Lec 16 Math 390.4

Polynomial regression was a type of "non-linear" linear modeling.

Here is another "non-linear" linear modeling: interactions.

Consider the case with $p = 2$. $X = \begin{pmatrix} 1 & 1 \\ x_1 & x_2 \\ \vdots & \vdots \end{pmatrix}$

..... hypothesis set

$$\mathcal{H} = \{ w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 : \vec{w} \in \mathbb{R}^4 \}$$

How to fit?

$$\Rightarrow X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n}x_{2n} \end{bmatrix}$$

dof. = 4

Why would you do this? Well, remember

$$\mathcal{H} = \{ w_0 + w_1 x_1 + w_2 x_2 \} \xrightarrow{A} g(\vec{x}) = b_0 + b_1 x_1 + b_2 x_2$$

Interp: if x_1 increases by one unit, $\hat{y}$ changes by b_1 units on average regardless of the value of x_2

$$\mathcal{H} = \{ w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 \} \xrightarrow{A} g(\vec{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$
$$= b_0 + x_1 (b_1 + b_3 x_2) + b_2 x_2$$

Interp: if x_1 ....., $\hat{y}$ .... $b_1 + b_3 x_2$ .....

Now $\Rightarrow$ You've given the model d.o.f. to have different slopes based on the value of the other predictors $\Rightarrow$ much more expressive!

If x_2 e.g. is binary $X_2 = \text{Is Male}$

the slope of x_1 is b_1 if female & $b_1 + b_3$ if male.

DEMO

Nar problem...

given the same X, there are many $\mathcal{H}$'s (and A's) to produce ~~diff~~ g's.

e.g.

$g_1 = b_0 + b_1 x_1$

$g_2 = b_0 + b_1 x_1 + b_2 x_2$

$g_3 = b_0 + b_1 \log(x_1) + b_2 x_2$

$g_4 = b_0 + b_1 x_1^2 + b_2 x_2$

$g_5 = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$

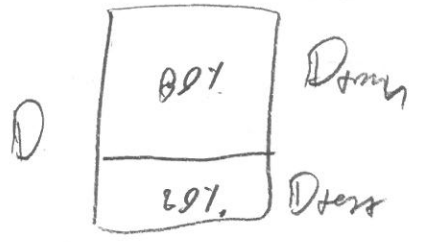
g_M ⋮

infinite models!!

How to "choose" which one to apply?

Model selection - one of the most fundamental questions in all of statistics (maybe even science - is $\tau = \frac{E}{m}$ the "best" model?)

Ideas? Pennsylvania



Purpose? Use Dtest to estimate OOS error (a consistent [honest estimate of future performance])

eg? $D_{train} \subset D$ if you call $D \Rightarrow$ know inside

Why not test $g_1, \ldots, g_M$ by fitting them on D_{train}, testing on D_{test} and picking the one with the least error?

Yes we can! But there is a problem!

What do we use now to estimate future prediction?

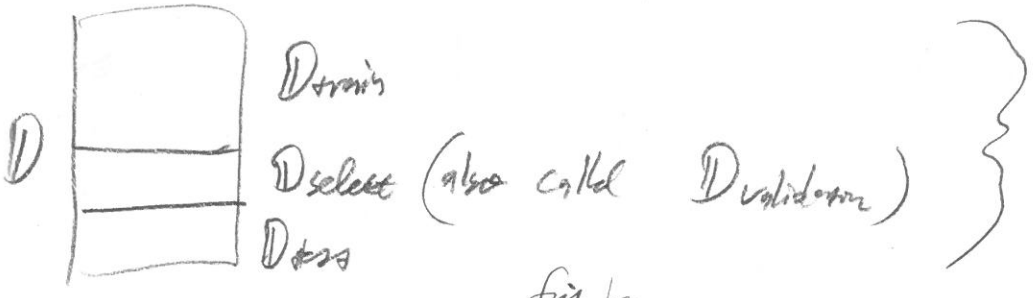
Why can we use $\mathcal{D}_{test}$ again?

Imagine

$g_1, \ldots, g_M$ where M is large.

I can find g_j s.t. error on $\mathcal{D}_{test}$ is near 0 by coincidence! This is not a reliable estimate of future performance!

Solution?



Randomly create 3 splits!

$K=5$ typical

60% train

20% select

20% test

Protocol:

first by defining $\mathcal{A}$ and $\mathcal{H}$.

3) for each model $j \in \{1 \ldots M\}$

build it $g_j = \mathcal{A}_j(\mathcal{H}_j, \mathcal{D}_{train})$

4) get error $oose_j = error(y_{select}, g_j(X_{select}))$ (usually se)

Steps 1 & 2 are like a given $\mathcal{A}_{combined}$

5) Select $j^* := \arg\min \{oose_1, \ldots, oose_M\}$

6) Compute $\underbrace{oose_{j^*}}_{\text{estimate of future performance}} = error(y_{test}, g_{j^*}(X_{test}))$ ✓ test results of $\mathcal{A}_{combined}$

7) Ship model built with steps 1-4 on full dataset!

This is called model selection by OOS error.

This is one such method. Other methods are "analytical" i.e. they rely on statistic/prob. models - we don't talk about those! AIC, BIC, C_p, etc....

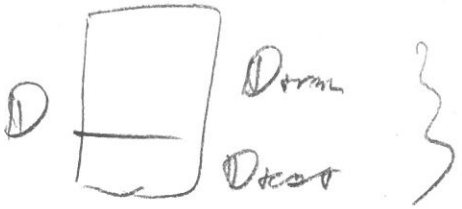
probably won't get to it this semester unfortunately.

DEMO

Midterm 2

FINAL

There is an extra step in validation that is usually employed. Remember before vacation we were testing out one model



} random 80%/20% split of n observations

what if you get an "unlucky" split?

e.g. all the "weird" observations in Dtest

⇒ your error estimate is much higher than it should be

all the "weird" ,,,,,, Dtrain

⇒ your error is much lower ,,,,,,,,

⇒ Variance in the estimate of future performance