

$$\rightarrow \boxed{y \leftarrow x_1 + x_2 + x_3 + x_4 + x_5} \rightarrow R^2, \text{Adj. } R^2$$

Stepwise
Regression

$$y \leftarrow \underline{x_1 + x_2 + x_3 + x_4} \rightarrow R^2, \text{Adj. } R^2$$

$$y \leftarrow x_1 + x_2 + x_4 + x_5 \rightarrow R^2, \text{Adj. } R^2$$

$$y \leftarrow x_1 + x_3 + x_4 + x_5 \rightarrow R^2, \text{Adj. } R^2$$

$$y \leftarrow x_2 + x_3 + x_4 + x_5 \rightarrow R^2, \text{Adj. } R^2$$

$$y \leftarrow x_1 + x_2 + x_3 \rightarrow R^2, \text{Adj. } R^2$$

$$y \leftarrow x_1 \rightarrow R^2, \text{Adj. } R^2$$

Backward elimination

Stepwise regression

backward elimination

forward selection

x_1, x_2, x_3, x_4, x_5

$$y \leftarrow x_1 \rightarrow \underline{R^2}$$

$$y \leftarrow x_2 \rightarrow \underline{R^2}$$

$$y \leftarrow x_3 \rightarrow \underline{R^2}$$

$$y \leftarrow x_4 \rightarrow \underline{R^2}$$

...

$$y \leftarrow x_1 + x_2 + x_3 + x_4 + x_5$$

①

$$y \leftarrow x_1 + x_2 + x_3 + x_4 + x_5 \rightarrow 0.9225$$

→ HthVaria

②

$$\underline{y \leftarrow x_1} \quad \leftarrow \underline{0.9223}$$

low variance

X

0003

↑

low

low

$$\underline{\underline{VIF}} = \frac{1}{1 - R^2} \rightarrow \underline{\underline{lm}}$$

$x_1 \ x_2 \ x_3 \ x_4$

↑

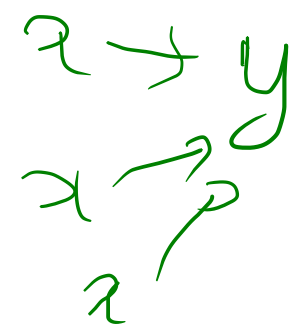
$$x_1 + x_2 + x_3 \xrightarrow{R} \underline{x_4} \rightarrow R^2$$

$$x_1 + x_2 + \underline{x_4} \xrightarrow{R} \underline{x_3} \rightarrow R^2$$

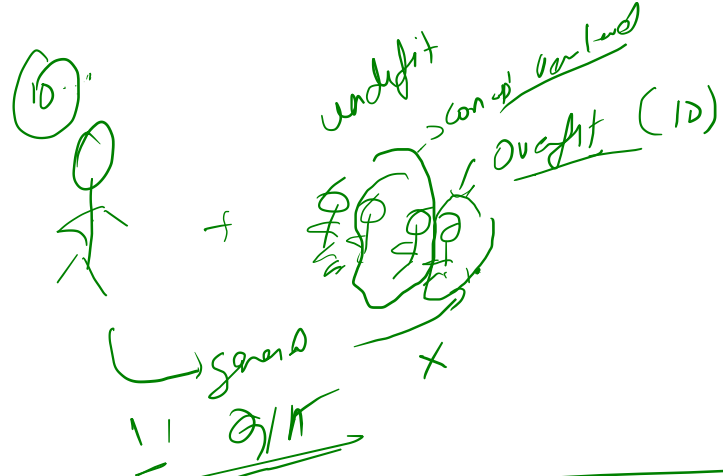
$$x_1 + x_3 + \underline{x_4} \xrightarrow{R} \underline{x_2} \rightarrow R^2$$

$$x_2 + x_3 + \underline{x_4} \xrightarrow{R} \underline{x_1} \rightarrow R^2$$

$$\begin{aligned} \rightarrow VIF &= \frac{1}{1 - R^2} \\ \rightarrow VIF & \\ \rightarrow VIF & \\ \rightarrow VIF & \end{aligned}$$



$VIF > 5$ highly affected by
multicollinearity

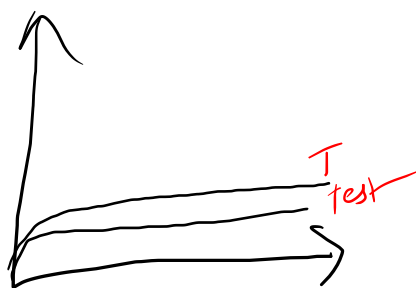


6%

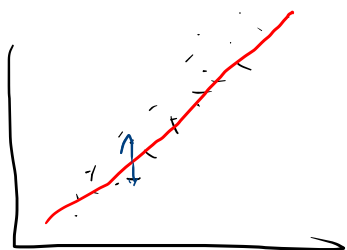
underfit



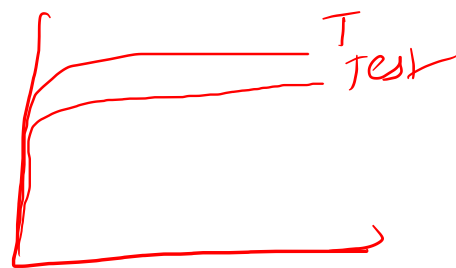
train accuracy ↓
test accuracy ↓



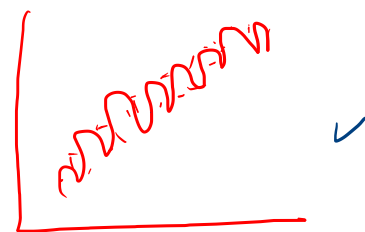
general fit



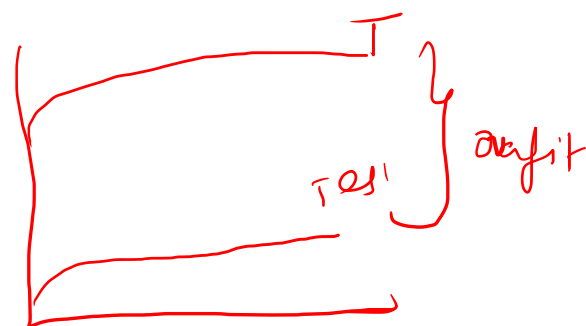
train accuracy ↑
test accuracy ↑



overfit



train accuracy ↑
x test accuracy ↓



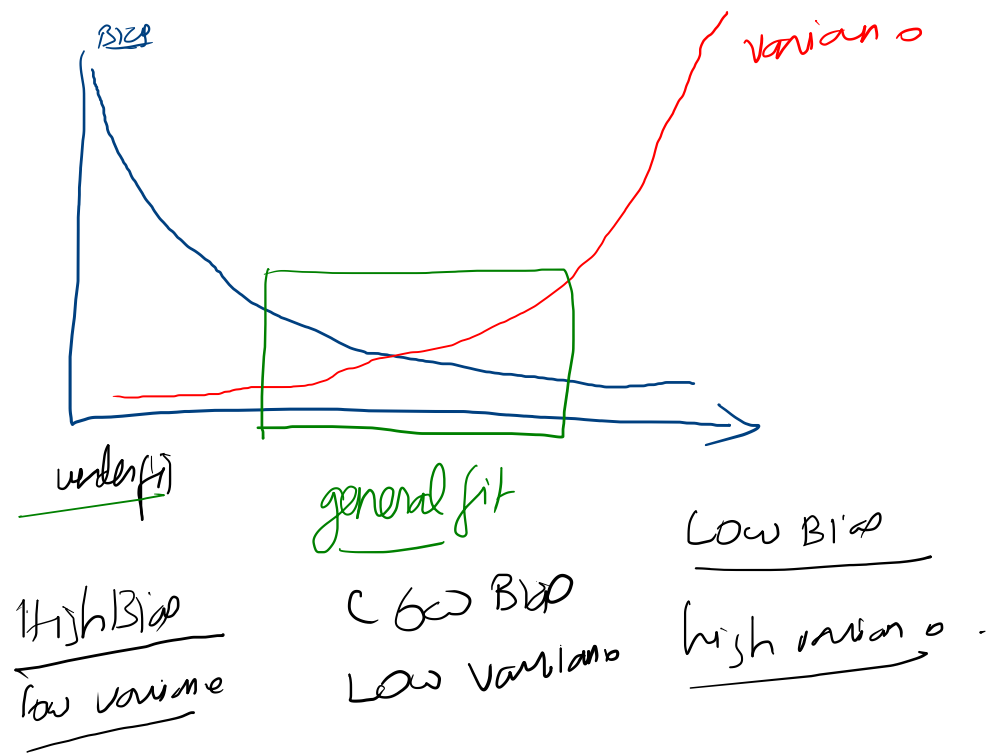
Bias → variance

$y \in \lambda_1$

$y \in \lambda_1 + \lambda_2 + \dots + \lambda_n$

$y \in \lambda_1 + \lambda_2 + \dots + \lambda_n$
variance

Bias (vs) variance tradeoff



- ① train with different algo.
- ② Add more variables
- ③ grid search, Hyper parameter tuning

⇒ generalized fit

Gross Validation

dimension reduction

regularisation

L_1

L_2

add penalty
w.r.t
coefficient

$$y = 1x_1 + 2x_2 + 3x_3$$

↓
penalty term

$$y = 0.5x_1 + 1.5x_2 + 2x_3$$