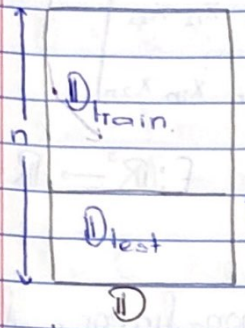
Lecture -16 04/02/2020



$\frac{1}{K} = \frac{n_{test}}{n} \Rightarrow K = \frac{n}{n_{test}}$

$n_{test} \in \{1, 2, \ldots, n/2, \ldots, n/3, \ldots, n/5, \ldots, n/10, \ldots, n-1\}$

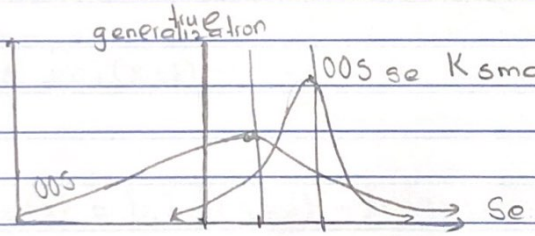
$\Rightarrow K \in \{n, n/2, \ldots, 2, \ldots, 3, \ldots, 5, \ldots, 10, \ldots, \frac{n}{n-1}\}$

What's the tradeoff? most common choices for K

Only a sample of the full distribution $\langle \vec{x}, y \rangle$

We want honest validation, we use $\mathcal{D}_{test}$ to estimate "generalization error" AKA model performance in the future.

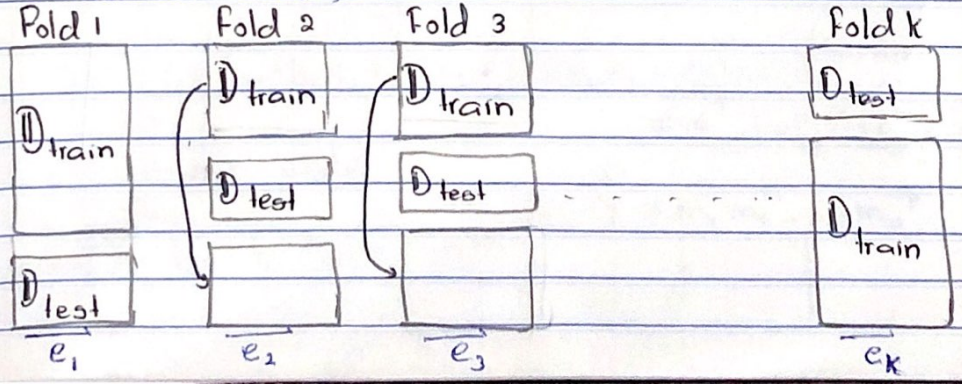
If n_{test} is small, the estimate is very variable but n_{train} is big thus ... my expected estimate is closer to the true performance of g_{final}.



Cross Validation (cv): do many train-test split.

→ K-Fold cv →

→ Another (Enter)



Let $\vec{e}_{CV} = \begin{bmatrix} \vec{e_1} \\ \vec{e_2} \\ \vdots \\ \vec{e_K} \end{bmatrix}$ has length n. Now compute OOS metrics on $\vec{e}_{CV}$

If K=n (n folds) AKA "leave-one-out cross validation (LOOCV)

Consider computing matrices on $\vec{e_1}, \vec{e_2}, \ldots, \vec{e_K} \Rightarrow$ $Se_1, Se_2, \ldots, Se_K$

$\overline{Se} = Avg(Se's)$, $Sse = StdError(Se's)$ $CI_{\theta e, 95\%} = [\overline{Se} \pm 2\, Sse]$

For this to be valid, $Se \overset{iid}{\sim} N(,)$. Is this true? Rnd? NO; K-large $\rightarrow$ normality from CLT.

Because $\vec{e}$'s are dependent. Since they are functions of g's which are dependant via sharing D train's.

Computing a valid CI for generalization error under general continuous is NOT possible.

Given D, you have many choice of model because there are many A's and many H's $g = A(D, H)$. Let's say you have choice of a finite # of models, m. How do we choose between $g_1, g_2, \ldots, g_m$?
e.g. Praw = 1., A = OLS, y = R

$g_1 = b_0 + b_1 X$
$g_2 = b_0 + b_1 X + b_2 X^2$.
$g_3 = b_0 + b_1 \ln(X)$
$g_4 = b_0 + b_1 \mathbb{1}_{X \in [0,1]} + b_2 \mathbb{1}_{X \in [1,2]} + \ldots$
$\vdots$

This is the fundamental problem of "model selection". In this class, we will provide one idea. It is not the only idea.

g_1 is fit with A_1, H_1 on $\mathcal{D}_{train}$ and Se_1 is computed on $\mathcal{D}_{test}$

g_2 is fit with A_2, H_2 on $\mathcal{D}_{train}$ and Se_2 is computed on $\mathcal{D}_{test}$

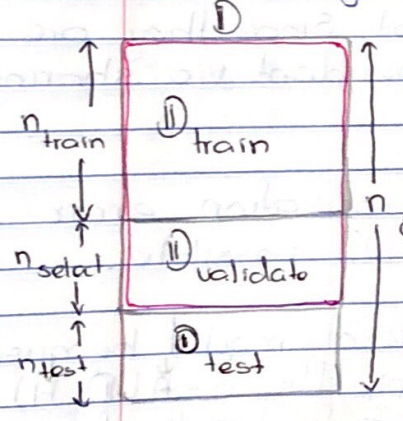
⋮

g_m is fit with A_m, H_m on $\mathcal{D}_{train}$ and Se_m is computed on $\mathcal{D}_{test}$

then select g_{m_*} which has lowest Se.

Problems (1) Se has high variance. So pick K reasonable

(2) We no longer have honest validation. Se is not honest because it could be overfit since $\mathcal{D}$ test was used many times.



Procedure

① Fit g_1 on $\mathcal{D}_{train}$, compute Se_1 on $\mathcal{D}$ select.

② Fit g_2 on $\mathcal{D}_{train}$, compute Se_2 on $\mathcal{D}$ select

⋮

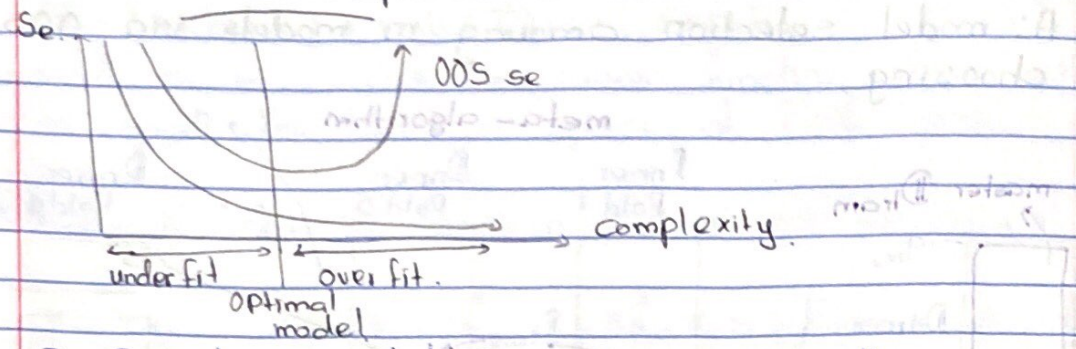
(m) Fit g_m on $\mathcal{D}_{train}$, compute Se_m on $\mathcal{D}$ select

(m+1) Find $m_* = \text{avgmax}\{Se_j\}$

(m+2) Fit g_{m_*} on $\mathcal{D}_{train} \cup \mathcal{D}_{select}$ and compute Se_* on $\mathcal{D}_{test}$

(m+3) Fit g_{final} on $\mathcal{D}$

Uses of this procedure.



① Consider models $g_1, g_2, \ldots g_m$ with increasing complexity. This is sometimes called stepwise modeling.