LIMITS OF MACHINE LEARNING

1. Limits of machine learning

00:00 - 00:12

So far you've seen what's possible with machine learning but there are some limitations as well. In this lesson, we'll talk about two, data quality and explainability.

2. Data quality

00:12 - 00:35

Let's start with data quality. A common phrase in machine learning is garbage in garbage out. This basically means that the quality of the output depends on the quality of the input. With bad data, applications that use machine learning will produce results that are inaccurate, incomplete or incoherent. Let's look at some examples.

3. How it can go horribly wrong

00:35 - 01:07

Amazon HR reportedly used an AI-enabled recruiting software between 2014 and 2017 to help review resumes and make recommendations. The model was found to prefer male applicants because it was trained on resumes submitted to Amazon over the past decade, when many more male candidates were hired. The model downgraded resumes that contain the word "women" or implied the applicant was female, for example because they had attended a women’s college.

4. How it can go horribly wrong

01:07 - 01:42

Microsoft made headlines in 2016 when they announced their new chatbot. Tay could automatically reply to people and engage in casual conversation on Twitter. As more people talked with Tay the chatbot would learn how to hold better conversations. Less than 24 hours after Tay launched, internet trolls had corrupted the chatbot’s personality. Tay started tweeting highly abusive and offensive things. Her in-built capacity to learn meant that she internalized some of the language she was taught by the trolls.

5. Beware

01:42 - 02:08

It should be clear by now that you shouldn't blindly trust your model. Be critical of its output. Awareness of the role of data is key. When working on a machine learning project it is extremely important to pay attention to the data that is used. So although machine learning can be very powerful, keep in mind, a machine learning model is only as good as the data you give it.

6. Quality assurance

02:08 - 02:33

Having high quality data for the task at hand requires: Data analysis including data characteristics, distribution, source, and relevance. A review of outliers, exceptions, and anything that stands out as suspicious. Domain expertise from experts to explain unexpected data patterns. And documentation. The process used must be transparent and repeatable.

7. Explainability

02:33 - 02:44

The second limitation we'll discuss is explainability. One of the biggest challenges in AI is that often machine learning models are considered black boxes.

8. Explainability

02:44 - 03:08

However, sometimes there is a need for AI systems to be transparent about the reasoning it uses, to increase trust, clarity, and understanding. For example, you will have to be able to explain your model to get business adoption from a customer, to prove you are adhering to laws regarding data, and allow for faster and better bias detection.

9. Explainable AI

03:08 - 03:34

Throughout this chapter we have been discussing deep learning. One big drawback that wasn't discussed yet is the lack of explainability. Although deep learning can make very accurate predictions, it's not always clear why the model is making a specific prediction. Methods that allow us to understand the factors that lead to each prediction are also known as explainable AI.

10. Example: Explainable AI

03:34 - 04:11

Let's examine a typical problem in Explainable AI. Suppose a hospital is using a traditional machine learning model to look at diabetes patient data. The trained model can tell us two things. First, it can predict the onset of Type 2 diabetes. Second, it can say which features were important in making this decision. This is the "explainable" part. This additional explainability can provide important insights for doctors, like if blood pressure is an important predictor of future diabetes.

11. Example: Inexplicable AI

04:11 - 04:35

Contrast that example with a typical deep learning problem. Suppose we want to recognize hand-written letters. We don't really care why a particular image was classified as an "a", as long as the predictions are highly accurate. Deep learning is a perfect solution to this problem because we don't care about explainability in this case.

12. Let's practice!

04:35 - 04:40

Let's move on to the final exercises!