Daniela Witten

Follow @daniela_witten 22.8K followers

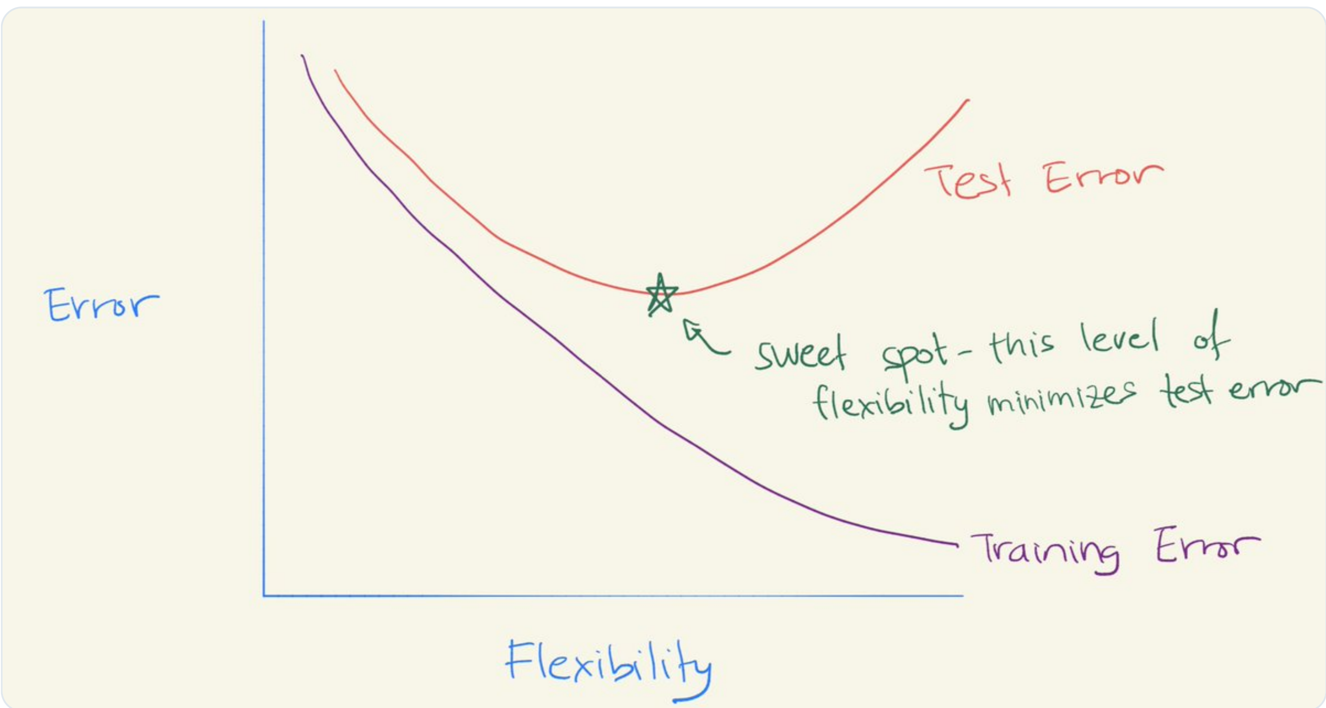
Aug 8th 2020, 23 tweets, 6 min read

The Bias-Variance Trade-Off & "DOUBLE DESCENT" 🧵

Remember the bias-variance trade-off? It says that models perform well for an "intermediate level of flexibility". You've seen the picture of the U-shape test error curve.

We try to hit the "sweet spot" of flexibility.

1/🧵



This U-shape comes from the fact that

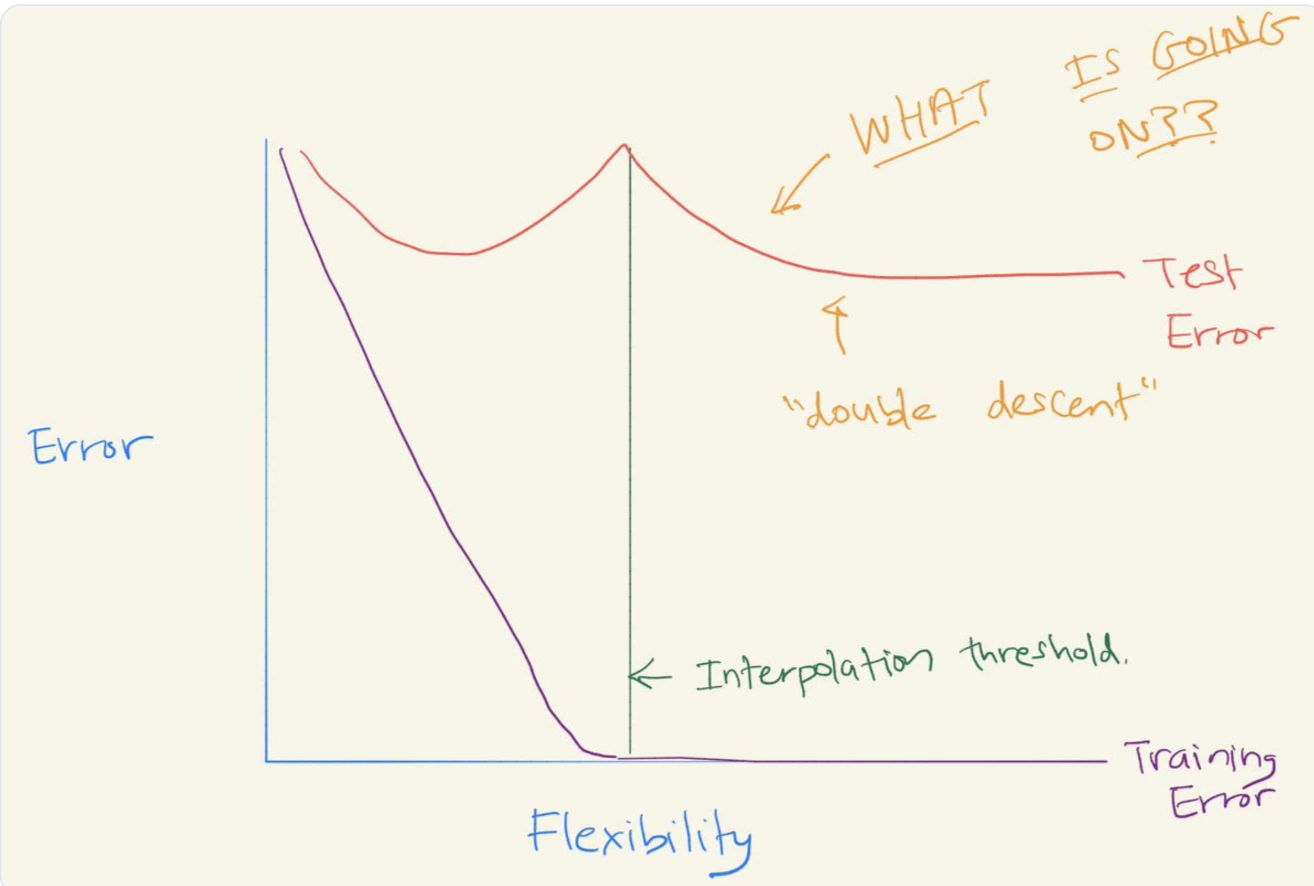
and variance -- i.e. a model with intermediate level of flexibility.

2/

In the past few yrs, (and particularly in the context of deep learning) ppl have noticed "double descent" -- when you continue to fit increasingly flexible models that interpolate the training data, then the test error can start to DECREASE again!!

Check it out:

3/



This seems to come up in particular in the context of deep learning (though, as we'll see, it happens elsewhere too).

What the heck is going on? Does the bias-variance trade-off NOT HOLD? Are the textbooks all wrong?!?!?!

Or is deep learning *magic*?

4/

I promise, the bias-variance trade-off is OK!

To understand double descent, let's check out a simple example that has nothing to do with deep learning: natural cubic splines.

5/

What's a spline? Basically, it's a way to fit the model Y=f(X)+epsilon, with f non-parametric, using very smooth piecewise polynomials.

To fit a spline, we construct some basis functions and then fit the response Y to the basis functions via least squares.

6/

The number of basis functions we use is the number of *degrees of freedom* of the spline.

The basis functions more or less look like this, but the details really aren't that important.

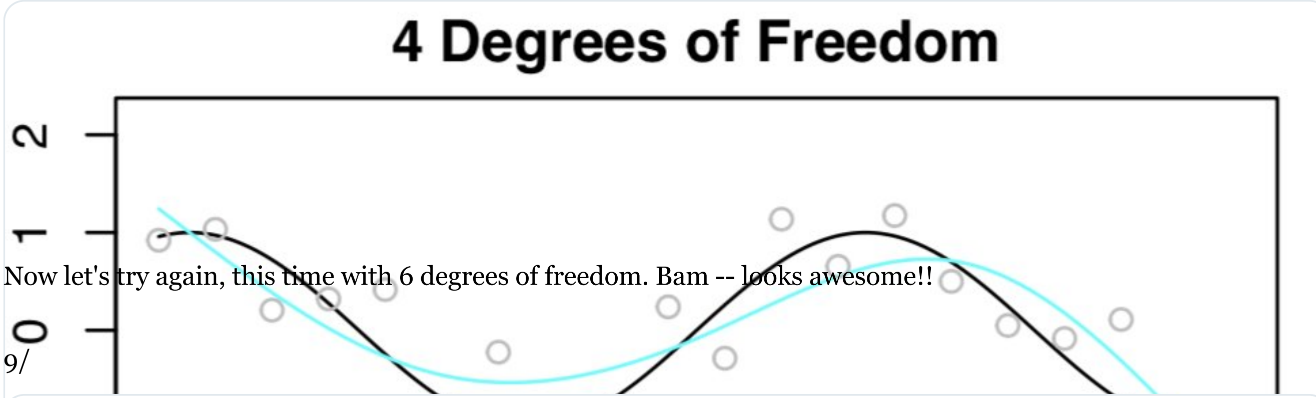
7/

$$(X - \psi_1)^3_+, \ldots, (X - \psi_K)^3_+$$

OK, so, suppose we have n=20 (X,Y) pairs, and we want to estimate f(X) in Y=f(X)+epsilon (here f(X)=sin(X)) using a spline.

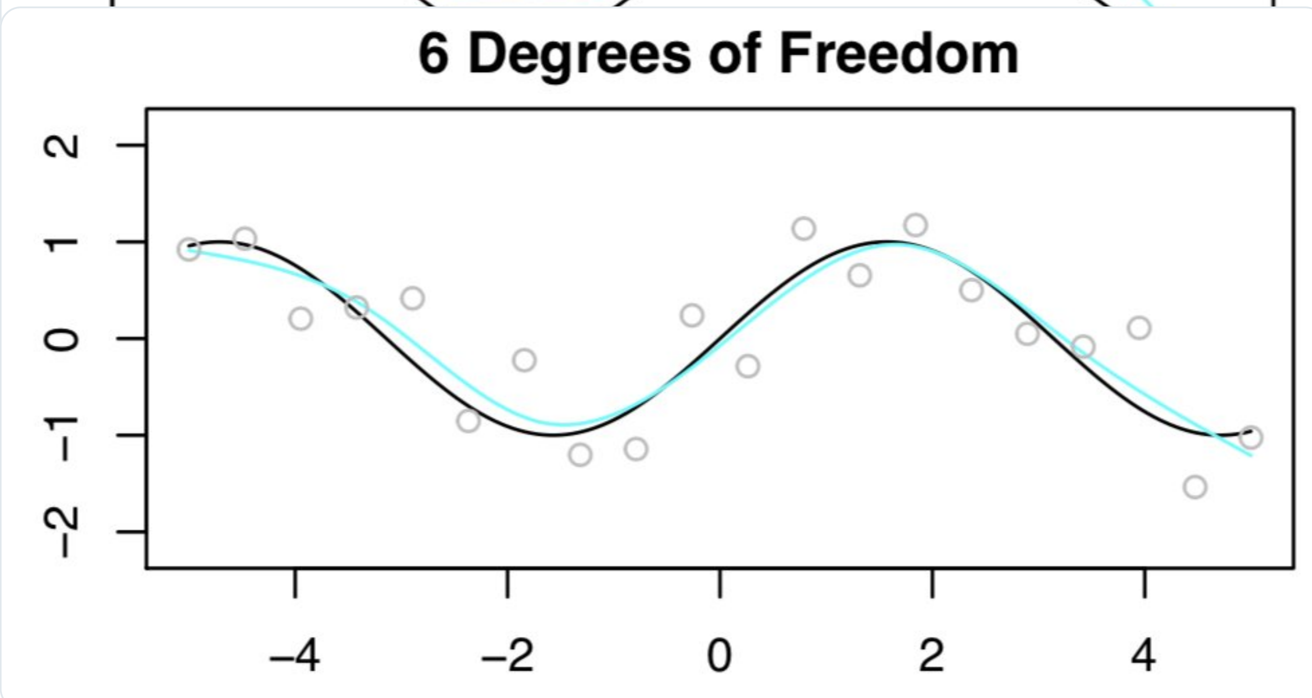
First we fit a spline w/ 4 DF. The n=20 observations are in gray, true function f(x) is in black, and the fitted function is in light blue. Not bad!

8/



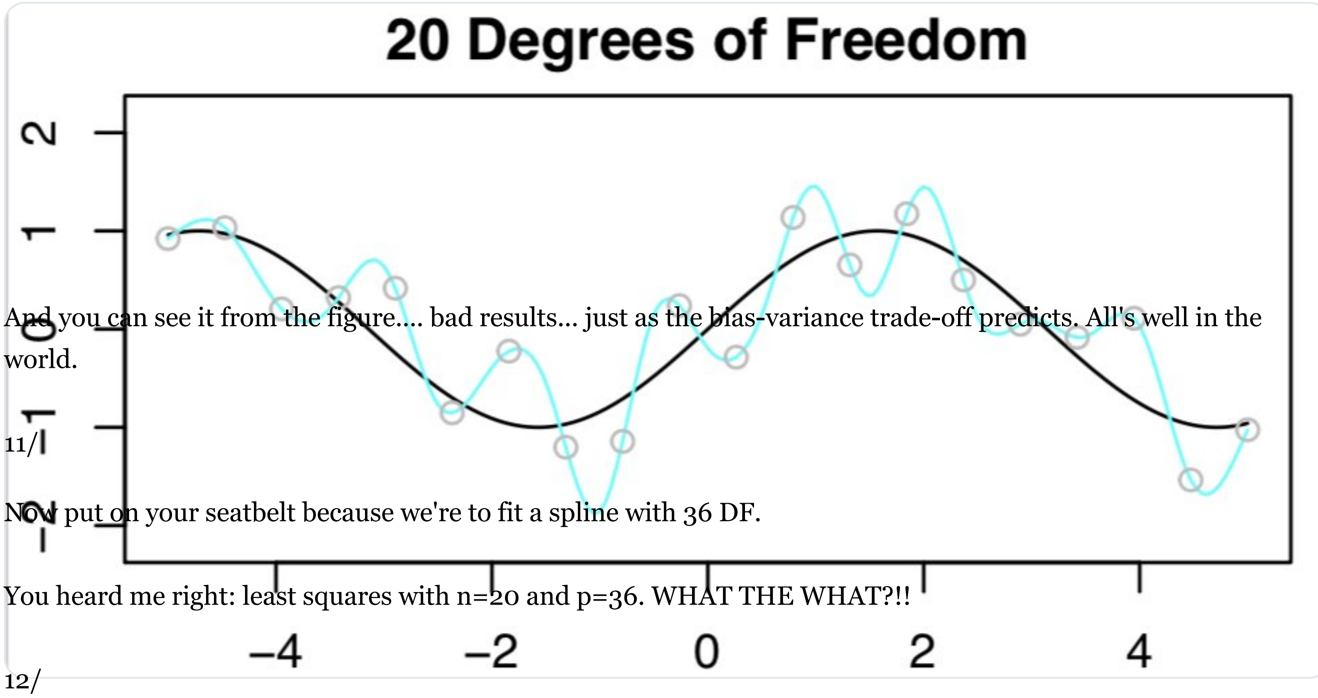
Now let's try again, this time with 6 degrees of freedom. Bam -- looks awesome!!

9/



Now what if we use 20 degrees of freedom? Ummm... this is a bad idea... because we have n=20 observations and to fit a spline with 20 DF I need to run least squares with 20 features!! We'll get ZERO training error (i.e. interpolate the training set) and bad test error!

10/



And you can see it from the figure.... bad results... just as the bias-variance trade-off predicts. All's well in the world.

11/

Now put on your seatbelt because we're to fit a spline with 36 DF.

You heard me right: least squares with n=20 and p=36. WHAT THE WHAT?!!

12/

W/ p>n the LS solution isn't even unique!

To select among the infinite number of solutions, I choose the "minimum" norm fit: the one with the smallest sum of squared coefficients. [Easy to compute using everybody's favorite matrix decomp, the SVD.]

Women in Statistics and Data Science
@WomenInStat

So yesterday I asked you all what you wanted to hear about from me this week, and one answer stood out from all the others: the SINGULAR VALUE DECOMPOSITION (SVD).

twitter.com/genomixgmailco...

1/

Pavitra Chakravarty @genomixgmailcom
Replying to @WomenInStat and @daniela_witten
Can we have your take on SVD? I have heard many people say it is the clearest explanation they ever heard! I could use some help with SVD

11:19 AM · Jul 21, 2020

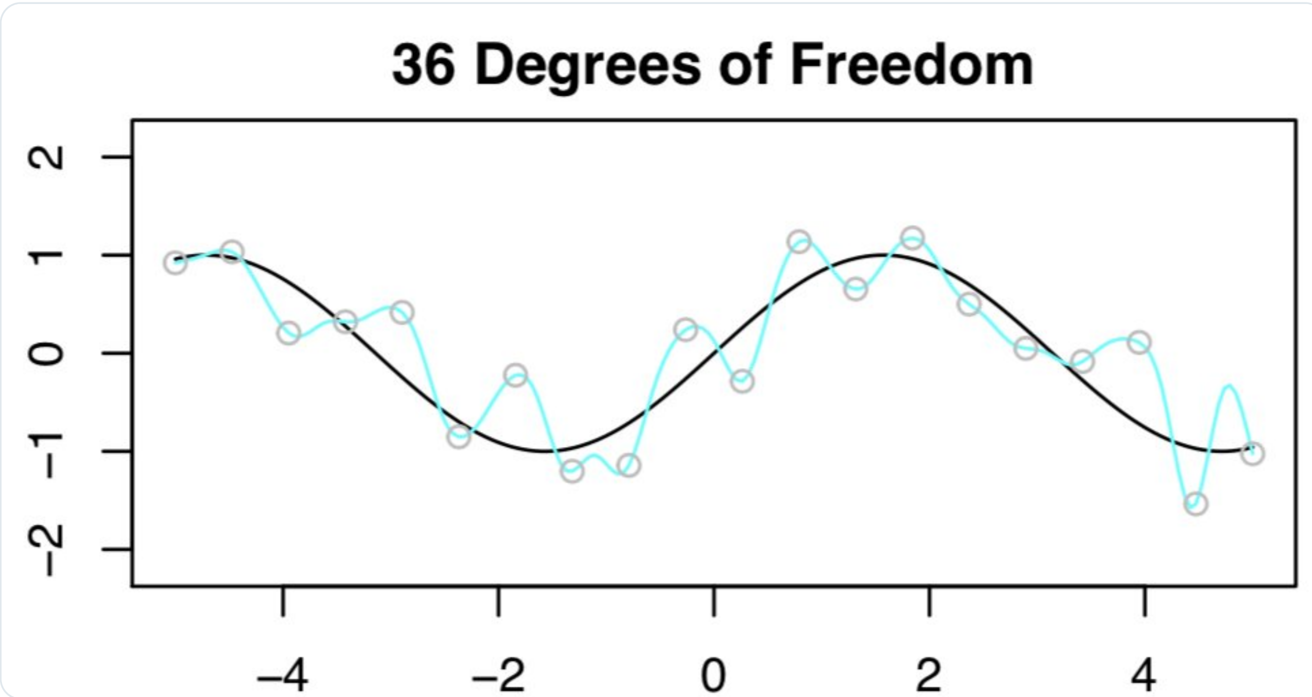
1.6K 459 people are Tweeting about this

The result will be HORRIBLE, because p>n, right??

Right??!!!!!!

Here's what we get:

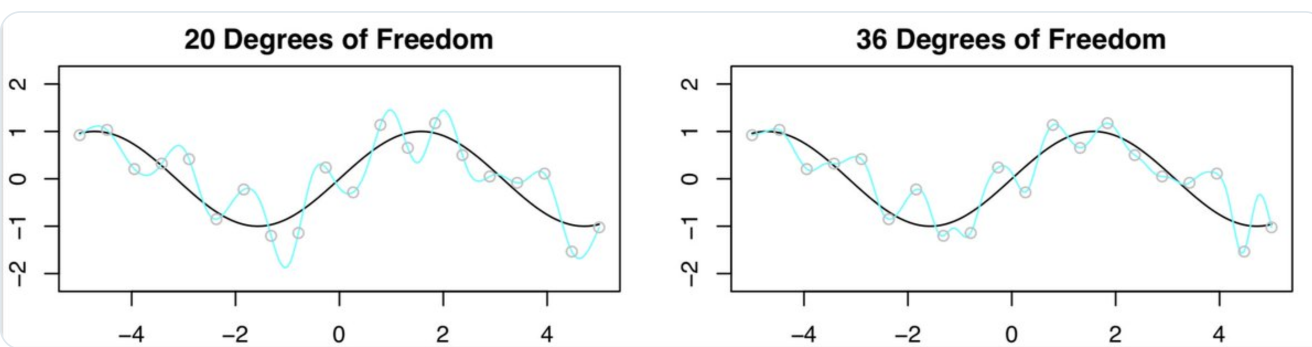
14/



Hmmm... not as bad as we expected... let's compare the results with 20 DF to 36 DF....

what is going on??? Shouldn't the fit with 36 DF look WORSE than the one with 20 DF? If anything, it looks a little BETTER!!

15/

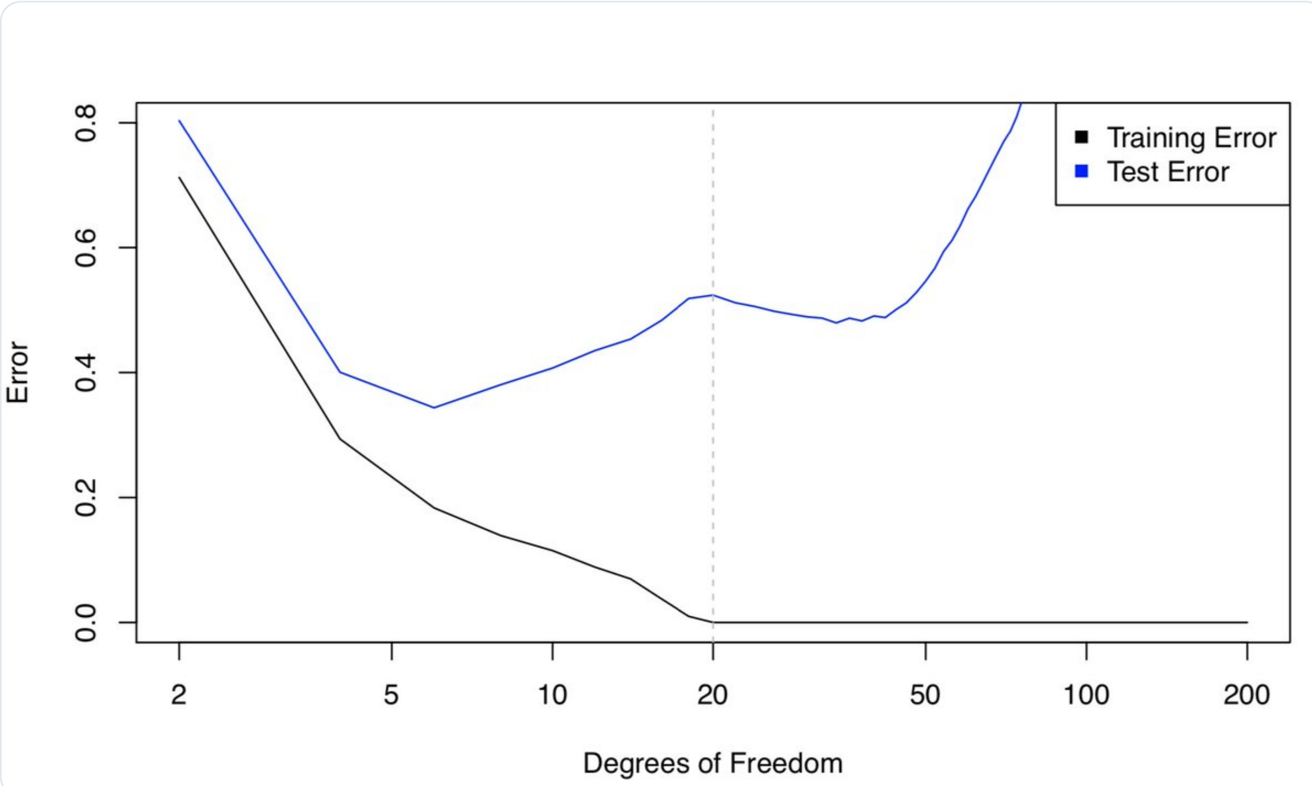


We can take a peek at the training and test error:

WHAT THE HECK IS HAPPENING?!??! Why did the test error (briefly) DECREASE when p>n? Isn't that literally THE OPPOSITE of what the bias-variance trade-off says should happen?

Should we burn our copies of ISL?!

16/



Calm down!! This actually makes sense.

The key point is with 20 DF, n=p, and there's exactly ONE least squares fit that has zero training error. And that fit happens to have oodles of wiggles.....

17/

.... but as we increase the DF so that p>n, there are TONS of interpolating least squares fits.

The MINIMUM NORM least squares fit is the "least wiggly" of those zillions of fits. And the "least wiggly" among them is even less wiggly than the fit when p=n !!!

18/

So, "double descent" is happening b/c DF isn't really the right quantity for the the x-axis: like, the fact that we are choosing the minimum norm least squares fit actually means that the spline with 36 DF is **less** flexible than

Crazy, huh?

19/

Now... what if had used a ridge penalty when fitting the spline (instead of least squares)?

Well then we wouldn't have interpolated training set, we wouldn't have seen double descent, AND we would have gotten better test error (for the right value of the tuning parameter!)

20/

How does this relate to deep learning?

When we use (stochastic) gradient descent to fit a neural net, we are actually picking out the minimum norm solution!!

So the spline example is a pretty good analogy for what is happening when we see double descent for neural nets.

21/

So, what's the point?

✅ double descent is a real thing that happens
✅ it is not magic 🚫
✅ it is understandable through the lens of stat ML and the bias-variance trade-off.

Actually, the B/V T/O helps us understand *why* DD is happening!

No magic ... just statistics

22/

But then again ... statistics is magical!! 💫 💫 💫

Thanks to my co-authors @robtibshirani @HastieTrevor and Gareth James for discussions leading to some of the ideas in this thread

24/24

Try unrolling a thread yourself!

1) Follow Thread Reader App on Twitter so you can easily mention us!

2) Go to a Twitter thread (series of Tweets by the same owner) and mention us with a keyword "unroll"

`@threadreaderapp unroll`

You can practice here first or read more on our help page!

More from @daniela_witten

see all

Daniela Witten
@daniela_witten
Aug 7th 2020

When I teach cross validation in my Intro to Statistical Learning course, I literally spend a class on "potential pitfalls of CV" and this is the main error I talk about. Happens all the time in published biology literature- not just for methylation data 1/

This error is particularly stressful to me as a statistician, because it means that my data analysis can be totally wrong due to data pre-processing that may have been performed before I ever saw the data and that I don't know about 2/

Read 8 tweets

Daniela Witten
@daniela_witten
Jun 24th 2020

The Fisher Lecture: an epilogue. I am delighted that COPSS decided to retire the name of the Fisher Award and Lecture, and to create a new COPSS Distinguished Achievement Award and Lecture. A lot has happened during the past 3 weeks since I posted this original thread. 1/

Read 16 tweets

Daniela Witten
@daniela_witten

much tragedy unfold. So much anguish from Black colleagues here on twitter. And so I've been trying to think of ways that *I* can improve my tiny corner of the world. A thread on why change is hard in academia 1/

Maybe you have heard of Ronald Fisher, "a genius who almost single-handedly created the foundations for modern statistical science" and "the single most important figure in 20th century statistics". (Geneticists: he is also well-known in

Read 19 tweets