Lecture 7: Model Evaluation

Need to be precise about what is random.

Consider training data $(\underset{\sim}{X}_n, Y_n) \overset{iid}{\sim} p$

So $T = \{(x_n, y_n)\}_{n=1}^{N}$

is random

Use this to fit

$$\hat{f} = \hat{f}_T$$

random

Let $\underset{\sim}{X}_0, Y_0$ be an independent (from training data) sample from p.

and let M be my perf. metric

e.g. $M(A, B) = (A-B)^2$

then let

$$Err_T = \mathbb{E}\left[M(Y_0, \hat{f}_T(\underset{\sim}{X}_0)) \mid T\right]$$

probably what we want to estimate

= given some T expected err. on new data

However, in practice often difficult to estimate.

Instead, easier to estimate

$$\text{Err} = E_T[\text{Err}_T]$$
$$= \mathbb{E}[M(Y_0, \hat{f}(\underset{\sim}{X}_0))]$$

= expected gen. perf. across all possible T

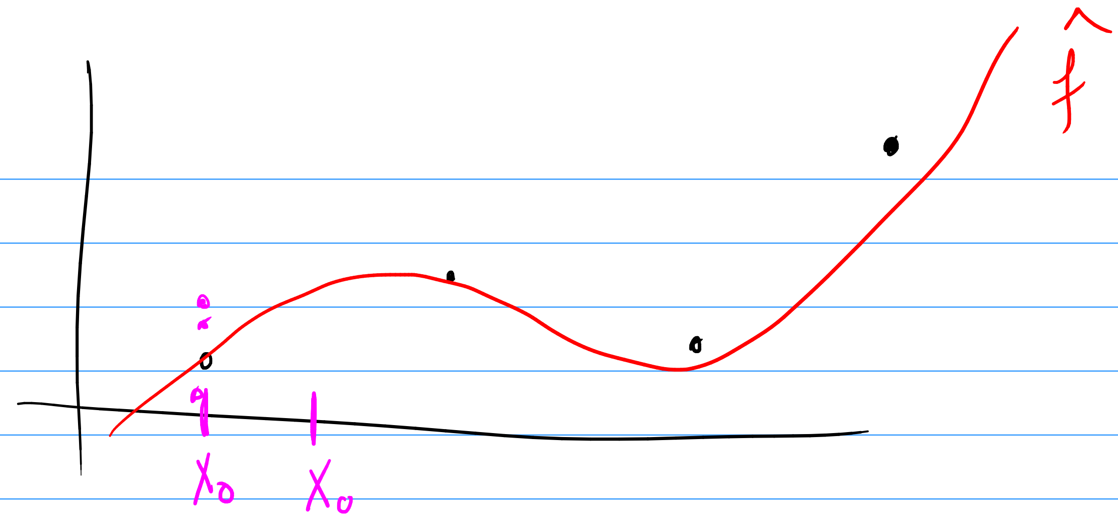
Also define

$$\overline{\text{err}} = \text{training error}$$
$$= \frac{1}{N} \sum_n M(y_n, \hat{f}(\underset{\sim}{x}_n))$$

Typically $\underbrace{\overline{\text{err}} < \text{Err}_T}_{\text{both eval. for fixed } T}$ b/c we over-fit

Part of the problem:

1. $\underset{\sim}{X}_0$ is different than training $\underset{\sim}{x}_n$
2. Y_0 is different than training y_n (even if $\underset{\sim}{X}_0 = \underset{\sim}{x}_n$)



Partially simplify by considering only Y_0 to be random.

In-sample error: $\mathrm{Err}_{in} = \frac{1}{N} \sum_n \mathbb{E}_{Y_0}\left[M(Y_{0n}, \hat{f}(\underset{\sim}{x}_n))\right]$

fix $\underset{\sim}{X}_0$s at training $\underset{\sim}{X}_n$s and calc. gen. err.

– like Err_T but fixing $\underset{\sim}{X}_0$s at training.

Define optimism

$$op = \mathrm{Err}_{in} - \overline{err}$$

in-sample – training.

typically $op > 0$ as $\overline{err}$ underestimates Err_{in}

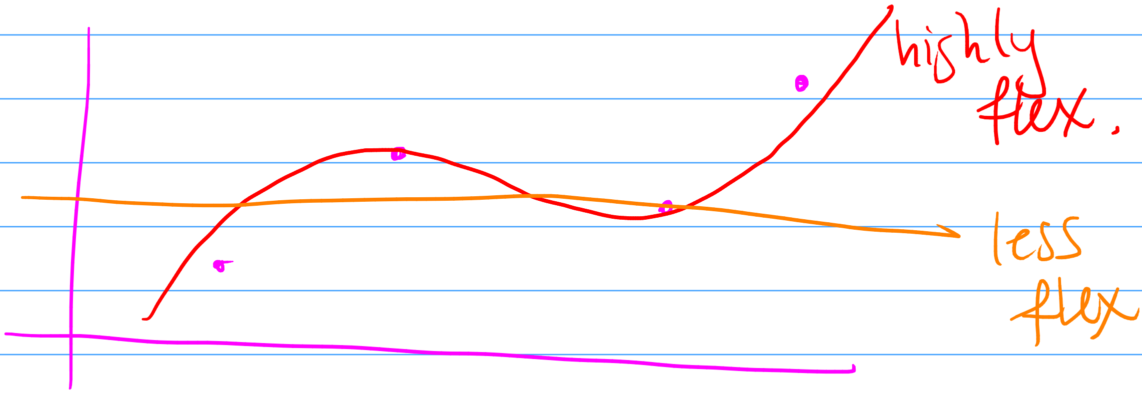
can calculate expected op over potential training data ($\underset{\sim}{X}_n$s)

$$\omega = \mathbb{E}[op]$$

In general

$$\omega = \frac{2}{N} \sum_n \operatorname{cov}(\hat{Y}_n, Y_n)$$

So ω the avg. amount that Y_n affects $\hat{Y}_n$



Notice

$$E[\mathrm{Err}_{in}] = E[\overline{\mathrm{err}}] + \omega$$

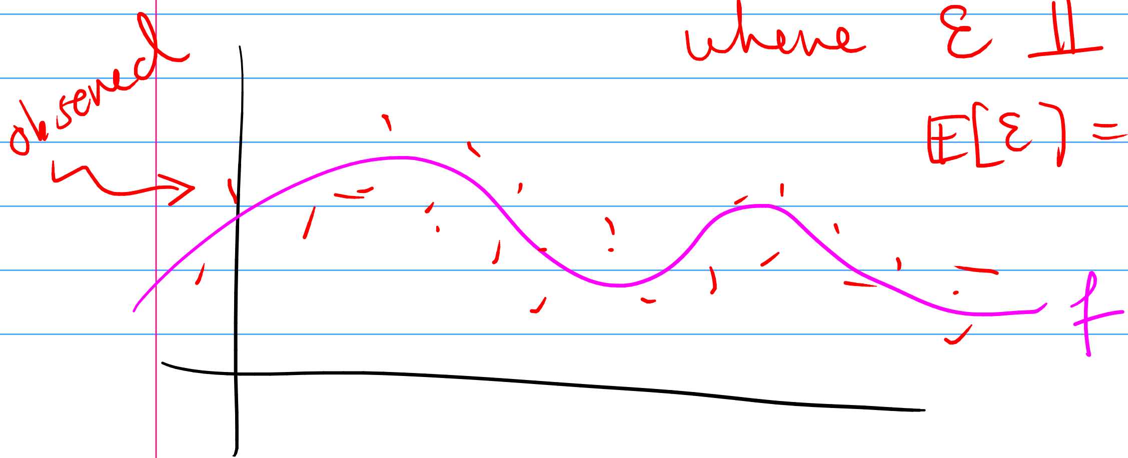
in some cases I can estimate ω as $\hat{\omega}$ and then est. my Err_{in}

$$\mathrm{Err}_{in} \approx \overline{\mathrm{err}} + \hat{\omega}$$

Ex. Assume $Y = f(\underline{X}) + \varepsilon$

where $\varepsilon \perp\!\!\!\perp \underline{X}, Y$

$E[\varepsilon] = 0, \operatorname{Var}(\varepsilon) = \sigma^2$



Then if I fit a lin. regression w/ P variables, then

$$\omega = \frac{2P}{N}\sigma^2.$$

In practice what we can do is estimate σ^2 as $\hat{\sigma}^2$ and then calc.

$$\widehat{Err}_{in} = \overline{err} + \frac{2P}{N}\hat{\sigma}^2.$$

mallow's C_p
generally AIC.

Tuning model complexity

(1) CV estimates Err
(2) estimate Err_{in}

neither estimating Err_T

Downside of CV is that it trains on a slightly smaller data set.

Do a K-fold X-val.

Can choose K. Can set K = N.

Called Leave-one-out X-val (LOO).

<u>Bias-Variance Tradeoff.</u>

Consider a model

$$Y = f(\underset{\sim}{X}) + \varepsilon$$

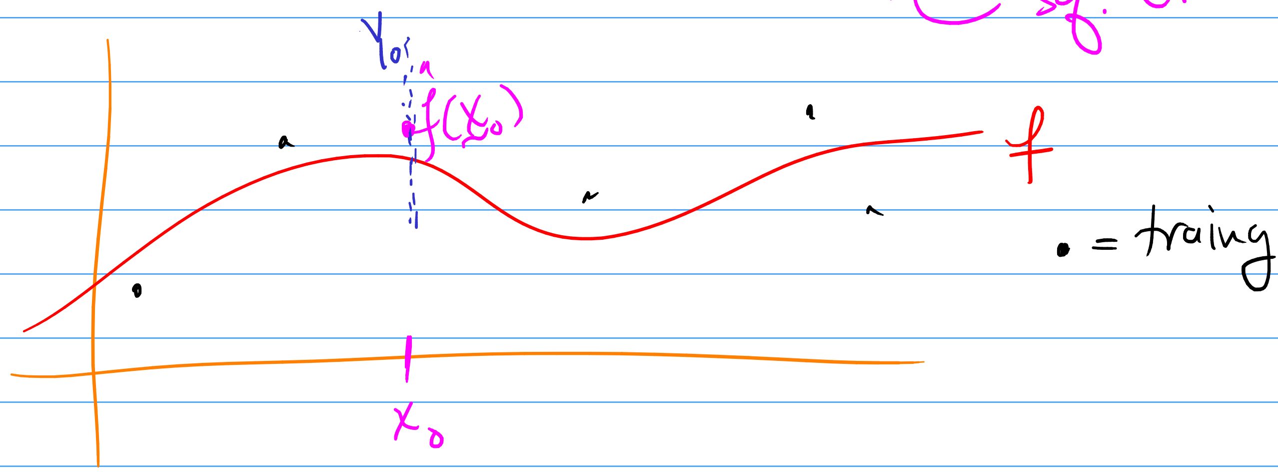
$$\varepsilon \perp\!\!\!\perp \underset{\sim}{X}, Y$$

$$\mathbb{E}\varepsilon = 0, \quad \mathrm{Var}\,\varepsilon = \sigma^2 < \infty.$$

Fix $\underset{\sim}{X}_0$ at some fixed $\underset{\sim}{x}_0 \in \mathbb{R}$ and calc

$$\mathrm{Err}(\underset{\sim}{x}_0) = \mathbb{E}\left[(Y_0 - \hat{f}(\underset{\sim}{x}_0))^2\right] = (*)$$

sq. err.



can decompose:

$$(*) = \mathbb{E}\left[\left(\underbrace{Y_0 - \overbrace{\mathbb{E}[\hat{f}(x_0)]}^{\text{expected over } T}}_{a} + \underbrace{\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)}_{b}\right)^2\right]$$

$$(a+b)^2 = a^2 + b^2 + 2ab$$

$$= \underbrace{\mathbb{E}\left[\left(Y_0 - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]}_{a^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right]}_{b^2}$$

$$\underbrace{+ 2\,\mathbb{E}\left[\left(Y_0 - \mathbb{E}[\hat{f}(x_0)]\right)\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\right]}_{+2ab}$$

Claim: $\mathbb{E}[a^2] = \text{Bias}(\hat{f})^2 + \sigma^2$

$\mathbb{E}[b^2] = \text{Var}(\hat{f})$

$\mathbb{E}[2ab] = 0$

i.e. $\text{Err}(x_0) = \text{Bias}^2 + \text{Var} + \underbrace{\sigma^2}_{\text{irreducible noise}}$

$\text{Bias} = \mathbb{E}[\hat{f}(x_0) - f(x_0)]$ = expected diff btwn est. and truth

$\text{Var} = \text{Var}(\hat{f}(x_0))$ = amt my $\hat{f}$ changes due to changing T

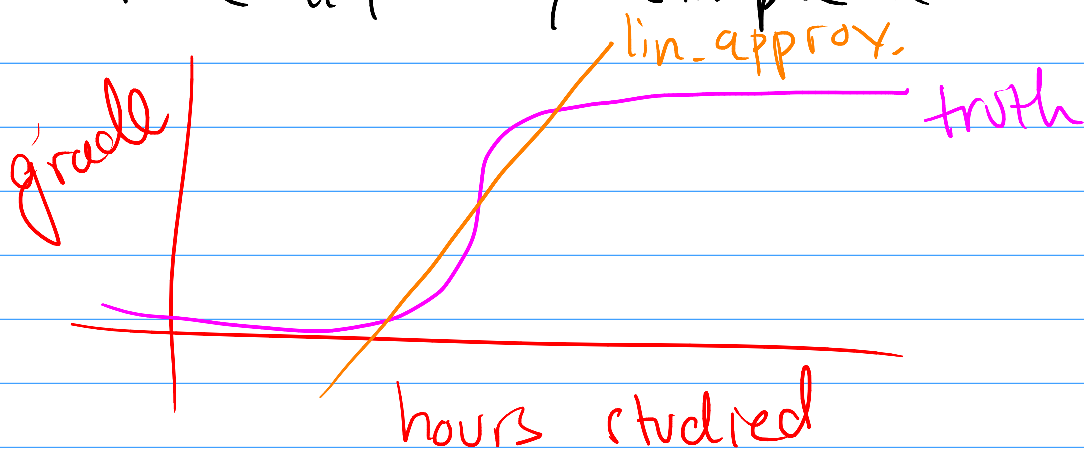
Typically:

low flex $\Leftrightarrow$ low var, high bias

high flex $\Leftrightarrow$ high var, low bias

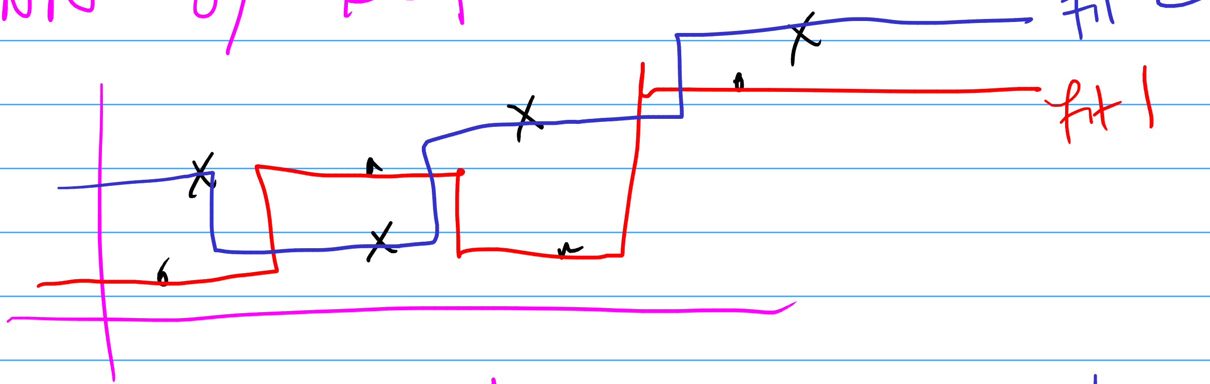
In laymens terms

bias = err incurred b/c approx. complicated real life w/ simple model



variance = sensitivity of fit

e.g. KNN w/ $k=1$



KNN w/ $k=N$

not sens. to T