Lec 17 4/11/18 Math 390.4

This is called model selection by oos error.

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

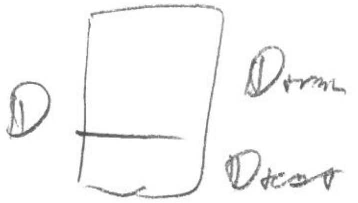
probably won't get to it this semester unfortunately.

DEMO

↑↑ Midterm 2

↓↓ FINAL

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



} random 80% / 20% split of n observations

what if you get an "unlucky" split?

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

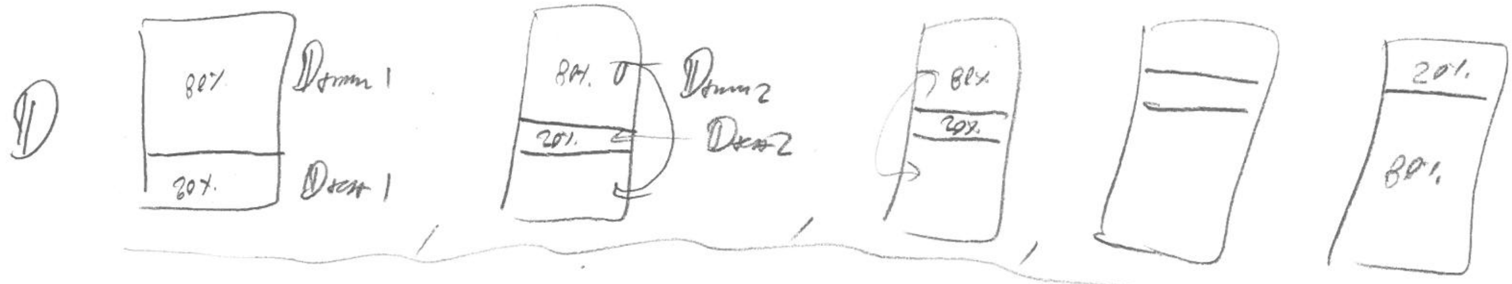
⇒ your error estimate is much higher than it should be

all the "weird" " " " " Dtrain

⇒ your error is much lower " " " " " " "

⇒ Variance in the estimate of future performance

Solution? "K-fold Cross Validation" (CV)



Now, each observation is in the $\mathbb{D}_{test}$ once. $K=5$

Protocol:

① fit $g_K = \mathcal{A}(\mathcal{H}, \mathbb{D}_{train,K})$

② Save $\vec{\hat{y}}_K = g_K(X_{test,K})$

③ Repeat 1-2 for each fold

③ Concatenate vertically $\vec{\hat{y}}_{CV} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_K \end{bmatrix}$

Still: maybe those 5 splits were idiosyncratic

$\Rightarrow$ repeated K-fold CV.
do the whole procedure many times and avg.

3) Compute $oose = error(\vec{y}, \vec{\hat{y}}_{CV})$

full $\vec{y}$ since each obs. represented across the K folds

Not dependent on a split now

EC)

CV gives you a lower variance estimate of future performance.

How ~~to~~ choose K? Let's see what happens with different values of K

No Theory exists to pick opt. value $\Rightarrow$ Open problem in Stat!!

$K=2$ 50% train, 50% test

Since I have 50% test ... I get a very nice estimate of error based on a model which has more est. error! than the final model

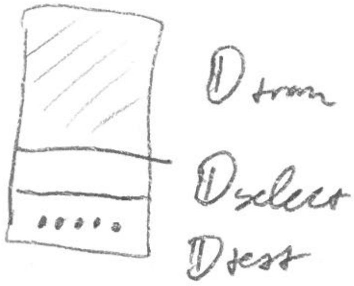
$K=5$ 80% train, 20% test

$K=n$ "leave one out CV" (LOOCV)

Totally random est. of error but a model with low est. error

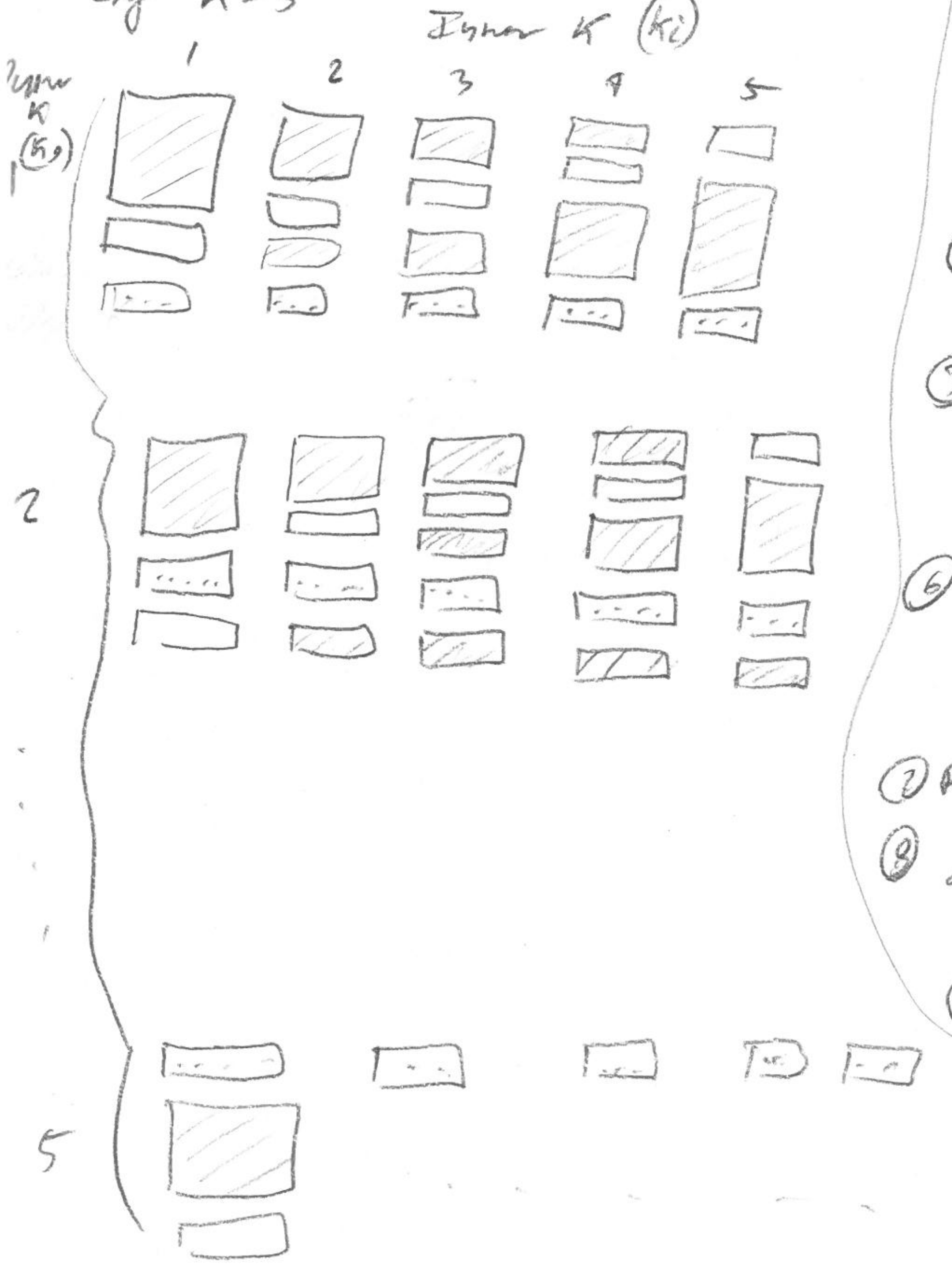
Balance between est. error of g est. error of oose

CV for model selection



There are two CV loops now!

e.g. $K=5$



Bootstrap Validation

sample n rows from $\mathbb{D}$ with replacement ... validate on "oob"

"bag"

For HW

Protocol

```
~~① for each outer loop: $(k_o)$~~
~~② for each each model $j \in \{1, \ldots, M\}$~~
~~③ for each inner loop $(k_i)$~~
    ① fit $g_{j,k_i,k_o} = \mathcal{A}_j(\mathcal{H}_j, \mathbb{D}_{train,k_i,k_o})$
    ② compute $\vec{\hat{y}}_{j,k_i,k_o} = g_{j,k_i,k_o}(\mathbb{D}_{select,k_i,k_o})$
③ Repeat 1-2 for all
   models $j \in \{1, \ldots, M\}$
④ Repeat 1-3 for all
   inner folds $k_i \in \{1, \ldots, 5\}$
⑤ Concatenate
   $\vec{\hat{y}}_{j,k_o} = \begin{bmatrix} \vec{\hat{y}}_{j,1,k_o} \\ \vdots \\ \vec{\hat{y}}_{j,5,k_o} \end{bmatrix}$
⑥ Select best model
   $j^*_{k_o} = \text{argmin}\{oose_{1,k_o}, \ldots oose_{M,k_o}\}$
⑦ Repeat 1-6 for $k_o \in \{1..5\}$
⑧ get $\vec{\hat{y}} = \begin{bmatrix} \hat{y}_{j^*_1,1} \\ \vdots \\ \hat{y}_{j^*_5,5} \end{bmatrix}$
⑨ Estimate $oose = error(\vec{\hat{y}}, \vec{y})$
⑩ Repeat steps 1-6
   to build final model $g$
   without $\mathbb{D}$ test. (only inner CV loop)
```