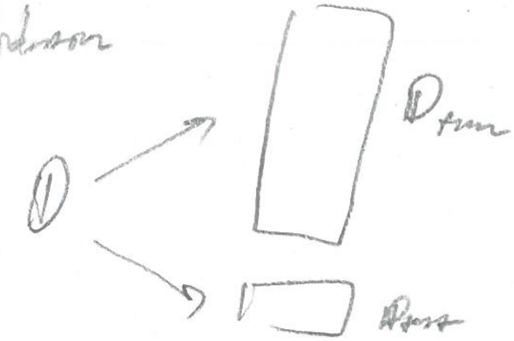
Lecture 17 Math 340.4 4/4/17

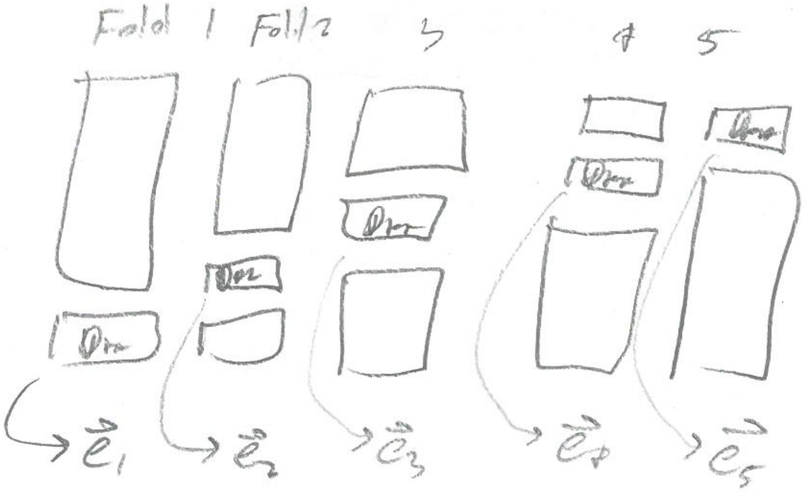
Validation



$K = 5 \Rightarrow \frac{n_{test}}{n} = \frac{1}{K} = 20\%$

If you compute s_e on $\mathbb{D}_{test}$.

$P(s_e \mid \mathbb{D}_{train}, \mathbb{D}_{test})$ meaning if you get lucky or unlucky in your choice of training and/or test split, your results are not accurate. Simple solution: "cross-validation" (CV). Do this procedure K times, moving the $\mathbb{D}_{test}$ through all n (called K "folds")



Set $\vec{e}_{CV} = \begin{bmatrix} \vec{e}_1 \\ \vec{e}_2 \\ \vdots \\ \vec{e}_K \end{bmatrix}$ and then compute error metrics on $\vec{e}_{CV}$

R^2, s_e, etc. . . this is a good est of gen. error of model

Also, s_e will be variable. Use the $K = 5$ estimates of s_e from $s_{e_1} = err(\vec{e}_1)$, $s_{e_2} = err(\vec{e}_2)$, ..., $s_{e_K} = err(\vec{e}_K)$ to obtain a std error of the estimate to see how variable it is. $s_{s_e} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{5}(s_{e_k} - \bar{s}_e)^2}$ Need independence of $s_{e_1}, \ldots, s_{e_K}$. they are "somewhat" $\perp$.

How to pick K? If K small then the error will be biased upwards since you are not using all data to estimate g. If K large, the estimates $s_{e_1}, \ldots, s_{e_K}$ will be very variable as they are not estimated with lots of n_{test}. Also $s_{e_1} \ldots s_{e_K}$ will lose independence.

$\Rightarrow$ Nobody knows the optimal K. 5 or 10 is default.