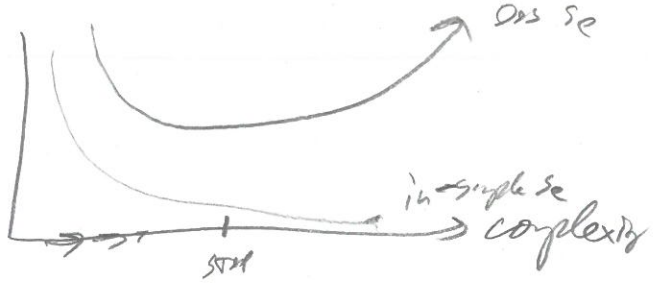
Lec 19 Math 390 4/11/19

What is a common use of this System?

(I)



Use this to choose complexity unit-by-unit until Se in Outer increases.

This is called greedy stepwise model selection. Here's how it works: ① pick a set of p known predictors: $x_1, \ldots, x_p, x_1^2, \ldots, x_p^2$, $\ln(x_1), \ldots, \ln(x_p)$, $x_1 x_2$, $x_1 x_3 \ldots$, $x_{p-1} x_p$, etc.

① For each of the p predictors, try each one by adding it to the model and select the model with maximum reduction in in-sample Se.

② For each of the remaining p-1 predictors, return to step 1. Stop if Se on validation set increases.

③ Using these features, build model on Dtrain ∪ Dvalidation and test on Dtest to get an OOS honest estimate.

(A) ' ' ' ' ' ' build model on D and skip g.

(II) Select Hyperparameters. ✓ do this first

Recall SVM which minimized:

$AHE + \lambda \|\vec{w}\|^2$ How to pick λ?

Hyperparameter selection AKA "tuning" which is same as model selection

① Select grid of λ's. $\lambda_{grid} = \{.00, .001, .01, .03, .1, .3, 1, 3, 10, 33, 100\}$

② Each λ is a diff. A_λ. Fit all models.

③ Select model with smallest error in validation set.

Return ② ↑

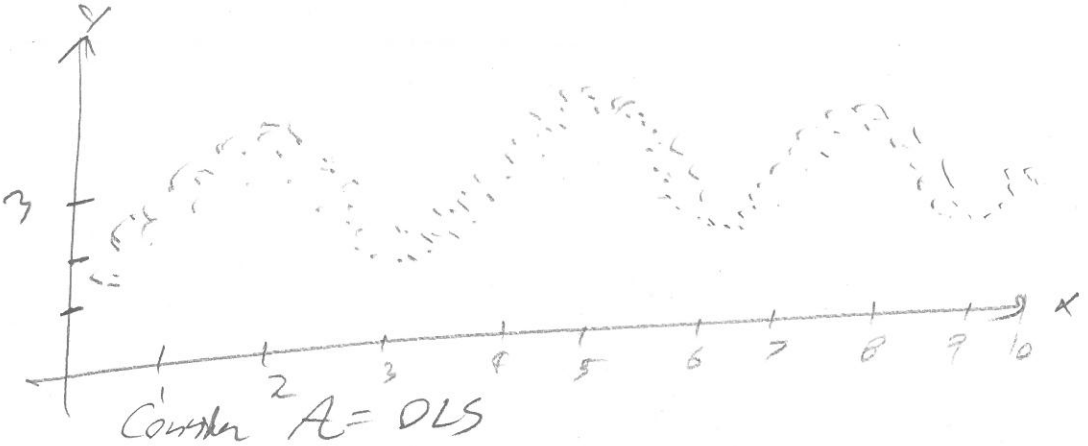
Previously we considered $\{g_1, \ldots, g_m\}$ and used OOS-validation to select a good model, i.e. min. misspec. error. Then did another OOS validation to obtain an honest estimate. But how to pick the models $\{g_1, \ldots, g_m\}$?

How about we create an A that flexibly makes it more complicated? So more $\lambda \to$ more d.f. And this A is immune to overfitting. This is the goal of non-parametric machine learning. We will study tree models, specifically

Classification & Regression Trees (CART, 1984)

$y = \{1, \ldots, L\}$ $\quad$ $y \subseteq \mathbb{R}$

Imagine data as:



Consider $A = OLS$

$$\mathcal{H}_1 = \{ w_0 + w_1 x : \vec{w} \in \mathbb{R}^2 \}$$

$$\mathcal{H}_2 = \{ w_0 + w_1 x + w_2 x_2 ; \vec{w} \in \mathbb{R}^3 \}$$

$$\mathcal{H}_3 = \{ w_0 + w_1 \sin(w_2 x_2) : \vec{w} \in \mathbb{R}^3 \}$$

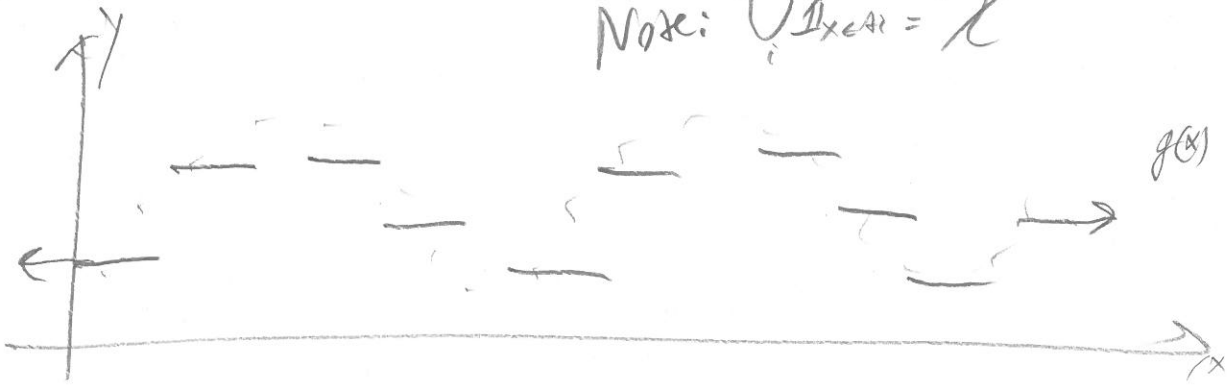
Easy to see $\mathcal{H}_3$ would win! But in high p — not easy!!

Now consider

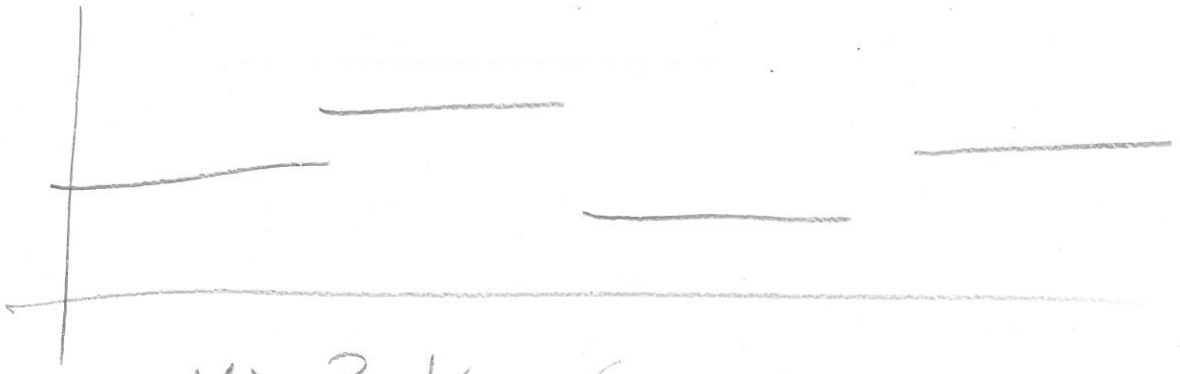
$$\mathcal{H}_4 = \{ w_1 \mathbb{1}_{x \in (\infty, 1)} + w_2 \mathbb{1}_{x \in [1,2)} + \dots + w_{10} \mathbb{1}_{x \in [9, \infty)} ; \vec{w} \in \mathbb{R}^{10} \}$$

the fits $b_1, \dots b_{10}$ is ground truth. The coefficients are just $\bar{y}$ within the bin.

Note: $\bigcup_i \mathbb{1}_{x \in A_i} = \mathcal{X}$



Let's make less than 10 parameters by using large bins



underfitting? Yes (misspecification error)

Max than 10 parameters.

Overfitting? Yes (estimation error) Not enough data in each bin to make a good decision

How to pick bin size? Use the model selection routine just discussed.