Jan B
Math 650.4

Late April

Theory of Bias-Variance Decomposition

Recall $y = g + \overset{\text{residual}}{e}$

$$y = g + \underbrace{(f-g)}_{\substack{\text{error due}\\ \text{to misspecification}\\ \text{\& estimation error}}} + \delta$$

$$e = y - g = f - g + \delta$$

so $e^2 = (f - g + \delta)^2$

r.v.
random variable

I wan to gauge "mean squared error" (MSE) for a new observation $\vec{x}^*$ and a model built from a single data set $\mathbb{D}$

to do mean, as in expectation, need random variables from Math 241

We need to assume a data generating process (DGP) (random variable model)

DGP Assumptions

1. $\mathbb{D} = (X, y)$ is fixed
2. $\vec{x}^*$ is fixed
3. δ is drawn from r.v. Δ with $E[\Delta \mid \vec{X} = \vec{x}^*] = 0$

δ is a r.v.

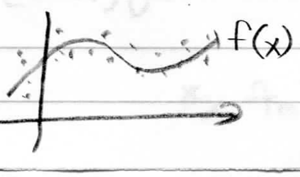
4. $Y = f(\vec{x}^*) + \Delta$

   Y is random through Δ only
   E is also a r.v. through Δ.

r.v.	realization
E	e
Y	y

$$\Rightarrow \underset{\uparrow \text{ expectation}}{E}[Y \mid \vec{X} = \vec{x}^*] = f(\vec{x}^*)$$ conditional expectation function

(expectation of Y is f)



Theory of Bias-Variance Decomposition

Only source of error is Δ^*

$$MSE = E[\mathcal{E}^2 \mid \vec{X} = \vec{x}^*] = E_{\Delta^*}[(Y^* - g(\vec{x}^*))^2 \mid \vec{X} = \vec{x}^*]$$

Imagine g is perfect, i.e. $g = f$

$$= E_{\Delta^*}[(Y^* - f(\vec{x}^*))] = E_{\Delta^*}[\Delta^2] = \sigma^2$$

everywhere from now on, but notation surpressed

irreducible squared error (theoretical best MSE)

Assume Δ is homoskedastic

$$Var[\Delta \mid X] = Var[\Delta] = E[\Delta^2] - (E[\Delta])^2 = E[\Delta^2] = \sigma^2$$

this is 0

$$= E_{\Delta^*}[Y^{*2}] - 2E_{\Delta^*}[Y^* g(\vec{x}^*)] + E_{\Delta^*}[g(\vec{x}^*)^2]$$

g is const so g is indep of Δ^*

$g = \mathcal{A}_{\mathcal{H}}(\mathbb{D})$

$g = \mathcal{A}_{\mathcal{H}}(X, y)$

since $\mathbb{D}$ is fixed, g is fixed

$$= \underbrace{E[(f(\vec{x}^*) + \Delta^*)^2]}_{\underbrace{E[f(\vec{x}^*) + 2f(\vec{x}^*) + \Delta^{*2}]}_{f(\vec{x}^*)^2 + \sigma^2}} - 2g(\vec{x}^*)f(\vec{x}^*) + g(\vec{x}^*)^2$$

by fact

bias of g $= g - f$

$$= \sigma^2 + \underbrace{(g(\vec{x}^*) - f(\vec{x}^*))^2}$$

$$= \sigma^2 + Bias[g(\vec{x}^*)]^2$$

$Bias[g(x)] = g(x) - f(x)$

upward bias if $g > f$

downward bias if $g < f$

(without homoskedicity, would be $= \sigma^2(\vec{x}^*) + Bias[g(\vec{x}^*)]^2$)

different MSE for every x^* ← MSE is a function of x^*

Modify Assumption 1

① X is fixed, y is drawn from $f(\vec{x}) + \Delta$

$$\mathbb{D} = \left([X], [y] \right) \qquad \vec{y} = f(\vec{x}^*) + \vec{\delta}$$

your D is just one realization (here y's are the r.v.'s)

$\vec{y}$ is drawn from $\underbrace{\vec{Y}}_{\text{r.v.'s}}$

$\delta_1, \delta_2, \dots, \delta_n$ is drawn from $\Delta_1, \Delta_2, \dots, \Delta_3$

$g = \mathcal{A}(x, y)$

Same x's different y's

↳ this is called Data set - Data Set Variability

Only source of error is $\Delta^*, \Delta_1, \Delta_2, \dots, \Delta_n$

Doing MSE over all possible y's (expectation over all possible data sets)

$$MSE(\vec{X}^*) = E_{\Delta_1, \dots \Delta_n, \Delta^*}\left[E^2 \mid \vec{X} = \vec{x}^* \right]$$

$$= E_{\Delta_1, \dots, \Delta_n, \Delta^*}\left[(Y^* - g(\vec{x}^*))^2 \mid \underbrace{\vec{X} = \vec{x}^*}_{\text{everywhere, but notation dropped}} \right]$$

$$= E_{\Delta^*}\left[Y^{*2} \right] - 2 E_{\Delta_1, \dots, \Delta_n, \Delta^*}\left[Y^* g(\vec{x}^*) \right] + E_{\Delta_1, \dots, \Delta_n}\left[g(\vec{x}^*)^2 \right]$$

g is not fixed now

$$= \underbrace{E_{\Delta^*}\left[(f(\vec{x}^*) + \Delta^*)^2 \right]} - 2 E_{\Delta^*}\left[Y^* \right] E_{\Delta_1, \dots, \Delta_n}\left[g(\vec{x}^*) \right] + E_{\Delta_1, \dots, \Delta_n}\left[g(\vec{x}^*)^2 \right]$$

$$E_{\Delta^*}\left[f(\vec{x}^*) + 2 f(\vec{x}^*) \Delta^* + \Delta^{*2} \right] - 2 f(\vec{x}^*) E\left[g(\vec{x}^*) \right] + Var\left[g(\vec{x}^*) \right] + E\left[g(\vec{x}^*) \right]$$

$$= \sigma^2 + \left(E[g(\vec{x}^*)] - f(\vec{x}^*) \right)^2 + Var\left[g(\vec{x}^*) \right]$$

$$= \sigma^2 + Bias(E[g(\vec{x}^*)])^2 + Var[g(\vec{x}^*)]$$

$$= \sigma^2 + \left(\text{Bias}\left[E[g(\vec{x}^*)]\right]\right)^2 + \text{Var}\left[g(\vec{x}^*)\right]$$

Now, change Assumption 1 again,

made X random too

① X is random, drawn from r.v. $\vec{X}$
Y is random, drawn from $f(\vec{x}) + \Delta$
r.v.'s $\Delta_1, \dots, \Delta_n$

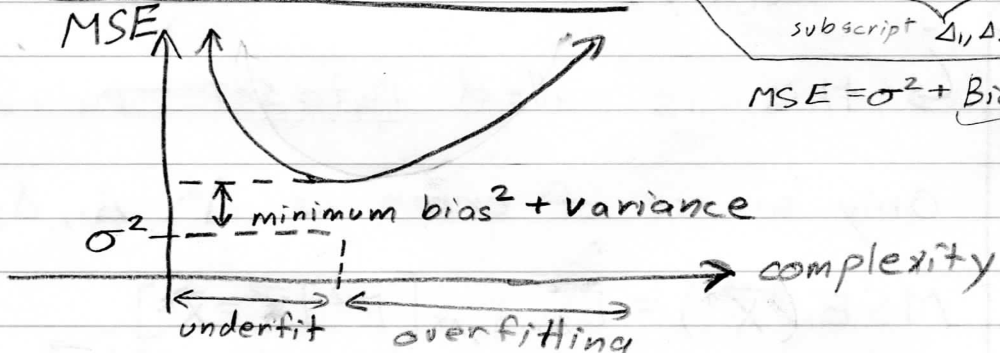
Now, $\underset{\substack{\uparrow \\ \text{generalization error}}}{MSE} = E_X\left[MSE(\vec{x})\right]$

$$= \sigma^2 + E_X\left[\left(\text{Bias}\left(\overbrace{E[g(\vec{x})]}^{\text{mean model}}\right)\right)^2\right] + E_X\left[\text{Var}[g(\vec{x})]\right]$$

subscript $\Delta_1, \Delta_2, \dots, \Delta_n$

$$MSE = \sigma^2 + \underbrace{\text{Bias}^2 + \text{Variance}}_{\text{want to minimize this part}}$$

Bias-Variance Tradeoff



Underfitting (complexity too low)

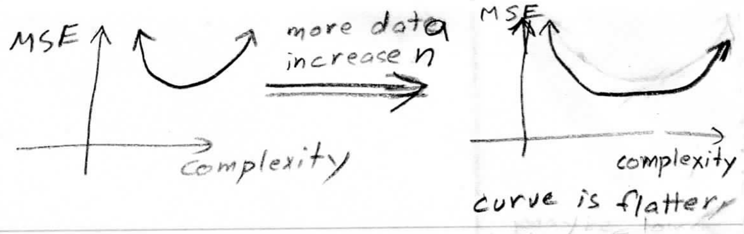
$E(g)$ is far from f

$\Rightarrow$ Bias^2 term is high
Var term is low

Overfitting (complexity too high)

$E(g)$ is approximately f

$\Rightarrow$ Bias^2 term is low
Var term is high



can make bias $\to 0$ by overfitting (more complexity)

but as complexity $\uparrow$, get variance $\uparrow$

Let's say we have a model that's somewhat overfit now

look at number of data points n

As $n \to \infty$, get $g \to E[g] \Rightarrow Var[g] \to 0$