Redesigning Common Mind & Business Towards Excellence

3P Purpose Process People Culture

SNS INSTITUTIONS www.snsgroups.com

1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

II YEAR IV SEM

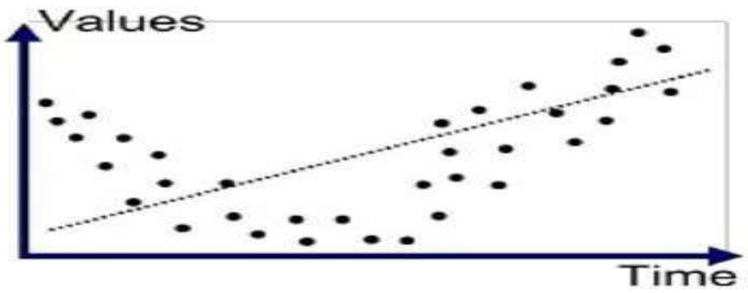
UNIT II – SUPERVISED LEARNING ALGORITHM

TOPIC 1 – Bias and variance Overfitting Underfitting

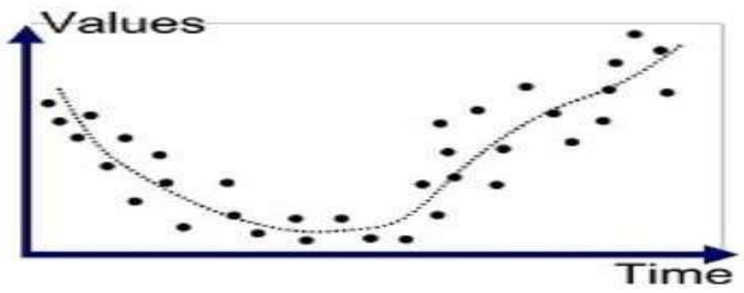
Introduction to under fitting and over fitting and best fit

Regression problem: Underfitting, overfitting, best fitting

1. **Underfitting:** A model or a ML algorithm is said to have underfitting when it cannot capture the underlying trend of the data.
2. **Overfitting:** A model or a ML algorithm is said to have overfitting when it capture the underlying trend of the data very accurately.
3. **Best fit:** A model or a ML algorithm is said to have best fit when it capture the underlying trend of the data moderately.



Underfitted



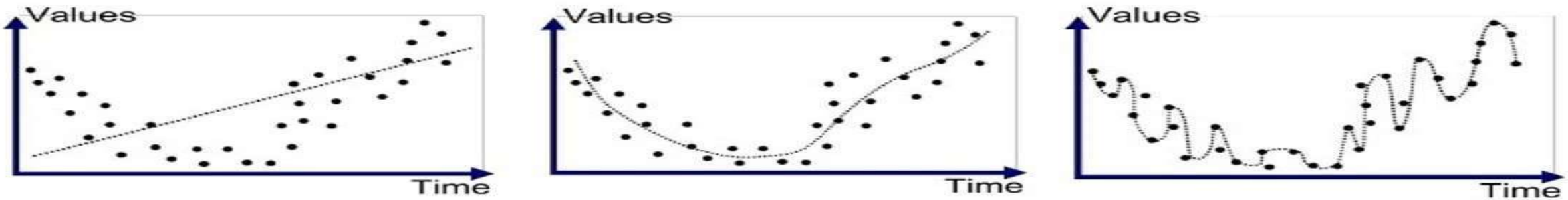
Good Fit/Robust

Overfitted

Bias and variance

1. **Bias:** It shows the degree of randomness of the training data.
   (a) Based on the training data a suitable model may be created for the regression or classification problem.
   (b) Regression: Linear, logistic, polynomial
   (c) Classification: Decision tree, random forest, naive baye's, KNN, SVM
2. **Variance:** It shows the degree of randomness of the testing data.
   (i) Testing data validates the accuracy of a model, that has been made with the help of training data set.
   (ii) Testing data is nothing but the unlabeled or unknown data.

Underfitting, overfitting, best fit and bias, variance



Note:

⇒ The objective of ML algorithm not only fit for the training data but also fit for the testing data.

⇒ In other words, low bias and low variance is the appropriate solution.

Underfitting	Overfitting	Best fit
High bias	Very low bias	Low bias
High variance	High variance	Low variance

Underfitting, overfitting, and best fit in classification problem

A classifier (Decision tree, random forest, naive baye's, KNN, SVM) works on two types of data

- Training data Testing
- data

Example:

Classifier comes under the category of underfitting, overfitting, and best fit based on its training and testing accuracy.

Underfitting	Overfitting	Best fit
Train error=25%	Train error=1%	Train error=8%
Test error= 27%	Test error= 23%	Test error=9%

Mathematical intuition

$$bias\ \hat{f}(x) = E[\hat{f}(x)] - f(x) \quad (1)$$

$$variance\ \hat{f}(x) = E\left[\hat{f}(x) - E[\hat{f}(x)]\right]^2 \quad (2)$$

⇒ $\hat{f}(x) \rightarrow$ output observed through the training model

⇒ For linear model $\hat{f}(x) = w_1 x + w_0$

⇒ For complex model $\hat{f}(x) = \sum_{i=1}^{p} w_i x^i + w_0$

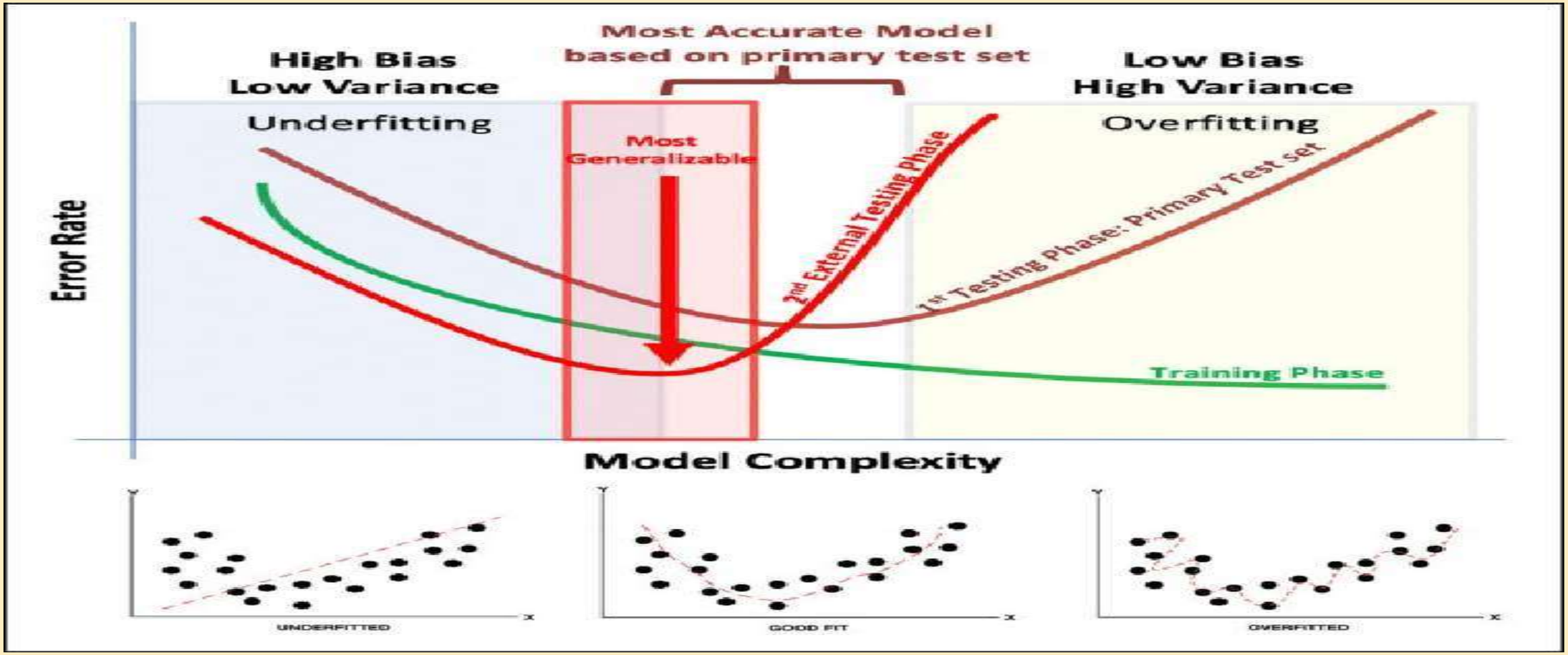
⇒ We don't have idea regarding the true $f(x)$.

⇒ **Simple model:** Low bias & high variance

⇒ **Complex model:** High bias & low variance

$$E[(y - \hat{f}(x))^2] = bias^2 + Variance + \sigma^2\ (Irreducible\ error) \quad (3)$$

as and variance trade-off relation



References

E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

J. Grus, *Data science from scratch: first principles with python*. O'Reilly Media, 2019.

T. M. Mitchell, *The discipline of machine learning*. Carnegie Mellon University, School of Computer Science, Machine Learning , 2006, vol. 9.

1. https://www.youtube.com/watch?v=iNmtkaeGxLg
   https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/
2. https://www.geeksforgeeks.org/linear-regression-python-implementation/?ref=lbp