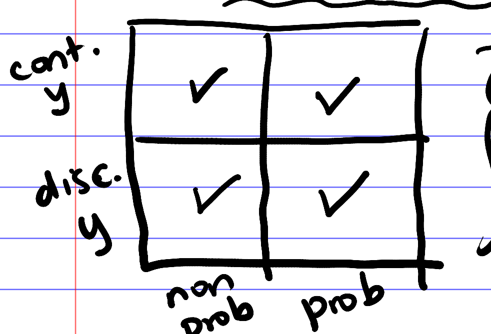
CS181: Model Selection

Book: 2.7, 2.8

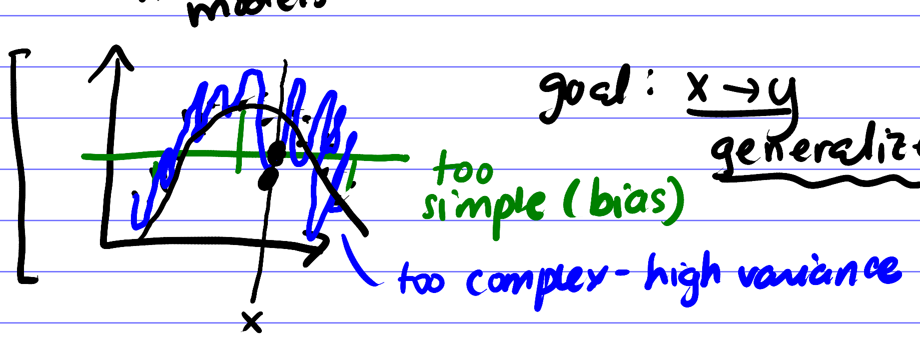


} supervised learning $x \to y$

Now: Selecting models ← TODAY
- expressive models
- loss function choices

Selecting models: 1st HW: sunspots vs. Republicans → different basis choices

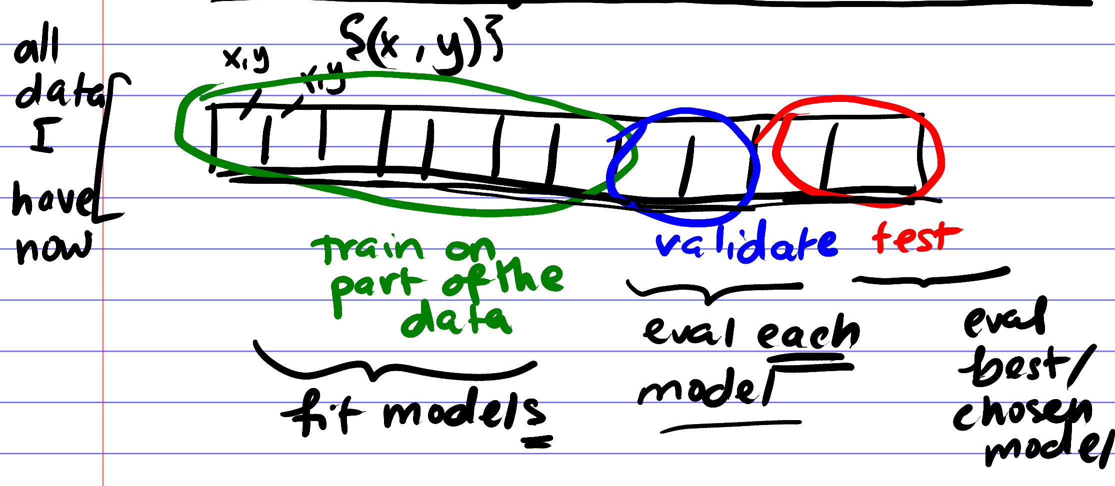
already had multiple models



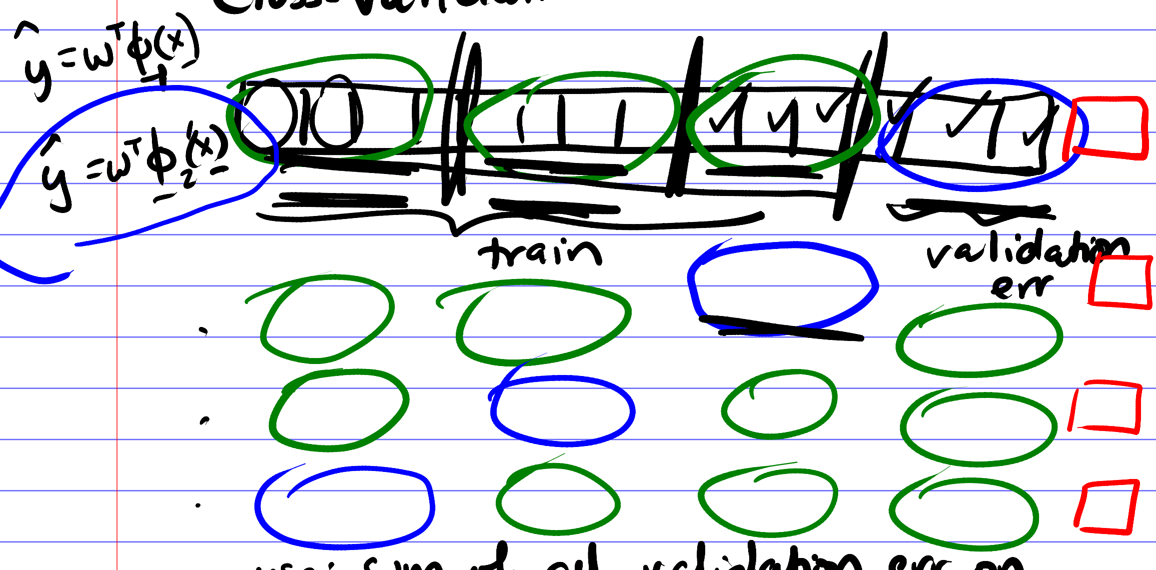
goal: $x \to y$ generalize

How can we get good generalization?
- basic techniques to select models that will generalize
- why models fail to generalize?
- ways to get around bias-variance trade-offs

Basic Strategies for Model Selection

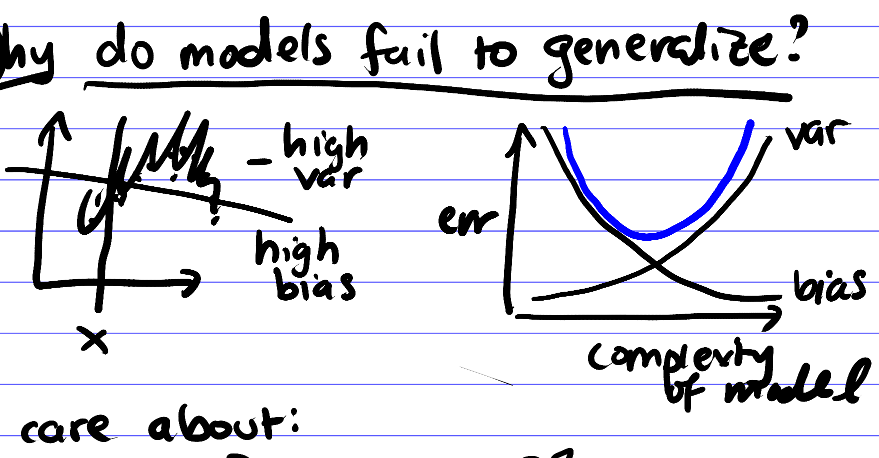


Cross-Validation:



use: sum of all validation err on each model as its score

Why do models fail to generalize?



we care about:

$$\mathbb{E}_{\mathcal{D}}\ \mathbb{E}_{y|x}\left[(y(x) - \hat{y}(x))^2\right]$$

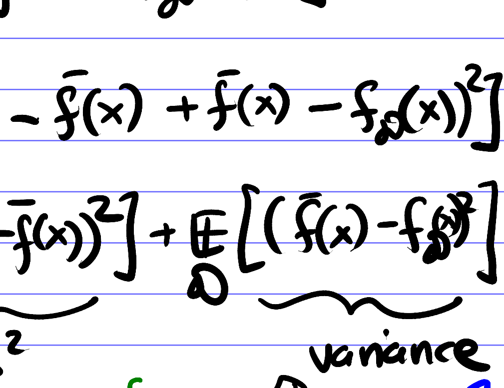
$f_{\mathcal{D}}(x)$

note: there exists some mean value $\bar{y}(x)$ for $y(x)$

$$= \mathbb{E}_{\mathcal{D}}\ \mathbb{E}_{y|x}\left[\left((y(x) - \bar{y}(x)) + (\bar{y}(x) - \hat{y}(x))\right)^2\right]$$

$$\underbrace{\mathbb{E}_{\mathcal{D}}\ \mathbb{E}_{y|x}\left[(y - \bar{y})^2\right]}_{\text{noise}} + \underbrace{\mathbb{E}_{\mathcal{D}}\ \mathbb{E}_{y|x}\left[(\bar{y} - \hat{y}(x))^2\right]}_{\text{err from our modeling}} + 2\,\mathbb{E}_{y|x}\,\mathbb{E}_{\mathcal{D}}\left[(\bar{y} - \hat{y})(y - \bar{y})\right]$$

0, 0



$$\mathbb{E}_{\mathcal{D}}\ \mathbb{E}_{y|x}\left[(\bar{y} - \hat{y}_{\mathcal{D}}(x))^2\right] \Rightarrow \mathbb{E}_{\mathcal{D}}\left[(\bar{y} - \hat{y}_{\mathcal{D}}(x))^2\right]$$

how will this model err term behave over multiple tries/multiple datasets?

$$\mathbb{E}_{\mathcal{D}}\left[(\bar{y}(x) - \hat{y}_{\mathcal{D}}(x))^2\right]$$

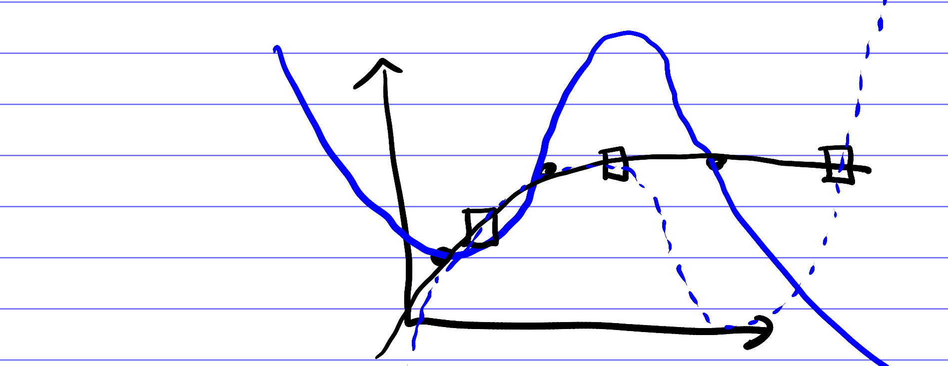
$$= \mathbb{E}_{\mathcal{D}}\left[(\bar{y}(x) - f_{\mathcal{D}}(x))^2\right]$$

let $\bar{f}(x)$ be $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)]$

$$= \mathbb{E}_{\mathcal{D}}\left[(\bar{y}(x) - \bar{f}(x) + \bar{f}(x) - f_{\mathcal{D}}(x))^2\right]$$

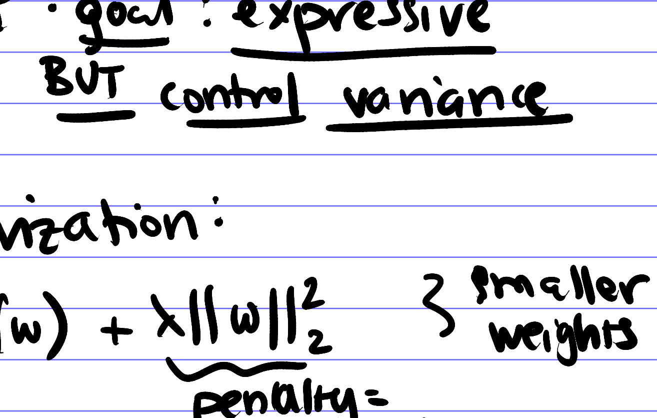
$$= \underbrace{\mathbb{E}_{\mathcal{D}}\left[(\bar{y}(x) - \bar{f}(x))^2\right]}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[(\bar{f}(x) - f_{\mathcal{D}})^2\right]}_{\text{variance}} + \mathbb{E}_{\mathcal{D}}\left[(\bar{y} - \bar{f})(\bar{f} - f_{\mathcal{D}})\right]$$

0



End up with:

$$\text{err} = \underbrace{\text{noise}}_{\text{not in our control}} + \underbrace{\text{bias}^2 + \text{variance}}_{\text{in our control}}$$



Ways to mitigate bias-var trade-off: goal: expressive models BUT control variance

1. Regularization:

$$\min_w \mathcal{L}(w) + \underbrace{\lambda\|w\|_2^2}_{\text{penalty = regularization}} \quad \} \text{ smaller weights}$$

$$+\lambda\|w\|_1 \quad \} \text{ sparse weights}$$

$$+R(w)$$

2. Ensembles: averaging over multiple models. [bagging, boosting, random forests, grad. boosted trees]