Why this lecture is important?

17 May 2023 07:16

$\boxed{17}$ → $\boxed{\text{top5}}$

Bias Var tradeoff — know
↳ understand X

$\boxed{\text{2hr}}$ → understand

→ Interview → 8 mins

→ 1% students →

→ Super useful → to all the ML algorithms → Deep learning

→ linear rg
↓
stats

What are we going to study?

17 May 2023 17:38

Bias Var tradeoff

- → Bias
- → Variance
- → Overfitting
- → underfitting
- → tradeoff
- → Bias Var decomposition

The Hidden Truth

17 May 2023 07:16

1000 students

x_1 cgpa	x_2 iq	lpa
8	80	8
9	90	9
.	.	.

1000 → model → predict → cgpa | iq → $\boxed{\text{lpa}}$

100

X Y → $\boxed{y = f(x)}$

x_1 x_2

$\boxed{\text{cgpa} \mid \text{iq}}$ → lpa

for all students

sample → pop predict

$y = \boxed{f(x)}$ → pop

$\boxed{y = f(x) + \boxed{\text{error}}}$

ml model

irreducible error

Irredu reducible

$\hat{y} = f'(x)$

reducible error

$y - \hat{y}$

$\boxed{f(x) - f'(x)}$

reduce

$f'(x) \to f(x)$

Bias Variance Tradeoff

Reducible error

$\boxed{\text{Reducible error} = \text{Bias}^2 + \text{Var}}$

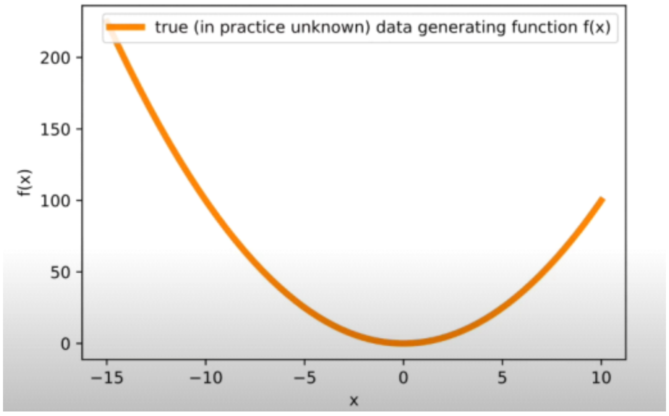
Bias variance decomposition → Regularization

Bias Variance Tradeoff

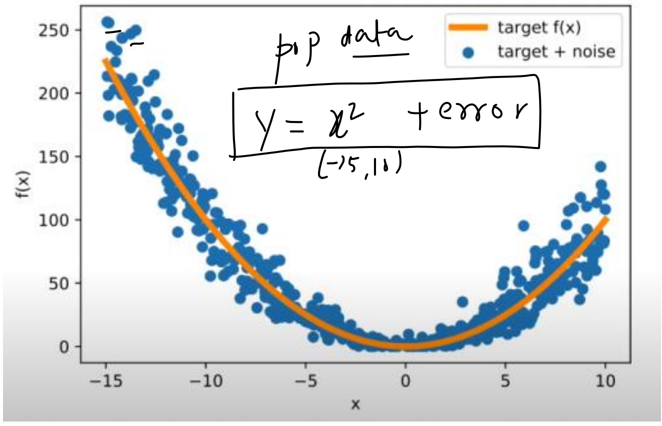
17 May 2023 07:16

$y = f(x) = x^2 \quad [-15, 10]$ $\boxed{y = x2}$

$y = x^2 + \text{error}$ pop $\boxed{1000 \text{ points}}$



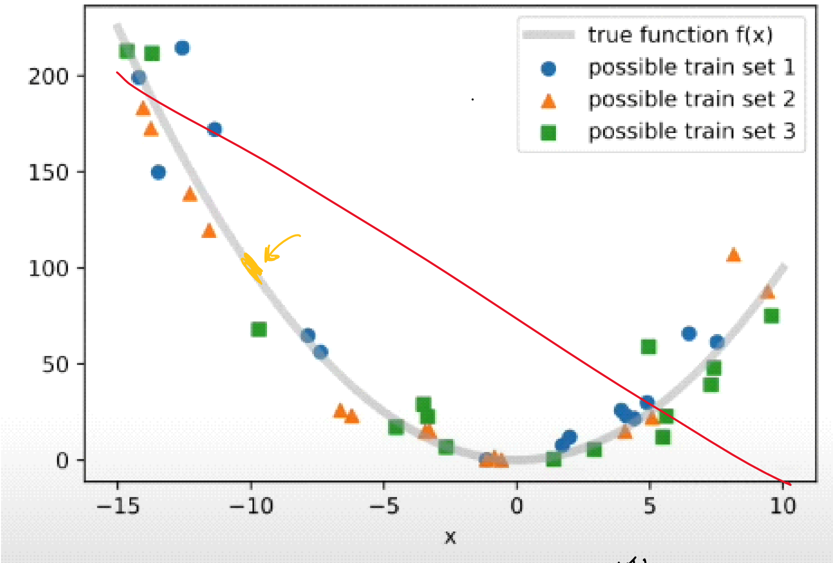
3 samples random

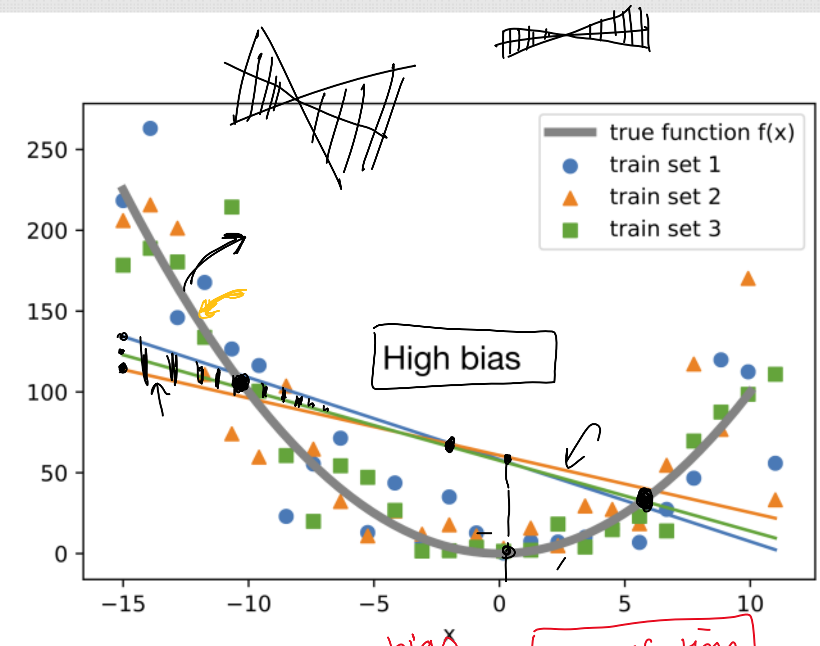


● → 1
▲ → 2
■ → 3

→ 3 student

linear model reg





high bias → $\boxed{\text{underfitting}}$

$\boxed{\text{Bias}}$ → the inability of a ML model to fit the training data

Variance

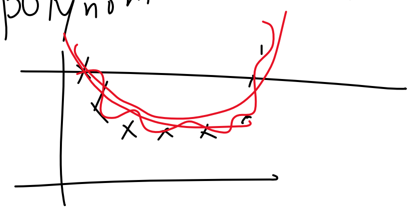
high bias ✓ $\boxed{\text{Low variance}}$

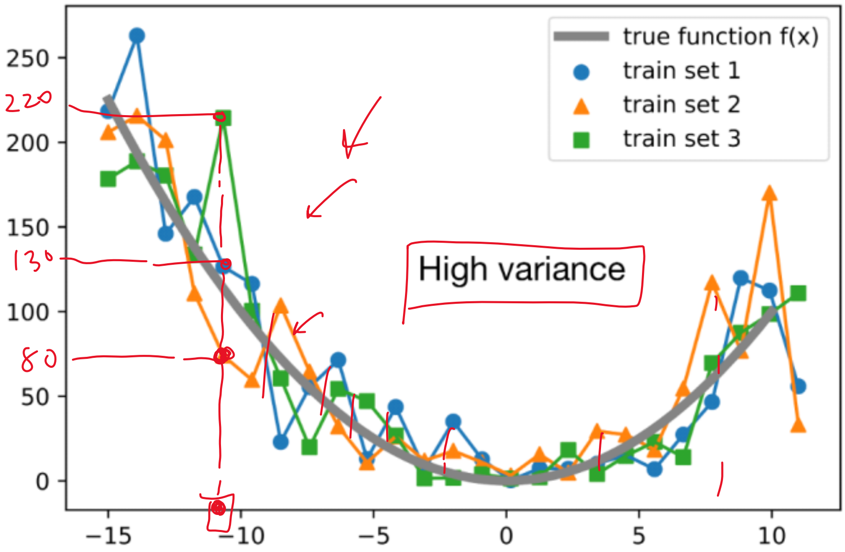
low bias

ml model predict when the training data is chan

polynomial

degree = 3





polynomial high degree

high bias

low bias ✓

Overfitting

high var

→ low bias high variance

high var → overfitting

low bias
low variance

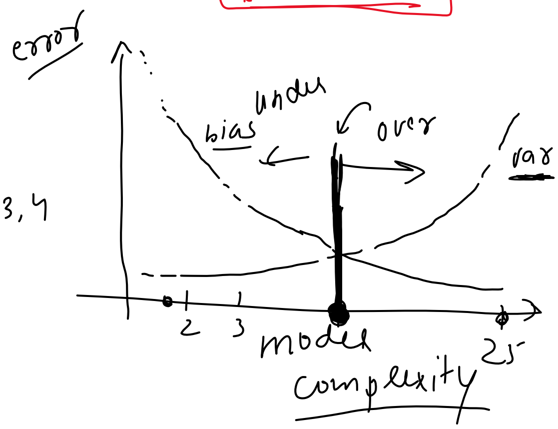
→ bias ↓ var ↑

var ↓ bias ↑

The "trade-off" in bias-variance trade-off refers to the fact that minimizing bias will usually increase variance and vice versa.

{ bias
variance }

poly → degree 2, 3, 4



Some questions

17 May 2023 07:16

1. How would you define bias and variance mathematically?
2. How is bias and variance related to overfitting and underfitting mathematically?
3. Why is there a tradeoff between bias and variance mathematically?

Expected Value and Variance

17 May 2023 07:17

pop mean $\boxed{\text{Prob}}$ ← expected value → pop mean

Expected value represents the average outcome of a random variable over a large number of trials or experiments. → pop mean

In a simple sense, the expected value of a random variable is the long-term average value of repetitions of the experiment it represents. For example, the expected value of rolling a six-sided die is 3.5 because, over many rolls, we would expect to average about 3.5.

$\boxed{\text{1 Lac roll}}$

1, 2, 3, 6, 4, 5

deterministic

$\boxed{\text{mean}}$ = ?

$\boxed{1/6}$ $\boxed{3.5}$

X = rolling a die 1 Lac

$$E[X] = x_1 p(x_1) + x_2 p(x_2) + \dots + x_n p(x_n)$$

$$= 1 \times \tfrac{1}{6} + 2 \times \tfrac{1}{6} + 3 \times \tfrac{1}{6} + 4 \times \tfrac{1}{6} + 5 \times \tfrac{1}{6} + 6 \times \tfrac{1}{6}$$

$$= \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

overall avg mean

$\boxed{4.1}$

$X \to x_1 \to 1$, $x_2 \to 2$, $x_3 \to 3$, x_n, $x_6 \to 6$

5

$\boxed{5\ 3\ 4\ 5\ 3\ 5}$ → mean

$$\frac{5+3+4+5+3+5}{6} = \frac{25}{6}$$

$$\frac{3(5) + 2(3) + 1(4)}{6}$$

$$= \frac{3}{6}(5) + \frac{2}{6}(3) + \frac{1}{6}(4)$$

$$= x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3)$$

$E[X]$ ←

mean → expected value

discrete random var (X)

$$E[X] = \sum_{i=1}^{n} x_i p(x_i)$$

→ expected value → $\boxed{\text{pop mean}}$ → $E[X]$

continuous random var

$$E[X] = \int x_i f(x_i)\, dx$$

$\boxed{\text{Var}(X)}$ → var of pop   pop var of X

$$\boxed{\text{Var}(X) = E[X^2] - (E[X])^2}$$

sample

sample

$$var = \frac{\sum (x_i - \bar{X})^2}{n} = E\left[(X - E[X])^2\right] = Var(x)$$

$\frac{1}{n}$

$E[X]$

avg

$E[X]$ = number

constant

$$Var(X) = E\left[(X - E[X])^2\right]$$

$E[X+Y] = E[X] + E[Y]$

$$= E\left[X^2 + (E[X])^2 - 2XE[X]\right]$$

$E[XY] = E[X]E[Y]$ given X and Y are independent

$$= E[X^2] + E\left[(E[X])^2\right] - E[2XE[X]]$$

$$= E[X^2] + E\left[(E[X])^2\right] - E[2]\,E[X]\quad E[E[X]]\quad E[E[X]]$$

$$= E[X^2] + E\left[(E[X])^2\right] - 2E[X]E[X]$$

$E[\text{constant}]$ = constant

$$= E[X^2] + E\left[(E[X])^2\right] - 2(E[X])^2$$

$$= E[X^2] + (E[X])^2 - 2(E[X])^2$$

$$= E[X^2] - (E[X])^2$$

$$\boxed{Var(X) = E[X^2] - (E[X])^2}$$

$E[X]$

discreet

continous

age

22

22

$$Var(X) = E[X^2] - (E[X])^2$$

$$= E\left[(X - E[X])^2\right]$$

feature

$[X]$

mean

$E[X]$

pop

$$\boxed{var = \frac{\sum (x_i - \bar{x})^2}{n}}$$

mean $\boxed{\frac{\sum}{n}} \rightarrow E$

pop

lg

mean

$$= E\left[(X - E[X])^2\right]$$

Bias ?

sample $E[X]$

$$\boxed{\begin{array}{l}\text{Bias ?}\\ \text{Var ?}\end{array}} \rightarrow \text{mathemati'}$$

$\underline{E[X]}$

What exactly are Bias and Variance Mathematically?

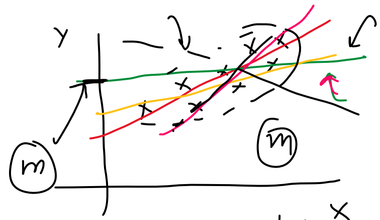
17 May 2023 07:17

Bias

In the context of machine learning and statistics, bias refers to the systematic error that a model introduces because it cannot capture the true relationship in the data. It represents the difference between the expected prediction of our model and the correct value which we are trying to predict. More bias leads to underfitting, where the model does not fit the training data well.

Variance

In the context of machine learning and statistics, variance refers to the amount by which the prediction of our model will change if we used a different training data set. In other words, it measures how much the predictions for a given point vary between different realizations of the model.



true function

$f(x) \rightarrow \boxed{f'(x)}$ $f'(x)$

100 samples → LR → m

$\boxed{E[f'(x)]}$ $E[m]$

$\boxed{E[m]} - m$ → bias

$m - m = 0$

$\boxed{bias = 0}$

$\boxed{Bias = E[f'(x)] - f(x)}$

$f(x) - f(x) = 0$

$\boxed{bias = 0}$

unbiased predictor

param: $b_0\ b_1\ b_2 \ldots b_n$ bias(b)

bias(b_0)) bias(b_1))

→ $Bias(f'(x)) = E[f'(x)] - f(x)$

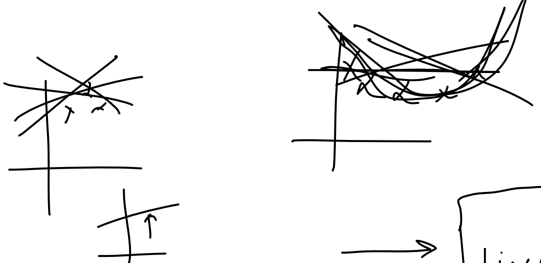
+ve −ve pop

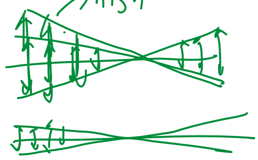
$f(x) \Rightarrow \boxed{y = x^2 + 5}$

$f'(x)$ $f(x) = x^2 + 5$

$x + 5$

high bias → underfitting





$$\boxed{bias(f'(x)) = E[f'(x)] - f(x)}$$

train → 90%
test → 80%

$$\boxed{Var(f'(x)) = E[(f'(x) - E[f'(x)])^2]}$$

high → data change change

overfitting

$E[f'(x)]$

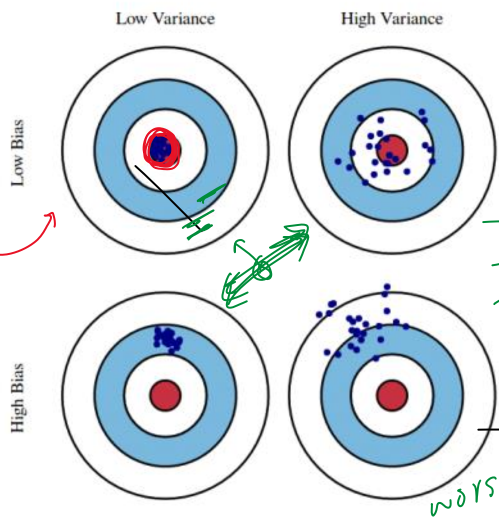
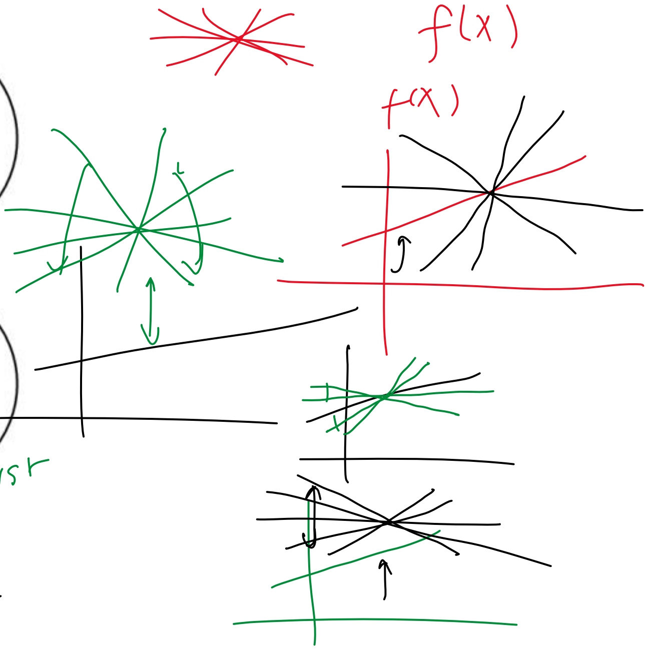


Fig. 1 Graphical illustration of bias and variance.

worst

$f(x)$

$f(x)$



$$var = \frac{\sum (x_i - \bar{x})^2}{n}$$

$y = f(x) + error$ mse $\hat{\theta}$ θ

var =

$\underline{n}$ $\hat{\theta}$ θ

$y = f(x) + \text{error}$ mse

$\hat{y} = f'(x)$

$\text{Bias}(f'(x)) = E[f'(x)] - f(x)$

$\text{Var}(f'(x)) = E[(f'(x) - E[f'(x)])^2]$

$\boxed{\text{Bias Variance Decomposition}}$ $E[(\hat{\theta} - E[\hat{\theta}])^2]$

$\text{Bias} = E[\hat{\theta}] - \theta$

Bias Variance Decomposition

17 May 2023 07:17

Bias-variance decomposition is a way of analysing a learning algorithm's expected generalization error with respect to a particular problem by expressing it as the sum of three very different quantities: bias, variance, and irreducible error.

1. Bias: This is the error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
2. Variance: This is the error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).
3. Irreducible Error: This is the noise term. This part of the error is due to the inherent noise in the problem itself, and can't be reduced by any model.

Loss

↓

mse

$y \to$ actual

$\hat{y} \to$ predicted

$$\boxed{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$$

Loss = bias + variance + irreducible

derive

$$Loss = \underbrace{[bias^2 + variance]}_{\text{reducible}} + \underbrace{[var(\epsilon)]}_{\text{irreducible}}$$

(reducible) noise

cgpa | iq | lpa | pred → irreducible

cgpa	iq	lpa	pred
—	—	8	9
—	—	8	8.1
—	—	7	6.9
—	—	9	10.1

→ 1 reducible

— 0.1 irred

lr → irreducible

$\sum_{i=1}^{n} (y - \hat{y})^2 \leftarrow$ minimize mse

irreducible error

$\hookrightarrow$ mean $= \theta$

var $= \sigma^2$

$y = \beta_0 + \beta_1 cgpa + \beta_2 iq$

$\boxed{\sigma^2}$

θ

Derivation

$(a+b)^2$

$$mse = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} = E[(y - \hat{y})^2]$$

$$(\underbrace{\theta - \hat{\theta}}_{a} + \underbrace{\epsilon}_{b})^2$$

$$E[(\underbrace{\theta + \overset{b}{\epsilon} - \hat{\theta}}_{a})^2]$$

$y = f(x) + \epsilon = \theta + \epsilon$

$\hat{y} = f'(x) = \hat{\theta}$

$\theta - \hat{\theta} + \epsilon$

ϵ eplison irreducible

$$= E[\underbrace{(\theta - \hat{\theta})^2}_{a^2} + \underbrace{\epsilon^2}_{b^2} + 2\epsilon(\theta - \hat{\theta})]$$

$E[x+y]$

$= E[x] + E[y]$

$$= E[(\theta - \hat{\theta})^2] + E[\epsilon^2] + E[2\epsilon(\theta - \hat{\theta})]$$

$$= E[(\theta - \hat{\theta})^2] + E[\epsilon^2] + E[2]\, E[\epsilon]\, E[\theta - \hat{\theta}]$$

$E[xy]$

$E[x]E[y]$

$\to E[\epsilon] = 0$

$$E[(\theta - \hat{\theta})^2] + E[\epsilon^2] \leftarrow$$

$E[XY]$

$E[X]E[Y]$

$$\boxed{mse} = \underbrace{E[(\theta-\hat{\theta})^2]} + \underbrace{E[\epsilon^2]}$$

$$E[\epsilon^2] \rightarrow \quad var(\epsilon) = \sigma^2 = E[(\epsilon - E[\epsilon])^2]$$

$E[\epsilon]\rightarrow 0$

$$E[\epsilon^2] = var(\epsilon) \quad = E[(\epsilon-0)^2] = E[\epsilon^2]$$

$$mse = \underbrace{E[(\theta-\hat{\theta})^2]} + \boxed{var(\epsilon)}$$

irreducible

$$E[(\theta-\hat{\theta})^2] = E[(\overbrace{\theta - E[\hat{\theta}]}^{a} + \overbrace{E[\hat{\theta}] - \hat{\theta}}^{b})^2]$$

$$E\left[(\theta-E[\hat{\theta}])^2 + \underbrace{(E[\hat{\theta}]-\hat{\theta})^2}_{b^2} + 2(\theta-E[\hat{\theta}])(E[\hat{\theta}]-\hat{\theta})\right]$$

$E[XY] = E[X]E[Y]$

$$\underbrace{E[(\theta-E[\hat{\theta}])^2]} + \underbrace{E[(E[\hat{\theta}]-\hat{\theta})^2]} + \boxed{E[2(\theta-E[\hat{\theta}])(E[\hat{\theta}]-\hat{\theta})]} = 0$$

$$E[2]\,E[(\theta-E[\hat{\theta}])]\,E[(E[\hat{\theta}]-\hat{\theta})]$$

$$2(\theta-E[\hat{\theta}])\{E[E[\hat{\theta}]]-E[\hat{\theta}]\}$$

$$E[\hat{\theta}]-E[\hat{\theta}] \qquad 0$$

$$\theta - E[\hat{\theta}]$$

$$E[(\theta-E[\hat{\theta}])^2] + \underbrace{E[(E[\hat{\theta}]-\hat{\theta})^2]}_{var}$$

$$(\theta - E[\hat{\theta}])^2$$

$$\rightarrow (\theta - E[\hat{\theta}])^2$$

$$(Bias)^2 + var$$

$$y = \theta + error$$

noise

$$mse = \underbrace{(bias)^2 + \boxed{variance}} + \text{irreducible } var(\epsilon)$$

reducible error

Bias_variance_decomposition

$$mse = \text{reducible} + \text{irreducible}$$

$$\rightarrow bias + var$$

cgpa	iq	lpa	0.1
8	80	8 —	8.1 → 0.05
7	70	7 —	6.9 0.1

0.05

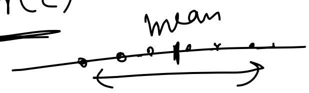
0.04

var(ε)

→ bias + var

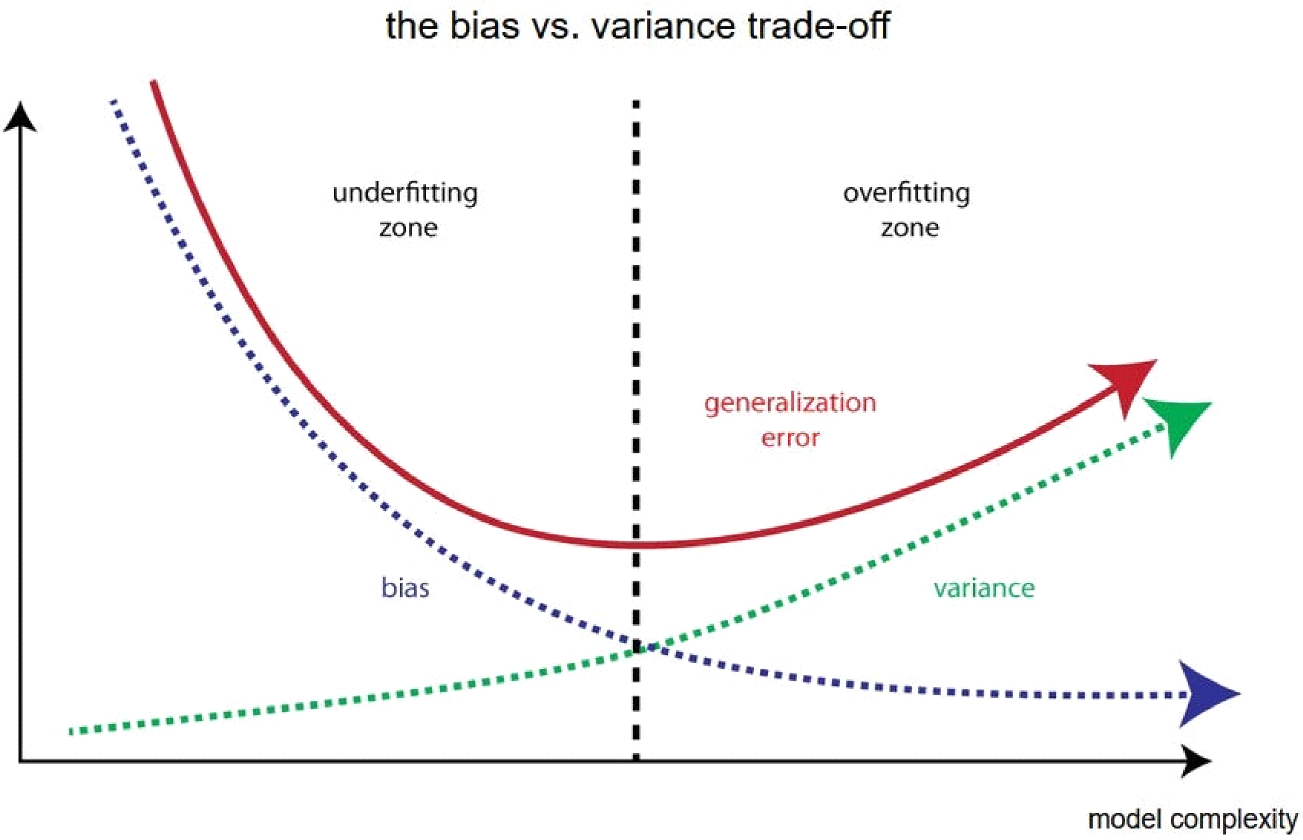
→ 7 70 7 — 6.1 0.04

Var(ε)



Diagram

17 May 2023 17:52



Analogy

17 May 2023 14:44

irreducible

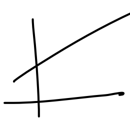
$$mse = (bias)^2 + variance + \boxed{variance(\epsilon) \neq 0} \qquad y = f(x)$$

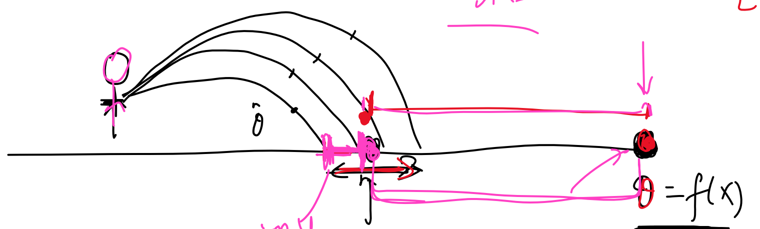
multiple shots

bias

$$E[\hat{\theta}] - \theta = bias$$

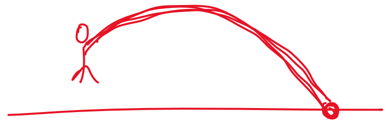
var on avg mean





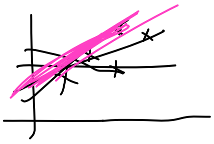
$\theta = f(x)$

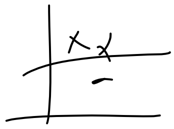
$\hat{\theta} = f(x)$



variance

variance



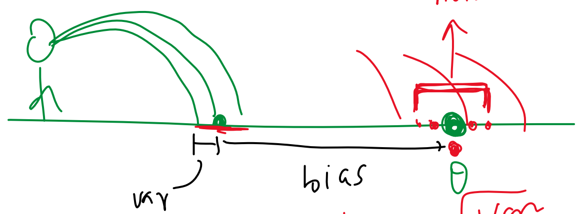


$Var(\hat{\theta})$ $\quad E[\hat{\theta}]$

noise

$var(\epsilon) = 0$

$\boxed{loss = 0}$

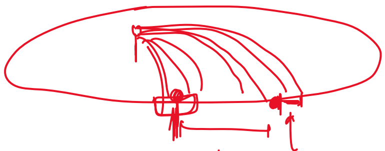


t

$$(y - \hat{y})^2 = (bias)^2 + var$$

bias + 2√var

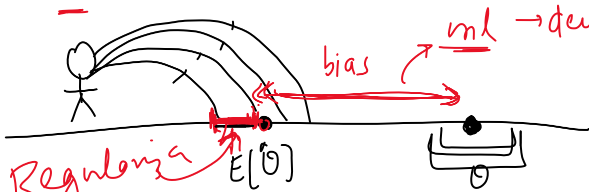
std



loss → bias, var, target

ml → decision $(y - \hat{y})$

bias

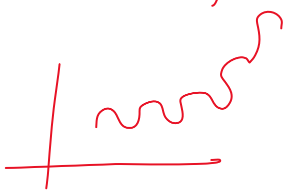


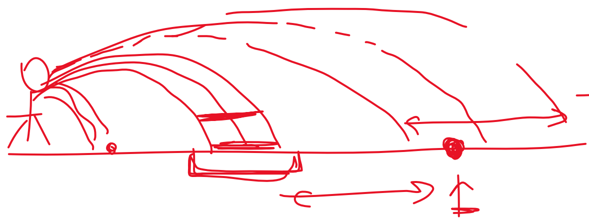
Regularization

Bias variance

bias

polynomial degree





$\boxed{\text{Regularise}}$

reduce variance

↳ reduce overfitting

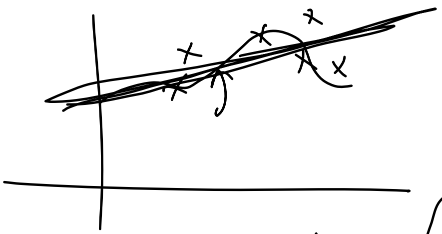
Code Example

17 May 2023 07:17

bias vari dec

bias vari

$$\text{loss} = \underbrace{(\text{bias})^2}_{} + \underbrace{\text{var}}_{} + \underbrace{\text{some error}}_{}$$



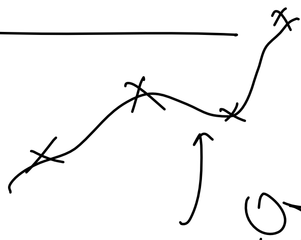
$(\text{bias})^2 + (\text{var}) \leftarrow$ technique

complex models

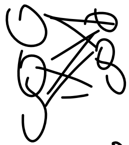
overfitting

Regularization

decision



linear → poly
decision →
neural



100 days of ML

Intro to reg

Ridge
LASSO
Elastic

2 session intro to reg

Smoth

SMOTE → imbal.

7 videos → Monday

3 hrs

→ Doubt clear