Lecture 8

Bias - Variance Decomp

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

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

and $\mathbb{E}\varepsilon = 0$

$\mathrm{Var}\,\varepsilon = \sigma^2$

← new pt – not traing

Fix $\underset{\sim}{X}_0$ at $\underset{\sim}{x}_0$ and look at

sq. err.

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

we can decompose this

$$= \mathbb{E}\Big[\big(\underbrace{Y_0 - \mathbb{E}[\hat{f}(\underset{\sim}{X}_0)]}_{a} + \underbrace{\mathbb{E}[\hat{f}(\underset{\sim}{X}_0)] - \hat{f}(\underset{\sim}{X}_0)}_{b}\big)^2\Big]$$

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

$$= \mathbb{E}\left[\left(Y_0 - \mathbb{E}[\hat{f}(\underset{\sim}{X}_0)]\right)^2\right] + \mathbb{E}\left[\left(\hat{f}(\underset{\sim}{X}_0) - \mathbb{E}[\hat{f}(\underset{\sim}{X}_0)]\right)^2\right]$$

1 2

$$+ 2\mathbb{E}\left[\left(Y_0 - \mathbb{E}\hat{f}(\underset{\sim}{X}_0)\right)\left(\hat{f} - \mathbb{E}\hat{f}(\underset{\sim}{X}_0)\right)\right]$$

3

1

$$Y_0 = f(\underset{\sim}{X}_0) + \varepsilon_0$$

constant

$$1 = \mathbb{E}\left[\left(f(\underset{\sim}{X}_0) + \varepsilon_0 - \mathbb{E}\hat{f}(\underset{\sim}{X}_0)\right)^2\right]$$

$$= \cancel{\mathbb{E}}\left[\left(f(\underset{\sim}{X}_0) - \mathbb{E}\hat{f}(\underset{\sim}{X}_0)\right)^2\right] + \mathbb{E}\varepsilon_0^2 + 2\mathbb{E}\left[\varepsilon_0\left(f(X_0) - \mathbb{E}\hat{f}(X_0)\right)\right]$$

$$\underbrace{\left(f(\underset{\sim}{x}_0)-\mathbb{E}\hat{f}(\underset{\sim}{x}_0)\right)}_{\text{Bias}(\hat{f})}{}^2 \qquad \sigma^2 \qquad +2\left(f(x_0)-\mathbb{E}\hat{f}(x_0)\right)\underbrace{\mathbb{E}[\varepsilon]}_{0}$$

$\text{Bias}(\hat{f})$ = diff btwn truth (f) and avg. est ($\mathbb{E}\hat{f}$)

$$① = \text{Bias}(\hat{f})^2 + \sigma^2$$

$$② \quad \mathbb{E}\left[\left(\hat{f}(\underset{\sim}{x}_0)-\mathbb{E}[\hat{f}(\underset{\sim}{x}_0)]\right)^2\right] = \text{Var}(\hat{f})$$

$$\text{Var}(z) = \mathbb{E}\left[(z-\mathbb{E}z)^2\right]$$

$$③ = 0$$

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

indep

can distribute b/c indep

$$= \mathbb{E}\left[Y_0-\mathbb{E}\hat{f}(\underset{\sim}{x}_0)\right]\underbrace{\mathbb{E}\left[\hat{f}(x_0)-\mathbb{E}\hat{f}(\underset{\sim}{x}_0)\right]}_{0}$$

$$\mathbb{E}\left[z-\mathbb{E}z\right] = \mathbb{E}z-\mathbb{E}z = 0$$

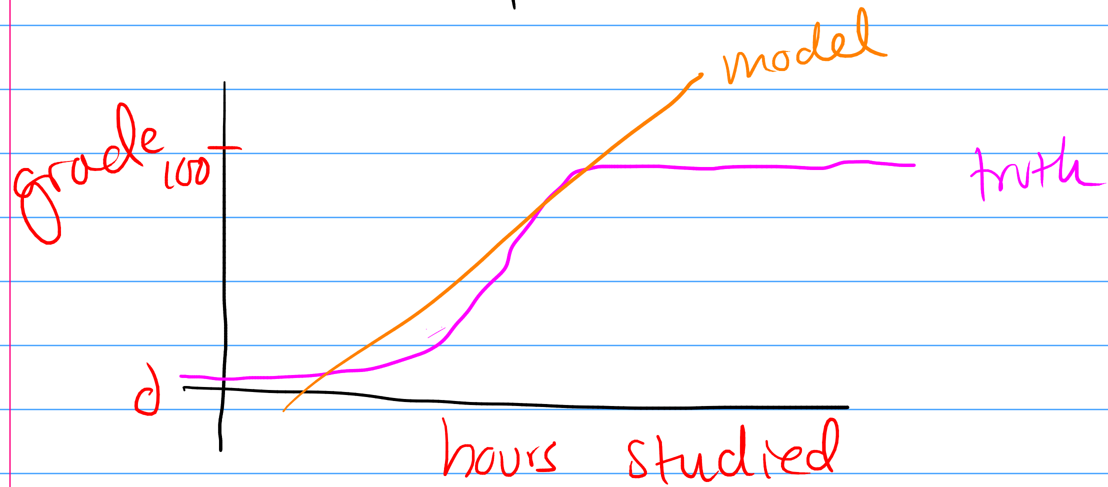
<u>All together:</u>

$$\text{Err}(\underset{\sim}{x}_0) = \text{Bias}\left(\hat{f}(\underset{\sim}{x}_0)\right)^2 + \text{Var}\left(\hat{f}(\underset{\sim}{x}_0)\right) + \sigma^2$$

irreducible var. term

Laymen's terms

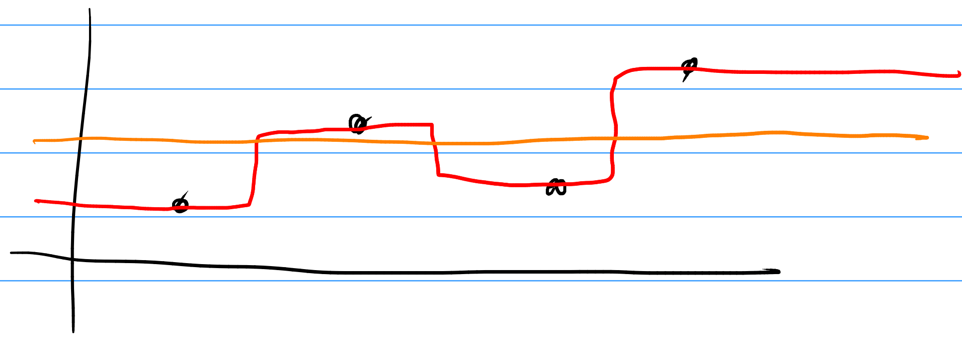
Bias: err. b/c we approx. a complicated real life fn w/ a simpler model



Variance of $\hat{f}$: how much my model changes when the traing data changes

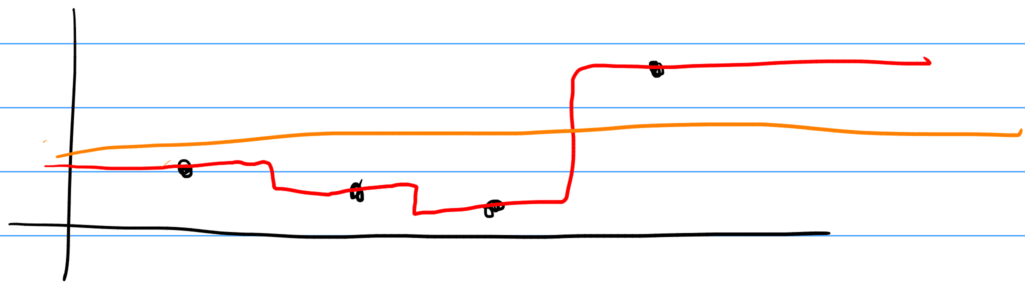
e.g. 1-NN

N-NN = kNN w/ k = #training



high variance

b/c my model changes a lot when my data does

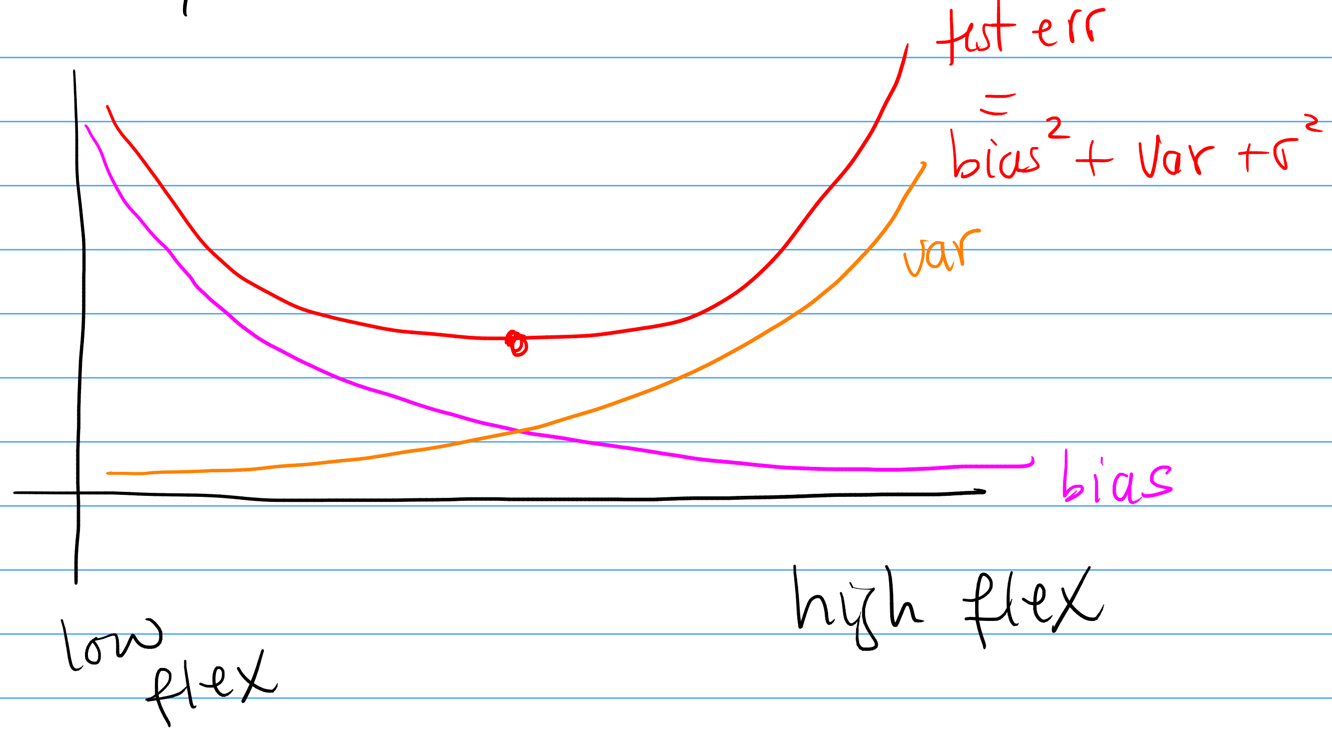


Typically:

low flex $\iff$ low var, high bias

high flex $\iff$ high var, low bias

Recall test/train err curve



Risk Minimization

Theoretically want

$$f^* = \underset{f}{\arg\min}\ \overbrace{\mathbb{E}[L(Y, f(\underset{\sim}{X}))]}^{\text{Risk}}$$

Theoretical minimum?

Let $(\underset{\sim}{X}, Y) \sim p$ ← joint density

$$\mathbb{E}[L(Y,f(\mathbf{X}))] = \iint L(y,f(\mathbf{x}))\,p(\mathbf{x},y)\,d\mathbf{x}\,dy$$

and $p(\mathbf{x},y) = p(y|\mathbf{x})\,p(\mathbf{x})$

$$= \int \underbrace{\int L(y,f(\mathbf{x}))\,p(y|\mathbf{x})\,dy}_{\mathbb{E}[L(Y,f(\mathbf{x}))\,|\,\mathbf{X}=\mathbf{x}]}\;p(\mathbf{x})\,d\mathbf{x}$$

$$\mathbb{E}_{\mathbf{X}}\Big[\qquad\qquad\Big]$$

iterated expectation $\longrightarrow$

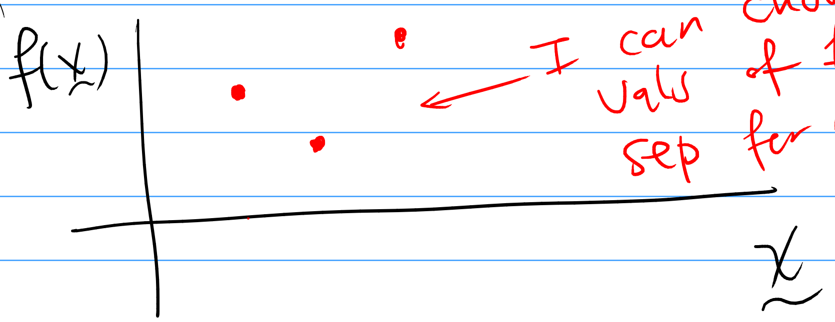
$$= \mathbb{E}_{\mathbf{X}}\Big[\underbrace{\mathbb{E}[L(Y,f(\mathbf{x}))\,|\,\mathbf{X}=\mathbf{x}]}_{A(\mathbf{x})}\Big]$$

Want to choose f to minimize.

$$= \int A(\mathbf{x})\,p(\mathbf{x})\,d\mathbf{x}$$

Consider <u>no restrictions</u> on f.

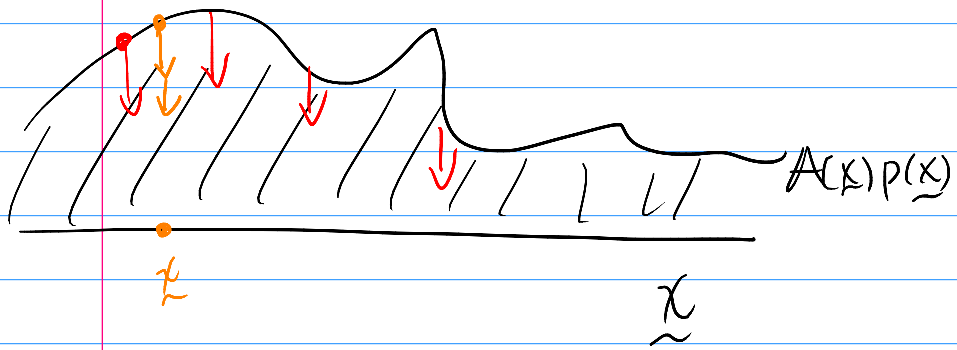
To choose f I need to tell you value of $f(\mathbf{x})$ for each $\mathbf{x}$.



I can choose vals of $f(\mathbf{x})$ sep for each $\mathbf{x}$

Want to choose $f(\underset{\sim}{x})$ to min. $\int A(\underset{\sim}{x})p(\underset{\sim}{x})d\underset{\sim}{x}$

depends on $f(\underset{\sim}{x})$



Game to do this is to sep. for each $\underset{\sim}{x}$ choose val. of $f(\underset{\sim}{x})$ to make $A(\underset{\sim}{x})p(\underset{\sim}{x})$ as small as possible.

Since f doesn't affect p we can just choose f to make $A(\underset{\sim}{x})$ as small as possible.

Punchline: to min risk I can choose $f(\underset{\sim}{x})$ sep. for each $\underset{\sim}{x}$ to min

$$A(\underset{\sim}{x}) = \int L(y, f(\underset{\sim}{x}))p(y|\underset{\sim}{x})dy$$

consider: $L(y, f(\underset{\sim}{x})) = (y - f(\underset{\sim}{x}))^2$

$$f^*(\underset{\sim}{x}) = \underset{f(\underset{\sim}{x})}{\operatorname{argmin}}\ \mathbb{E}[(Y - f(\underset{\sim}{x}))^2 | \underset{\sim}{X} = \underset{\sim}{x}]$$

$$= \underset{c}{\operatorname{argmin}}\ \mathbb{E}[(Y - c)^2 | \underset{\sim}{X} = \underset{\sim}{x}]$$

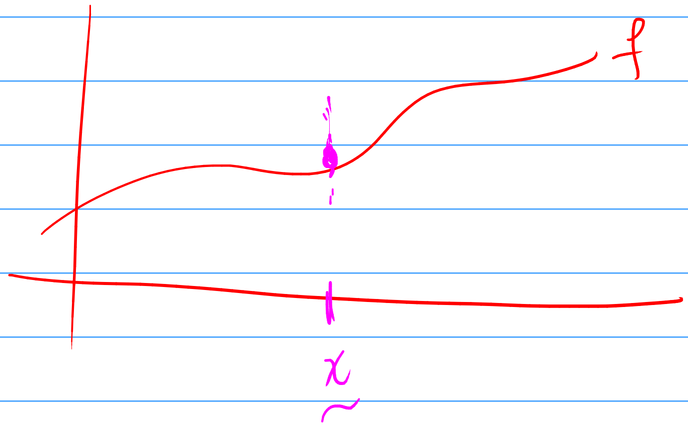
$$\arg\min_{c} \mathbb{E}[(Z-c)^2] = \mathbb{E}Z$$

Why? $\mathbb{E}[Z^2 + c^2 - 2Zc] = \mathbb{E}[Z^2] + c^2 - 2c\,\mathbb{E}Z$

$$\frac{\partial}{\partial c}[\cdots] = 2c - 2\mathbb{E}Z = 0$$

$$\Rightarrow c = \mathbb{E}Z$$

So... $f^*(\underset{\sim}{x}) = \mathbb{E}[Y \mid \underset{\sim}{X} = \underset{\sim}{x}]$



What about other losses?

$$L(y, f(\underset{\sim}{x})) = |Y - f(\underset{\sim}{x})|$$

$$f^*(\underset{\sim}{x}) = \text{median}(Y \mid \underset{\sim}{X} = \underset{\sim}{x})$$

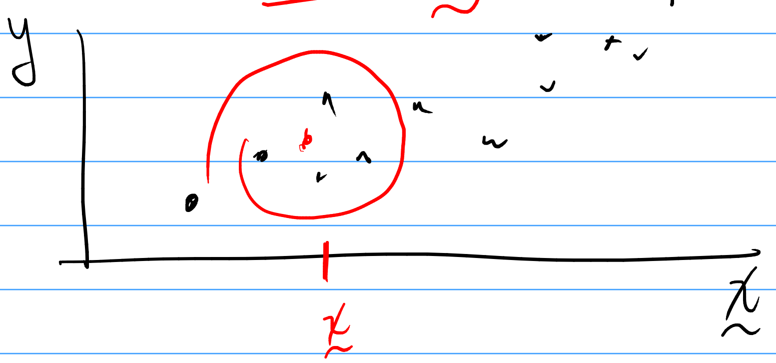
Problem: I don't know $p(\underset{\sim}{x}, y)$.

All I have is training data.

Let's approximate $\mathbb{E}[Y \mid \underset{\sim}{X} = \underset{\sim}{x}]$.

Ex. (1) $\hat{f}(\underset{\sim}{x}) =$ avg. of y_n's for $\underset{\sim}{x}_n$ near $\underset{\sim}{x}$

KNN regression



(2) Maybe assume $\mathbb{E}[Y \mid \underset{\sim}{X} = \underset{\sim}{x}] = \underset{\sim}{x}^T \beta$

and then build $\hat{f}$ by est. β.

OLS regression

Stat. Learning

- Supervised
  - regression ← so far
  - classification ← now
- unsupervised

Setup: Classification

$\underset{\sim}{x} \in \mathbb{R}^p$ and $y \in \mathcal{C}$

$$\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$$

set of possible classes.

Goal: find some $\hat{f}$ so that $\hat{f}(\underset{\sim}{x}) \approx y$

Need a loss fn   0-1 loss

$$L(y, \hat{f}(\underset{\sim}{x})) = \mathbb{1}(y \neq \hat{f}(\underset{\sim}{x}))$$

$$= \begin{cases} 0 & \hat{f}(\underset{\sim}{x}) = y \\ 1 & \hat{f}(\underset{\sim}{x}) \neq y \end{cases}$$

What is f^*?

$$f^*(\underset{\sim}{x}) = \underset{c}{\arg\min}\, \mathbb{E}[L(y, c) \mid \underset{\sim}{X} = \underset{\sim}{x}]$$

$$\mathbb{E}[\mathbb{1}(z \in A)] = \mathbb{P}(z \in A)$$

$$= \underset{c}{\arg\min}\, \mathbb{P}(Y \neq c \mid \underset{\sim}{X} = \underset{\sim}{x})$$

$$= \arg\min_{c} \left(1 - P(Y=c \mid \underline{X} = \underline{x}\right)$$

$$\boxed{f^*(\underline{x}) = \arg\max_{c} P(Y=c \mid \underline{X} = \underline{x})}$$

Bayes' classifier

predict class c w/ highest cond. prob.

e.g. $K=3$

$$P(Y=1 \mid \underline{X} = \underline{x}),\ P(Y=2 \mid \underline{X} = \underline{x}),\ P(Y=3 \mid \underline{X} = \underline{x})$$

pick the largest