Understanding human-level performance

(DESCRIPTION)
Text, Comparing to human-level performance. Understanding human-level performance.

(SPEECH)
The term human-level performance is sometimes used casually in research articles.

But let me show you how we can define it a bit more precisely.

And in particular, use the definition of the phrase, human-level performance, that is most useful for helping you drive progress in your machine learning project.

(DESCRIPTION)
New slide, Human-level error as a proxy for Bayes error. Medical image classification example. An x-ray of a human hand is present as an inset photo.

(SPEECH)
So remember from our last video that one of the uses of this phrase, human-level error, is that it gives us a way of estimating Bayes error.

What is the best possible error any function could, either now or in the future, ever, ever achieve?

So bearing that in mind, let's look at a medical image classification example.

Let's say that you want to look at a radiology image like this, and make a diagnosis classification decision.

And suppose that a typical human, untrained human, achieves 3% error on this task.

A typical doctor, maybe a typical radiologist doctor, achieves 1% error.

An experienced doctor does even better, 0.7% error.

And a team of experienced doctors, that is if you get a team of experienced doctors and have them all look at the image and discuss and debate the image, together their consensus opinion achieves 0.5% error.

So the question I want to pose to you is, how should you define human-level error?

Is human-level error 3%, 1%, 0.7% or 0.5%?

Feel free to pause this video to think about it if you wish.

And to answer that question, I would urge you to bear in mind that one of the most useful ways to think of human error is as a proxy or an estimate for Bayes error.

So please feel free to pause this video to think about it for a while if you wish.

But here's how I would define human-level error.

Which is if you want a proxy or an estimate for Bayes error, then given that a team of experienced doctors discussing and debating can achieve 0.5% error, we know that Bayes error is less than equal to 0.5%.

So because some system, team of these doctors can achieve 0.5% error, so by definition, this directly, optimal error has got to be 0.5% or lower.

We don't know how much better it is, maybe there's a even larger team of even more experienced doctors who could do even better, so maybe it's even a little bit better than 0.5%.

But we know the optimal error cannot be higher than 0.5%.

So what I would do in this setting is use 0.5% as our estimate for Bayes error.

So I would define human-level performance as 0.5%.

At least if you're hoping to use human-level error in the analysis of bias and variance as we saw in the last video.

Now, for the purpose of publishing a research paper or for the purpose of deploying a system, maybe there's a different definition of human-level error that you can use which is so long as you surpass the performance of a typical doctor.

That seems like maybe a very useful result if accomplished, and maybe surpassing a single radiologist, a single doctor's performance might mean the system is good enough to deploy in some context.

So maybe the takeaway from this is to be clear about what your purpose is in defining the term human-level error.

And if it is to show that you can surpass a single human and therefore argue for deploying your system in some context, maybe this is the appropriate

(DESCRIPTION)
The B definition, typical doctor, 1 percent error, is circled.

(SPEECH)
definition.

But if your goal is the proxy for Bayes error, then this is the appropriate

(DESCRIPTION)
The D definition, team of experienced doctors, 0.5 percent error, is circled.

(SPEECH)
definition.

To see why this matters, let's look at an error analysis example.

(DESCRIPTION)
New slide, Error analysis example. Training error, and dev error are listed.

(SPEECH)
Let's say, for a medical imaging diagnosis example, that your training error is 5% and your dev error is 6%.

And in the example from the previous slide, our human-level performance, and I'm going to think of this as proxy for Bayes error.

Depending on whether you defined it as a typical doctor's performance or experienced doctor or team of doctors, you would have either 1% or 0.7% or 0.5% for this.

And remember also our definitions from the previous video, that this gap between Bayes error or estimate of Bayes error and training error is calling that a measure of the avoidable bias.

And this as a measure or an estimate of how much of a variance problem you have in your learning algorithm.

So in this first example, whichever of these choices you make, the measure of avoidable bias will be something like 4%.

It will be somewhere between I guess, 4%, if you take that to 4.5%, if you use 0.5%, whereas this is 1%.

So in this example, I would say, it doesn't really matter which of the definitions of human-level error you use, whether you use the typical doctor's error or the single experienced doctor's error or the team of experienced doctor's error.

Whether this is 4% or 4.5%, this is clearly bigger than the variance problem.

And so in this case, you should focus on bias reduction techniques such as train a bigger network.

Now let's look at a second example.

Let's see your training error is 1% and your dev error is 5%.

Then again it doesn't really matter, seems but academic whether the human-level performance is 1% or 0.7% or 0.5%.

Because whichever of these definitions you use, your measure of avoidable bias will be, I guess somewhere between 0% if you use that, to 0.5%, right?

That's the gap between the human-level performance and your training error, whereas this gap is 4%.

So this 4% is going to be much bigger than the avoidable bias either way.

And so they'll just suggest you should focus on variance reduction techniques such as regularization or getting a bigger training set.

But where it really matters will be if your training error is 0.7%.

So you're doing really well now, and your dev error is 0.8%.

In this case, it really matters that you use your estimate for Bayes error as 0.5%.

Because in this case, your measure of how much avoidable bias you have is 0.2% which is twice as big as your measure for your variance, which is just 0.1%.

And so this suggests that maybe both the bias and variance are both problems but maybe the avoidable bias is a bit bigger of a problem.

And in this example, 0.5% as we discussed on the previous slide was the best measure of Bayes error, because a team of human doctors could achieve that performance.

If you use 0.7 as your proxy for Bayes error, you would have estimated avoidable bias as pretty much 0%, and you might have missed that.

You actually should try to do better on your training set.

So I hope this gives a sense also of why making progress in a machine learning problem gets harder as you achieve or as you approach human-level performance.

In this example, once you've approached 0.7% error, unless you're very careful about estimating Bayes error, you might not know how far away you are from Bayes error.

And therefore how much you should be trying to reduce aviodable bias.

In fact, if all you knew was that a single typical doctor achieves 1% error, and it might be very difficult to know if you should be trying to fit your training set even better.

And this problem arose only when you're doing very well on your problem already, only when you're doing 0.7%, 0.8%, really close to human-level performance.

Whereas in the two examples on the left, when you are further away human-level performance, it was easier to target your focus on bias or variance.

So this is maybe an illustration of why as your pro human-level performance is actually harder to tease out the bias and variance effects.

And therefore why progress on your machine learning project just gets harder as you're doing really well.

(DESCRIPTION)
New slide, Summary of bias, variance with human-level performance. Listed are human-level error, training error, and dev error.

(SPEECH)
So just to summarize what we've talked about.

If you're trying to understand bias and variance where you have an estimate of human-level error for a task that humans can do quite well, you can use human-level error as a proxy or as a approximation for Bayes error.

And so the difference between your estimate of Bayes error tells you how much avoidable bias is a problem, how much avoidable bias there is.

And the difference between training error and dev error, that tells you how much variance is a problem, whether your algorithm's able to generalize from the training set to the dev set.

And the big difference between our discussion here and what we saw in an earlier course was that instead of comparing training error to 0%, And just calling that the estimate of the bias.

In contrast, in this video we have a more nuanced analysis in which there is no particular expectation that you should get 0% error.

Because sometimes Bayes error is non zero and sometimes it's just not possible for anything to do better than a certain threshold of error.

And so in the earlier course, we were measuring training error, and seeing how much bigger training error was than zero.

And just using that to try to understand how big our bias is.

And that turns out to work just fine for problems where Bayes error is nearly 0%, such as recognizing cats.

Humans are near perfect for that, so Bayes error is also near perfect for that.

So that actually works okay when Bayes error is nearly zero.

But for problems where the data is noisy, like speech recognition on very noisy audio where it's just impossible sometimes to hear what was said and to get the correct transcription.

For problems like that, having a better estimate for Bayes error can help you better estimate avoidable bias and variance.

And therefore make better decisions on whether to focus on bias reduction tactics, or on variance reduction tactics.

So to recap, having an estimate of human-level performance gives you an estimate of Bayes error.

And this allows you to more quickly make decisions as to whether you should focus on trying to reduce a bias or trying to reduce the variance of your algorithm.

And these techniques will tend to work well until you surpass human-level performance, whereupon you might no longer have a good estimate of Bayes error that still helps you make this decision really clearly.

Now, one of the exciting developments in deep learning has been that for more and more tasks we're actually able to surpass human-level performance.

In the next video, let's talk more about the process of surpassing human-level performance.