Regularization

Hidden truth?

In ML we try to find relationship for $x \rightarrow y$ for all universal points of X to predict y with given only sample data points.

$\therefore$ $y = f(x) + \underline{error}$

error → irreducible error ⇓ can't be manageable

Reducible error ⇓ $f(x) - \hat{f}(x)$ ⇓ we can reduce it.

Bias variance Tradeoff

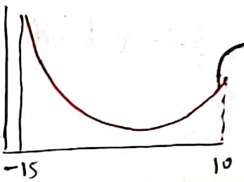
↳ Closely related with reducible error

Mostly in ML we have sample data → we predict population data.

[but]

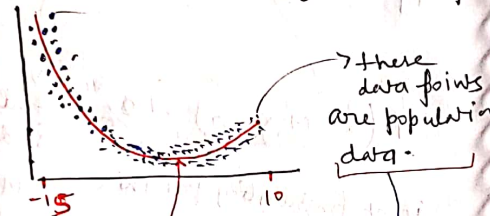
~~let's take we trained model for population data~~

let's $f(x) = x^2$ $[-15, 10]$ → population function.



$f(x) = x^2$ is function we chosen.

let's take 1000 points as population data {generally we never know population data & we Here, taking to know y}



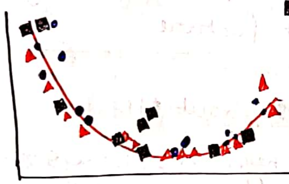
→ these data points are population data.

population data eqn as $y = x^2 + \underline{error}$

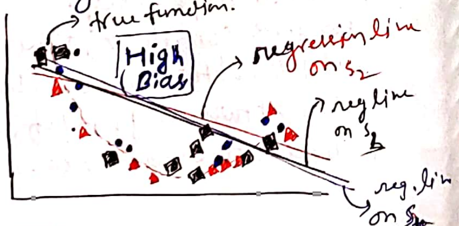
$[-15, 10]$

error → distance or diff. b/w x^2 & y.

Let's we took 3 random sample from population data ● → S1, ▲ → S2, ■ → S3



Now, we given these above 3 samples of data to 3 people without giving knowledge of actual (x^2) line eqn. below is Model (regression line) fitted by them.



Bias:- the inability of a ML Model to fit the training data.

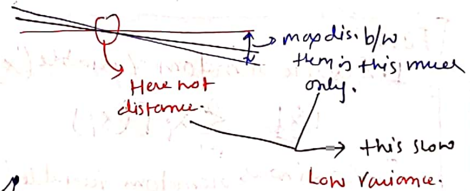
~~Variance~~

All above three model is much much away from training data set → High bias.

Variance:- How correctly ML model or changes it's prediction when training data changes.

↳ Here change means unseen data or test data.

Since, reg. line of all three Model for true sample data from same population data. is overlapping to each other it means → Low Variance.



∵ Here distance b/w 3 model reg. line has been increased s.c. Called → High Variance. & Low bias

~~High Var / Low bias → overfitting (∵ performance measured on training data)~~

~~High bias → underfitting (∵ Not performing on training data)~~

High variance → overfitting (∵ Not performing well on test data. ∴ variance b/w test sample's model line is high)

High bias → underfitting (∵ Not performing well on train data, as shown After training on 3 samples as train data it is away from actual pop. curve)

What is meaning of "trad-off"?

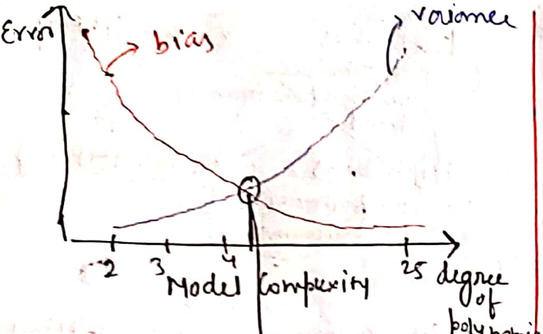
It refers to the fact, that minimizing bias will usually increase variance and vice versa.

Variance: Means test pe error ka decrease karo to train pe bias High ho jata hai & vice versa

Means:-

bias-variance trade off:

variance ↓ → bias ↑
Variance ↑ ← bias ↓



as data ~~ba~~ scientist we have to find this point.

Some question

Q1. How would you define bias & variance mathematically?

Q2. How is bias & variance related to overfitting & ~~variance~~ underfitting mathematically

Q3. why is there a tradeoff b/w bias & variance mathematically?

***

Expected value & Variance

→ represents the average outcome of random variable over a large number of trails or experiments.

⇓

in simple word

⇓

Ex:- If we toss dice 1 lakh times want to find avg. value of it's outcom → is it deterministic?

→ yes.

⇓

it's average would be 3.5.

X := rolling dice → 1 lag times.

$E(X) = x_1 P(x_1) + x_2 P(x_2) + \dots + x_n P(x_n)$

$= 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + \dots 6\times\frac{1}{6}$

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

~~Here~~

Term "over a large number of trails or experiments"

⇓

same as

Population

⇓

∴ Expected value → same as population mean.

For

1. Descrete random variable (x)

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

2. Continous random variable (x)

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

~~Variance~~

Variance of population $(Var(X))$

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

↳ its proof

As we know that

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

$\bar{x}$ → sample mean.

represent all data point

$=$

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

$= Var(X)$

$x_i \rightarrow X$

$\bar{X} \rightarrow E[X]$

∴ pop. mean is same as $E[X]$

∵ $\sum \frac{1}{n} \rightarrow E[X]$

Let's expand it.

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

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

∵ $E[X+Y] = E[X] + E[Y]$

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

∵ $E[XY] = E[X]\, E[Y]$ give X & Y independent

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

∵ $E[\text{const.}] = \text{const.}$ & $E[X] \rightarrow$ constant

∵ $E[E[X]] = E[X]$

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

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

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

***

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

↓

works well for descrete & continuous.

What exactly are Bias and Variance Mathematically?

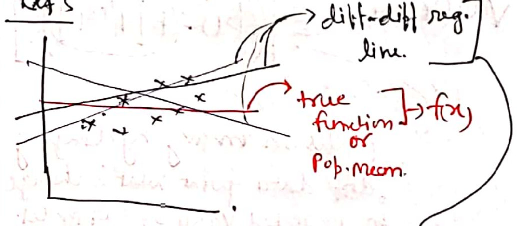
Bias:-

↳ 1. Systematic error that a model introduces because it cannot capture true relationship in the data.

*** 2. diff (expected Pred. of our model, Correct value trying to predict)

3. ↑ bias → underfit.

Let's



***

$f'(x), f'(x), f''(x)$

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

↳ ~~when we means~~ If we make ML model and run over ~~go~~ again & again on diff sample data → find ~~the~~ expected value for each sample's outcome, it → $f(x)$

***

$$\boxed{\text{Reducible error} = \text{bias} + \text{variance}}$$

Scanned with OKEN Scanner

Variance :-

1. refers to the amount by which the prediction of our model will change if we used a diff-diff training set (or test)
2. Prediction for a given point vary b/w diff realization of the model

$$Var(f'(x)) = E\left[\left(f'(x) - E[f'(x)]\right)^2\right]$$

unseen data point (test data)

Expected mean after giving diff-diff sample on Model.

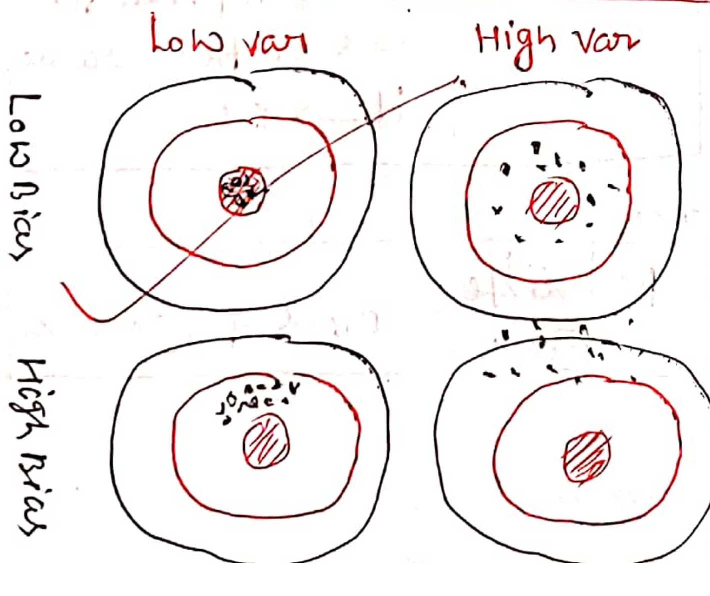
It tells on an avg by changing diff data point what is change in expected value of our model.

It came from $Var = \frac{\sum (x_i - \bar{x})^2}{n}$

***

$$\therefore \; Bias(f'(x)) = E[f'(x)] - f(x)$$

$$Variance(f'(x)) = E\left[\left(f'(x) - E[f'(x)]\right)^2\right]$$



* Bias variance decomposition

$$\boxed{Loss = bias + Variance + irreducible\ error.}$$

further simplified as

$$Loss = \underbrace{bias^2 + variance}_{reducible\ Error.} + \underbrace{Var(\epsilon)}_{irr.\ Error}$$

Let's

x_1 x_2 ...	y	$\hat{y}$	Error
	8	9	1
	8	8.1	0.1
	7	6.9	0.1
	9	10.1	1.1

Error: reducible / irreducible.

Let's assume (assumption)

$mean_{irreducible} = 0$

$Variance_{irreducible} = \sigma^2$ (const)

spread of irreducible Error

Let's take Loss function as MSE

$$\therefore MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}$$

$$= E\left[(y - \hat{y})^2\right]$$ { we can write as in terms of expected val. }

Scanned with OKEN Scanner