Lec 21 4/30/19 Math 390.4

Regression Trees $(y \subseteq \mathbb{R})$ review with R demo

Theory of Bias-Variance Decomposition

Recall... $y = g + e = g + \underbrace{(f-g)}_{\text{error due to mis. + estimation}} + \underbrace{\delta}_{\text{error due to ignorance}}$

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

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

What if I want to measure the "mean squared error" (MSE) for a new observation $\vec{x}_*$?

I need to assume a r.v. model somewhere!

Let's assume Δ is the r.v. where δ is realized from.

and $\vec{x}$, $f(\vec{x})$ are constant.

$\overset{\text{r.v.}}{Y} = f(\vec{x}) + \Delta$

this is equivalent to assuming

(I) $E[Y \mid \vec{X} = \vec{x}] = f(\vec{x})$ ← conditional expectation function

$\Rightarrow E[Y \mid \vec{X} = \vec{x}] = E[f(\vec{x}) + \Delta \mid \vec{X} = \vec{x}] = E[f(\vec{x}) \mid \ ] + E[\Delta \mid \ ] = f(\vec{x})$

$\Rightarrow E[\Delta \mid \vec{X} = \vec{x}] = 0.$

the expected Y at any x is $f(\vec{x})$ $\Rightarrow$ the error due to ignorance must have mean center of 0.

(II) Also assume variance does not depend on $\vec{x}$. $Var(\Delta \mid \vec{X} = \vec{x})$

$$= Var(\Delta) = \sigma^2$$

$$\Rightarrow E[\Delta^2] = \sigma^2$$

Not strictly necessary! But makes life easy here.

Back to MSE for a new obs. x^*. If we know f...

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

$$= E(\Delta^{*2}) = \sigma^2$$ expectation taken over Δ^*

If we don't know f... then $MSE(\vec{x}^*) \geq \sigma^2$. Proof:

drop this notation...

$$E[(Y - g(\vec{x}^*))^2 \mid \ldots] = E[Y^2 - 2Y g(\vec{x}^*) + g(\vec{x}^*)^2]$$ the only randomness is in Δ^*!

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

$$= E[(f + \Delta)^2] - 2E(Y)E[g] + E[g^2]$$

$$= E[f^2 + 2f\Delta + \Delta^2] - 2fg + g^2$$

$$= f^2 + \sigma^2 - 2fg + g^2$$ expected sqd. errors additive

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

Now... instead of taking expectation over just Δ^*, we take expectation over $\Delta_1, \Delta_2, \ldots, \Delta_n, \Delta^*$ which means the randomness in $\mathbb{D}$ itself.

"Dataset-dataset variability"

Assume! $\vec{X}, Y$ are r.v.'s and $(\vec{X}_1, Y_1), (\vec{X}_2, Y_2), \ldots$ are all iid realizations

$$MSE(\vec{x}^*) = E_{\Delta_1, \ldots, \Delta_n, \Delta^*}\left[\left(\underbrace{Y^*}_{\text{just a function of } \Delta^*} - \underbrace{g(\vec{x}^*)}_{\text{is a r.v. now based on } \Delta_1, \ldots, \Delta_n}\right)^2 \,\middle|\, \vec{X} = \vec{x}^*\right]$$

independent! why $\mathbb{D} \perp (\vec{x}^*, y^*)$!

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

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

subscripts dropped everywhere

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

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

$$= \underbrace{\sigma^2}_{\text{irreducible gen. err.}} + \underbrace{Bias\left[g(\vec{x}^*)\right]^2}_{\text{How far is } g \text{ from } f \text{ on average?}} + \underbrace{Var\left[g(\vec{x}^*)\right]}_{\text{How variable is } g \text{ from its center measured in mean sqd. distance.}}$$

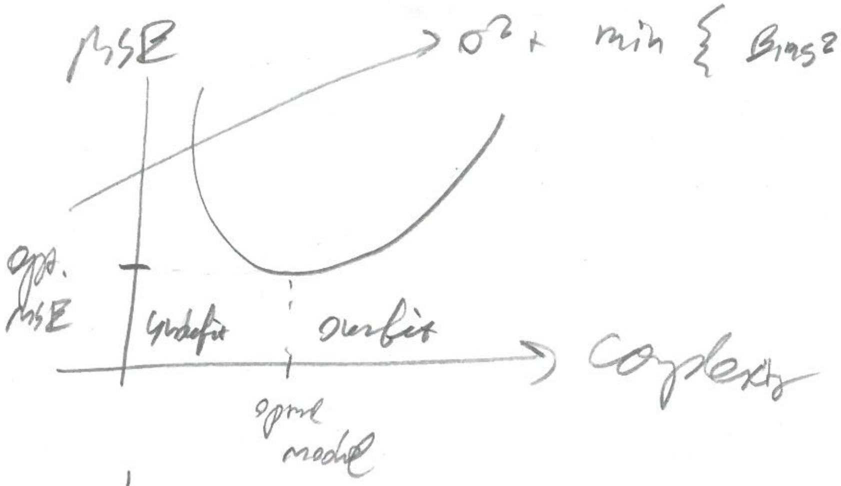
One more small point... this is all for one new obs. $\vec{x}^*$. Let's assume $\vec{x}^*$ is realized from $P(\vec{x})$. Then...

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

$$= \underbrace{\sigma^2}_{\text{irred. err.}} + \underbrace{E_X\left[Bias\left[g(\vec{x}^*)\right]^2\right]}_{\text{expected sqd. model bias}} + \underbrace{E_X\left[Var\left[g(\vec{x}^*)\right]\right]}_{\text{expected model variance}}$$

DEMO

Is there a "bias-variance tradeoff"? Yes and no.
Yes in the extremes. No in not the extremes.



What happens during underfitting? $E_X\left[Bias\left[g(\vec{x})\right]^2\right] = E\left[\left(g(\vec{x}) - f(\vec{x})\right)^2\right]$ is high since $f(\vec{x})$ likely way more complicated than $g(\vec{x})$

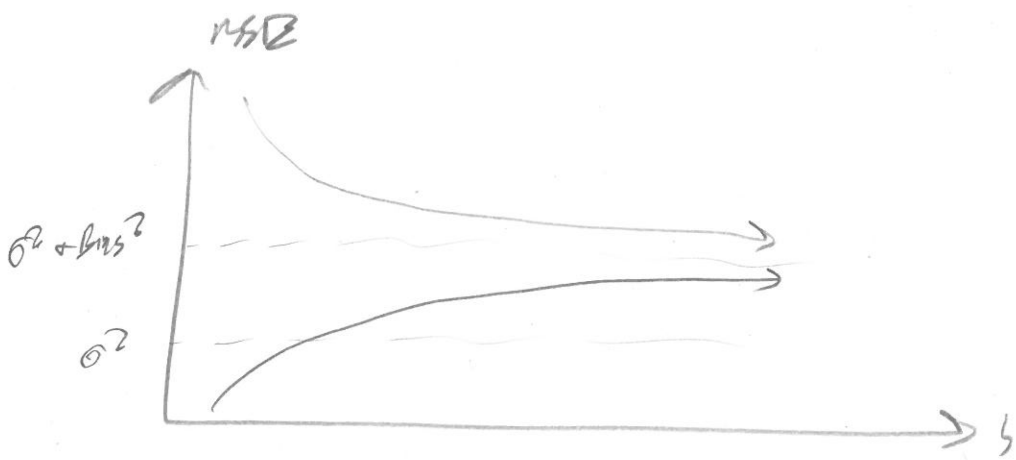
$E_X\left[Var\left[g(\vec{x})\right]\right]$ is low since if $\mathbb{D}$ changes, fit doesn't change too much

What happens during overfitting?

$E_x[\text{Bias}(g(x))^2]$ is low since $g(x)$ will be complicated enough to locate $f(x)$ here

$E_x[\text{Var}(g(x))]$ is high since it is tightly fitting $\mathcal{D}$ which change dataset – dataset.

For any fixed complexity. (fixed algorithm)



as $n \to \infty \Rightarrow \text{Var}[g(x)] \to 0$ Why? But bias does not budge. Why?

So there is a bias variance tradeoff ~~but~~ there is a spot where both together are minimized.
Each algorithm has a different curve.