to be really quick and smooth, that means the maintainability is really well. Then I get a pile on my shoulder, but if it didn’t succeed I need to improve it.

P-9

But, for Machine Learning I would guess it would be some form of statistical measure that you will apply as well. You will probably provide in tests lot of input data to your trained neural network and then you will measure what is the decision that your neural network will take, maybe you will accept a certain number of wrong decisions over time.

P-9

I mean for safety requirements; it will be too much the way I mentioned for hardware. If you have the strictest hardware requirements, I think is that you can only have a random failure as a probability of 10^-9 per hour. So, it’s extremely extremely low. And I think something similar you can define for neural networks probably as well, that the only problem then is of course that you have to provide a neural network with plausible input data that represents the operation that you actually want to do with the neural network. And that is what I meant when you need a lot of testing which is expensive.

P-8

Yes, of course, you have to.

P-10

I think, yes, there are a lot of metrics collected each time continuously, let’s say, for smart trains, there is a thing that said there are some trains deployed in France by Thales they want to check the timing, the where the train is now. So, the train is continuously giving some feedback, although there was any latency or not, it was supposed to come to one station at this time, how late it was. So, we try to get these things. This is kind of a scheduling problem and try to make it better. So, in that case, we are also trying to take into account the Machine Learning systems. So continuously we are trying to measure.

P-10

And if you say that, what kind of metrics? It is a continuous process in these applications. We are regularly updating the metrics. We are trying to, you know, make it better and maybe some other things happening. We are also trying to make this better. So, for now, to have a very good system, we try to collect every metrics that can affect the performance of the service. So, it’s a continuous process and we analyze every metrics, and we try to then propose a new roadmap that how we can make it improved or not. So, I think the metric collections if you say that, what kind of metrics? It’s I will say that some kind of background evaluation metrics of the system based on different performance measures, t’s not only about accuracy.

P-3

Yes, some of these NFRs are measured.

P-3

Lots of these are quantifiable, like accuracy is quantifiable. Repeatability is also quantifiable. Consistency of execution, for example, that is also verifiable.

P-5

When you talk about the performance that’s very important. We have a had a lot of problems and that has been a tough one, that’s not specific for machine learning. It’s more when you come to when you have for the example your ipads in the ambulance and you get disconnected and connected and there’s a lot of performance optimization that have been done there and I guess that’s also non-functional in a way. And we try to measure up. And it’s difficult to off course because clients are in different servers and it may work really good that even if we are having a large load on our system and still it can be different thing in the client environment.

P-1

Yes, evaluating ML models is something we do all the time and I think that is NFRs. We are doing that as part of the ML component also measuring the whole digital system.

P-1

Just evaluating ML models that is usual, like you look at different metrics and cross validation etc. So that is general thing. Then if you wanna try this line in prospective study the whole system that has influence on your business or if you wanna like predict the patient outcome or something. If you set up a clinical trial of something then you compare with or without machine or with a doctor’s judgement with machine then compare those and in the end if you do statistical analysis to se whether it is significant difference.

P-6

As I said we do not have software on machine learning yet. But If I see this in perspective, I would like to measure it for sure.

P-6

I think that approach will be similar to the generic software. They will be specialized for machine learning but the approach will be similar.

P-6

I’m just thinking about the testability. I can test the machine learning algorithm. There are some specific ways for providing sample data for training and providing or using data finally for verification.