Bias-variance tradeoff

- Y = response
- $\underline{X} = (x_1, \ldots, x_p)^T$ = feature vector
- ε = mean zero random error

Model: $Y = f(\underline{x}) + \varepsilon$

Given data, have $\hat{f}$, want to predict a new Y for features $\underline{X}_0$, use $\hat{Y} = \hat{f}(\underline{x}_0) = \hat{f}$

predicted Y!

How good is $\hat{f}$ at predicting Y? → the unknown Y

$$MSE = E[(Y-\hat{f})^2] = \cdots =$$

$$= E[(f-\hat{f})^2] + E(\varepsilon^2) + E(2\varepsilon(f-\hat{f}))$$

1 2 3

$$③ = E\left(2\underset{\text{random}}{\overset{\text{not}}{\cancel{\varepsilon}}}(f-\underset{\text{random}}{\overset{\text{not}}{\hat{f}}})\right) = 2E\left(\varepsilon(f-\hat{f})\right) = 2\,\underbrace{[E\varepsilon]}\,\underbrace{[E(f-\hat{f})]}$$

Assume ε + $\hat{f}$ are uncorrelated

$$= 2 \cdot 0 \cdot \text{something} = 0$$

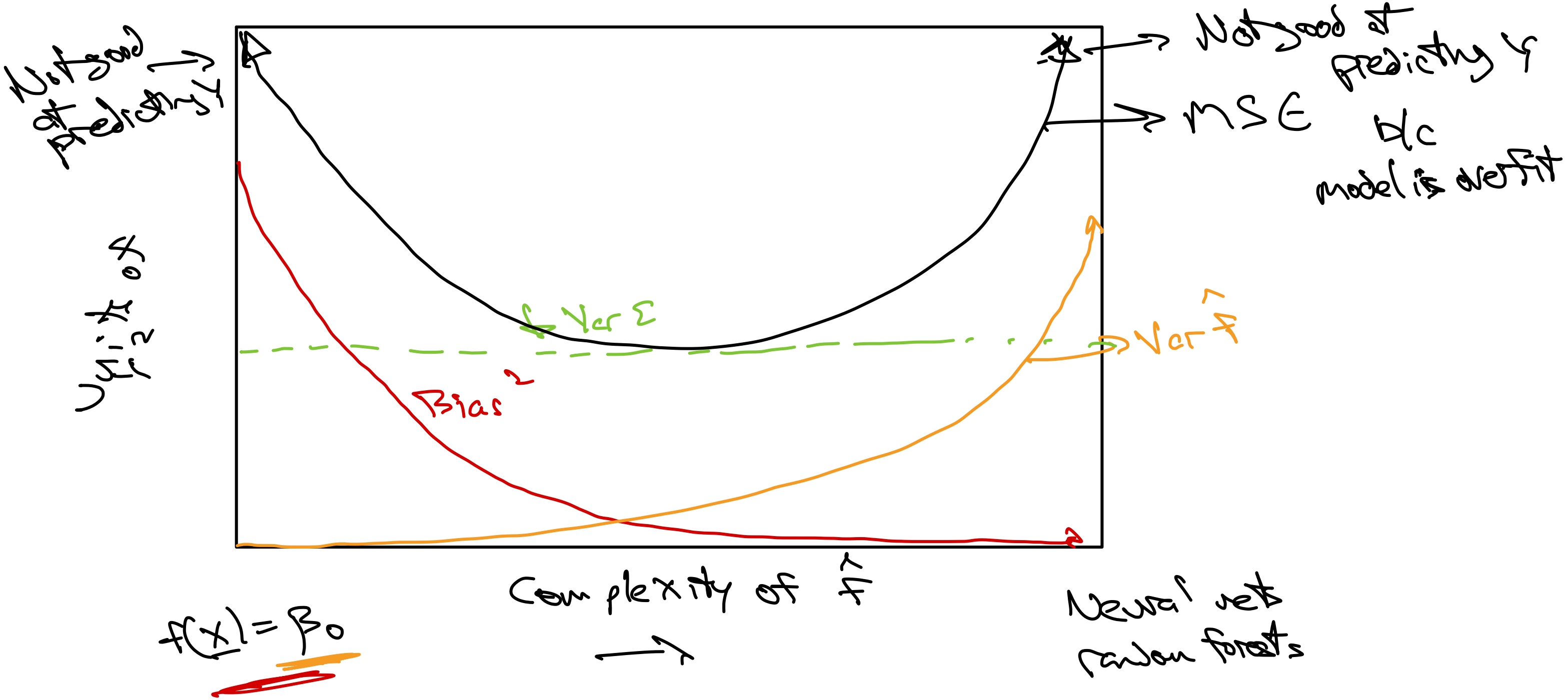
$$② = E(\varepsilon^2) = E(\varepsilon^2) - 0 = E(\varepsilon^2) - (E\varepsilon)^2 = \operatorname{Var}\varepsilon$$

$$① = E\left[(f-\hat{f})^2\right] = E\left[\left(\underbrace{f - E\hat{f}} + \underbrace{E\hat{f} - \hat{f}}\right)^2\right]$$

$$= E\left[(f-E\hat{f})^2\right] + E\left[(\hat{f}-E\hat{f})^2\right] + 0$$

$$\Rightarrow MSE = (f-E\hat{f})^2 + E(\hat{f}-E\hat{f})^2 + \operatorname{Var}\varepsilon$$

$$= \text{Bias}^2 + \operatorname{Var}(\hat{f}) + \operatorname{Var}\varepsilon$$



Not good at predicting Y
Not good at predicting Y b/c model is overfit
MSE
Var ε
Var f̂
Bias²
Complexity of f̂
f(x) = β₀
Neural nets
random forests

Ch2 : Matrix review : Homework!

3 Random Vectors

If X & Y are random variables, then

$$\mathrm{Cov}(X,Y) = E\big[\underbrace{(X-EX)}_{\text{units of } X}\underbrace{(Y-EY)}_{\text{units of } Y}\big] \in \mathbb{R}$$

and

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}X \cdot \mathrm{Var}Y}} = \frac{\mathrm{Cov}(X,Y)}{\underbrace{SD(X)}\cdot\underbrace{SD(Y)}} \in [-1,1]$$

→ unit less

Properties

- $\mathrm{Var}X = \mathrm{Cov}(X,X)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$

(Z random) • $\mathrm{Cov}(X+Y, Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$

a, b, c, d fixeds

- $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$

Need similar ideas for random vectors.

$$\underset{n\times 1}{\underline{X}} = \begin{pmatrix} X_1 \to \text{r.v.} \\ X_2 \to \text{r.v.} \\ \vdots \\ X_n \to \text{r.v.} \end{pmatrix} \qquad \underset{m\times 1}{\underline{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} \Bigg\} \text{ collection of R.V.s}$$

$$\text{Cov}(\underline{X}, \underline{Y}) =$$

$n \times m$ matrix

not square

$$\begin{pmatrix} \text{Cov}(X_1, Y_1) & \text{Cov}(X_1, Y_2) & \cdots & \text{Cov}(X_1, Y_m) \\ \text{Cov}(X_2, Y_1) & \text{Cov}(X_2, Y_2) & \cdots & \text{Cov}(X_2, Y_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, Y_1) & \text{Cov}(X_n, Y_2) & & \text{Cov}(X_n, Y_m) \end{pmatrix}$$

called a covariance matrix

$\boxed{\text{Properties}}$

Variance-covariance matrix

- $\operatorname{Var} \underline{X} = \operatorname{Cov}(\underline{X}, \underline{X}) = \begin{pmatrix} \operatorname{Var} X_1 & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var} X_2 & & \vdots \\ \vdots & & \ddots & \\ \operatorname{Cov}(X_n, X_1) & \cdots & & \operatorname{Var} X_n \end{pmatrix}$

  $n \times n$ matrix

- $\operatorname{Cov}(A\underline{X} + \underline{\mu}, B\underline{Y} + \underline{\nu}) = A \operatorname{Cov}(\underline{X}, \underline{Y}) B^T$

- $\operatorname{Cov}(\underline{X}, \underline{Y}) = \operatorname{Cov}(\underline{Y}, \underline{X})^T$

- $\operatorname{Cor}(\underline{X}) = \begin{pmatrix} 1 & \operatorname{Cor}(X_1, X_2) & \cdots & \operatorname{Cor}(X_1, X_n) \\ \vdots & 1 & & \vdots \\ \vdots & & \ddots & \\ \operatorname{Cor}(X_n, X_1) & \cdots & & 1 \end{pmatrix}$

Useful case: Suppose $\varepsilon_1, \dots, \varepsilon_n$ are uncorrelated mean zero RVs with common variance $Var\, \varepsilon_i = \sigma^2, \forall i$.

If $\underset{n\times 1}{\underline{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ then $E\underline{\varepsilon} = \begin{pmatrix} E\varepsilon_1 \\ E\varepsilon_2 \\ \vdots \\ E\varepsilon_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

and

$Cov(\varepsilon_1, \varepsilon_2) = 0$ b/c

$Var\,\varepsilon_1$

$Var\,\varepsilon_2$

$$\underset{n\times n}{Var\,\underline{\varepsilon}} = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma^2 \end{pmatrix} = \sigma^2 I$$

$Var\,\varepsilon_n$

4 Linear Regression [R]

<u>Setup</u>: Have features $X_1, \dots, X_p$ & continuous response Y.

- $\exists$? relationship between X_i & Y?
- Can we predict Y for new set of Xs?