ADVANCED DATA SCIENCE

Lecture 3: Data Preprocessing, Bias–Variance Trade off

By

Dr. Sumit Kumar Singh

Associate Professor

SR University

WHAT IS DATA PREPROCESSING?

- Data preprocessing prepares raw data for analysis.

Key steps:

- Handling missing data
- Noise removal
- Normalization
- Encoding categorical data
- Data splitting

NOISE IN DATA

Noise = random variation in data caused by:

- Sensor or measurement errors
- Transmission faults
- Human entry errors

Impact: reduces model accuracy and reliability.

Noise = Random, meaningless variation in data that hides the true pattern.
It's not an error in the data structure — it's **unwanted fluctuations** that reduce model accuracy.

Example:

Student	Age	Marks	Attendance
A	20	80	0.9
B	21	?	0.85
C	19	85	0.8
D	25	105	0.95

- Missing value (“?”)
- Wrong data (Marks = 105)

NOISE REMOVAL TECHNIQUES

Common approaches:

- Binning or smoothing
- Moving averages
- Regression-based filtering
- Outlier detection

We’ll use one common dataset for explanation:

Hour	Temperature (°C)
1	30
2	31
3	32
4	31
5	29
6	28
7	32
8	33
9	34
10	35

Binning (Smoothing by Mean/Median)

- Group data into bins and replace values with **bin mean**.
  - **Example:**
    Bins of size 3 →
    [30, 31, 32] → mean = 31
    [31, 29, 28] → mean = 29.3
    [32, 33, 34] → mean = 33
    [35] → mean = 35
  - Smoothed data: [31,31,31,29.3,29.3,29.3,33,33,33,35]
  - Binning reduces random variation and reveals general trends.

Moving Average (Time Series Smoothing)

Formula:

$$Y_i' = \frac{Y_{i-1} + Y_i + Y_{i+1}}{3}$$

Example:

At hour 4:

$$Y_4' = \frac{32 + 31 + 29}{3} = 30.7$$

Smooths the curve by averaging neighbors.

NORMALIZATION

➢ Normalization ensures features are on the same scale.

1. Min–Max Scaling: X' = $\frac{(X - X_{min})}{(X_{max} - X_{min})}$

2. Z-score Scaling: Z = (X − μ) / σ

Outlier Detection (Z-Score Method)

Outliers = extreme values far from the mean.
Z-score formula:

$$Z = \frac{X - \mu}{\sigma}$$

If $|Z| > 3$ → outlier.

Example:
Mean = 31.5, SD = 2.0
For X = 36: $Z = (36 - 31.5)/2 = 2.25$ →borderline high.

EXAMPLE: BEFORE AND AFTER NORMALIZATION

Person	Age (years)	Salary (₹)	Experience (years)
A	22	25,000	1
B	35	60,000	8
C	47	85,000	12
D	52	120,000	20

- Notice how the **Salary** column is much larger in scale (thousands), while **Age** and **Experience** are small.
- Machine learning models using distance will give more weight to Salary.

Apply Min–Max Normalization

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

This scales all values between **0 and 1**

Person	Age	Salary	Experience	**Age'**	**Salary'**	**Experience'**
A	22	25,000	1	0.00	0.00	0.00
B	35	60,000	8	0.43	0.35	0.37
C	47	85,000	12	0.83	0.60	0.58
D	52	120,000	20	1.00	1.00	1.00

All features are now on the **same scale (0–1)**.

BIAS–VARIANCE TRADEOFF

- When we train a machine learning model (like Linear Regression, Decision Tree, etc.), we want the model to:
- Fit the training data well work accurately on new (unseen) data

But achieving both is *hard* — because models can make **two types of errors**:

1. **Bias error** → comes from **wrong assumptions** (model too simple)

1. **Variance error** → comes from **too much sensitivity** to training data (model too complex)

Together, they form the **Bias–Variance Trade-off**.

- Bias: Error from overly simple model (underfitting)
- Variance: Error from overly complex model (overfitting)

Goal: Find balance to minimize total error.

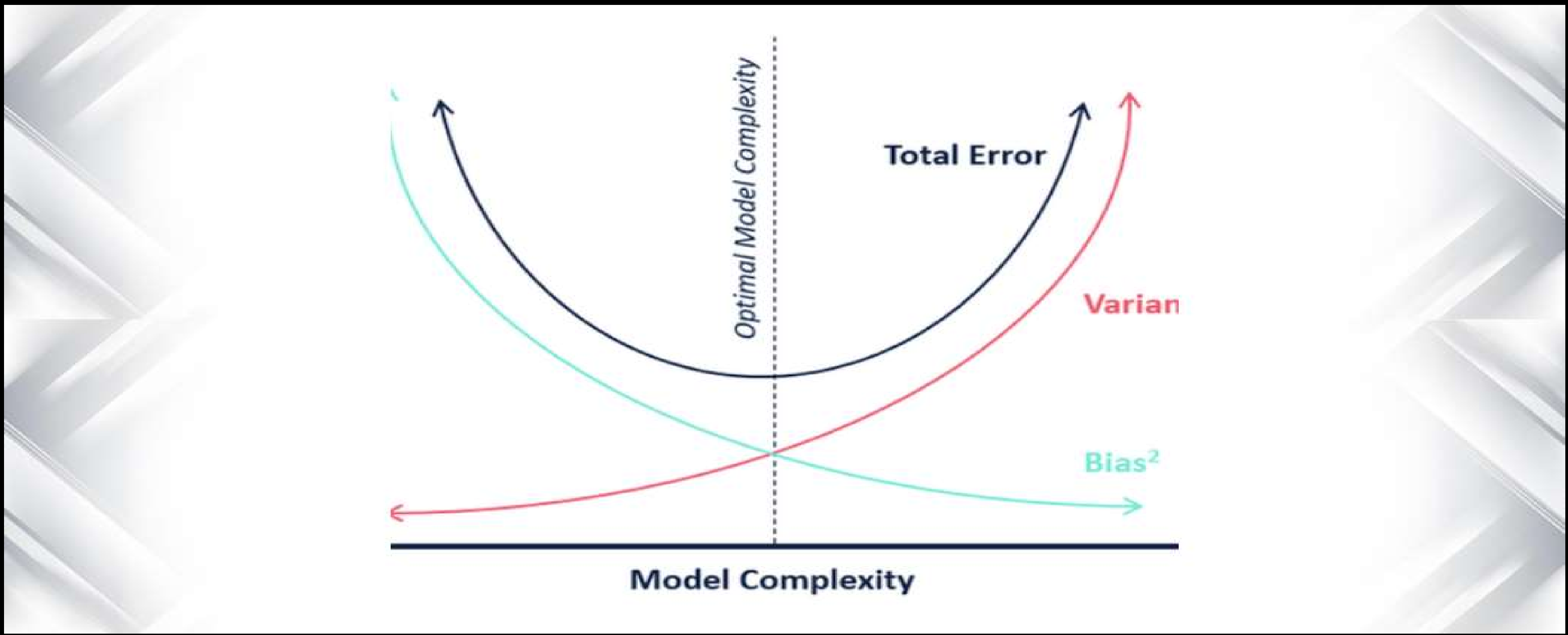
Input (X)	True Y	Predicted Y
1	2	1.2
2	4	2.3
3	9	3.4
4	16	4.2

- The model's predictions are always lower than the true values.
  That's **high bias** — the model is **too simple**.

TOTAL ERROR FORMULA

- $E[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$

- High Bias → underfitting
- High Variance → overfitting
- Irreducible Error → inherent randomness



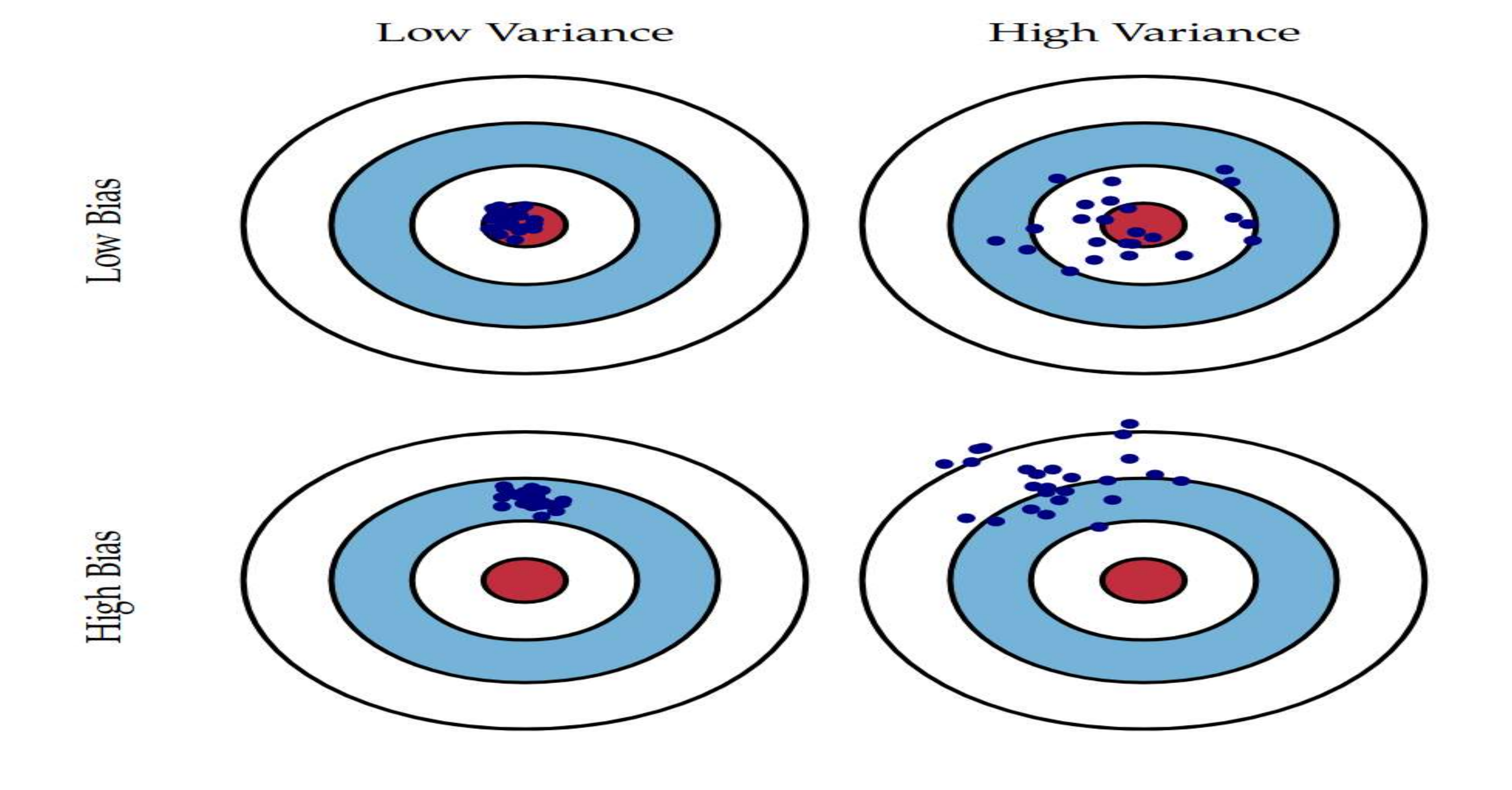


Fig. 1 Graphical illustration of bias and variance.