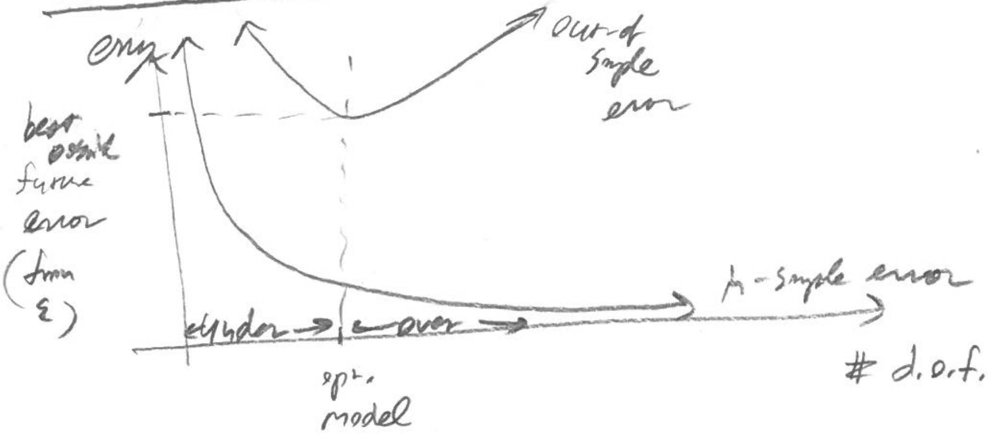
Lec 15 Math 390.4 3/20/18

Under vs. Overfitting

1



p59 (Learning from Data)

In sample error: SSE, s_e computed from $\mathbb{D}$

Out-of-sample error: oosSSE ooss_e computed from new future data $\mathbb{D}^*$

(oos)

your score on practice exam that you get to see

" " " " " the real exam " " " " doing " " " "

In-sample error can go to zero since you can merely memorize all $i=1 \ldots n$ answers, the $\{y_1, \ldots, y_n\}$. If you've merely memorized, then you haven't learned the theory / concepts / principles which is $f(x)$ or at the very least, $h^*(x)$.

<10N, 3m/s²>, ...

Imagine if Newton only memorized force-mass-accel data? Instead of memorizing, the law F=ma is the dimension reduction, the principle that is learned.

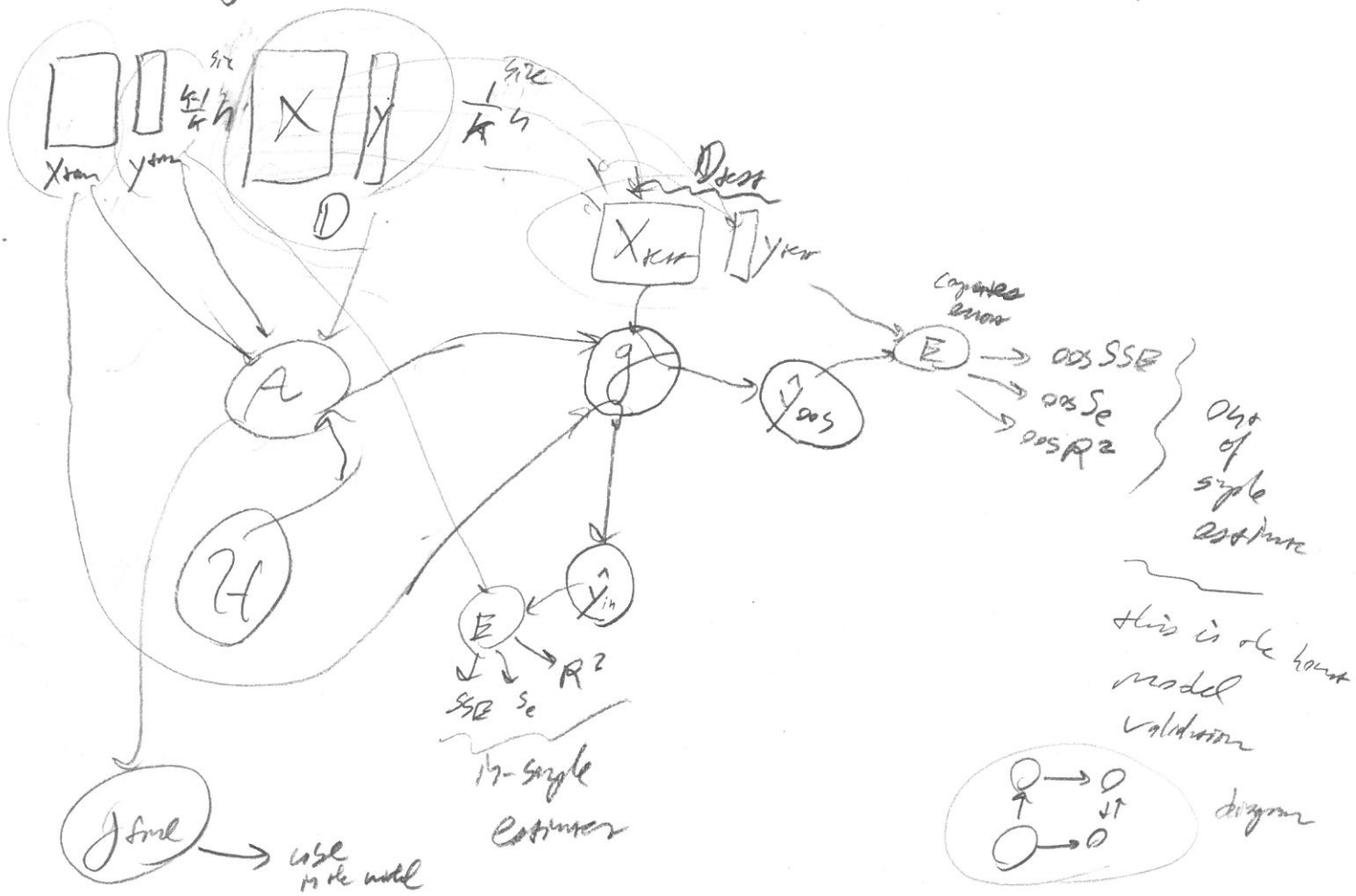
In-sample error metrics are not honest. How do we get Validation / accountability?

Need new data where the y_i's are known so we can compare to $\hat{y}_i$'s from g. ($\notin \mathbb{D}$ never seen by $\mathcal{A}$ ⇒ cannot have been incorporated into model g) How do we do this if $\mathbb{D}$ is all we got ???

Split $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}$ randomly.

Typical splits are 90% train & 10% test ($K=10$) or 80% train & 20% test ($K=5$).

Use $g = \mathcal{A}(\mathcal{D}_{train}, \mathcal{H})$ and use $\mathcal{D}_{test}$ to predict OOS.



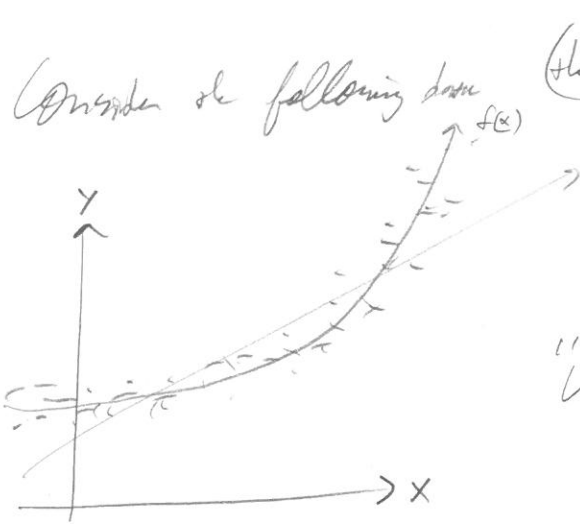
In order for the OOS error estimates to be honest $\mathcal{D}_{test}$ can not be looked at until g is fully constructed. And it can be good once to get OOS estimates and can never be looked at again.

It's like passing to the Internet... once passed... it is forever private again.

For the final shipped model, use all $\mathcal{D}$. This final model is not validated directly but is assumed to be more accurate than g since now it has less estimation error.

DEMO

Consider the following data (Solution to the problem)



$h^*(x)$ if $\mathcal{H} = \{w_0 + w_1 x\}$

2 d.o.f.

"Underfitting" → not enough d.o.f.

However, there is misspecification error. We know (because we can see) $f(x)$ — this is possible only in one dimension. As a reason why does $y \neq f(x)$? $\Rightarrow$ ignorance error.

How do we allow $f(x)$ to be $\in \mathcal{H}$? Make $\mathcal{H}$ "richer". How?

We need to add more terms beyond the linear terms.

Consider the Weierstrass Approx. Thm:

For every cont. function x in region $\in \mathcal{X} = [a,b]$, $\exists$ a polynomial function $p(x)$ s.t. $\forall \varepsilon \; \forall x \in \mathcal{X} \; |f(x) - p(x)| < \varepsilon$.

This means any cont. function can be approximated with a polynomial.

$\Rightarrow \mathcal{H} = \{w_0 + w_1 x + w_2 x^2 : \vec{w} \in \mathbb{R}\}$

Let's begin with polynomials of degree 2. In the above example $p = 1$

Since

$$X = \begin{bmatrix} x_{11} \\ \vdots \\ x_{1n} \end{bmatrix} \quad p = 1$$

add intercept d.o.f.

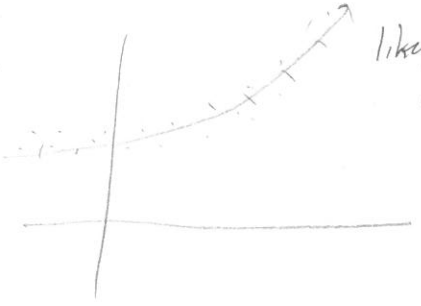
$$\Rightarrow \quad X = \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \quad p+1 = 2 \;\&\; \text{full rank}$$

add poly degree two d.o.f.

$$\Rightarrow X = \begin{bmatrix} 1 & x_{11} & x_{11}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{1n}^2 \end{bmatrix} \quad 2p+1 = 3 \;\&\; \text{full rank} \rightarrow \text{why?}$$

Now use LS as usual! This is why I call this a "linear non-linear model" (asymptotically).

$$\vec{b} = (X^T X)^{-1} X^T \vec{y}$$



likely $h^*(x) \neq f(x)$ or much closer than before!

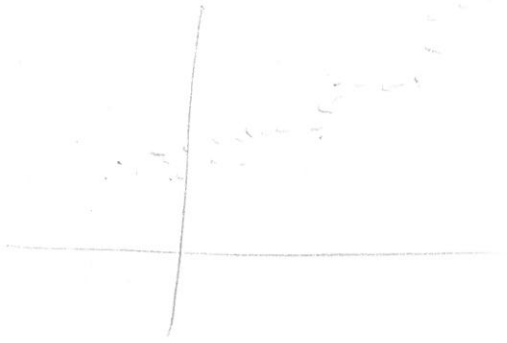
Can we get better?

Why not approx of better by using degree 3?

$$X = \begin{bmatrix} 1 & x_{11} & x_{11}^2 & x_{11}^3 \\ \vdots & & & \\ 1 & x_{1n} & x_{1n}^2 & x_{1n}^3 \end{bmatrix}$$

What would happen?

DEMO



⇒ If n large, not much ... the x^3 predictor would function as one nonsense predictor ... not so bad.

What about many more polynomial terms? ⇒ BAD

We just saw $n = 5$ and a 4-degree polynomial fit it. Ideas?

Vandermonde matrix

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & & & & \vdots \\ 1 & x_5 & x_5^2 & x_5^3 & x_5^4 \end{bmatrix}}_{X} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{bmatrix}$$

Is this invertible? $\det(X) = \prod_{i=1}^{n} \prod_{j=i+1}^{n} x_j - x_i \neq 0$ only if all x_i's unique

Yes n data pts can be fit by a n-1 degree polynomial.

Instead of polynomials, could you make logs? Sins/cosines? Exponentials? Others? Yes! The sky is the limit! Hopefully we can talk more about alternatives to polynomials...