MA477: Data Science

Lesson Outline — 8 January 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

---

1 Administrative

- Linking Notebook (well use this later)
- Problem Set 1
- Gen AI Quiz

2 Student Review

3 Key Lesson Topics

- prediction vs. inference
- reducible vs. irreducible error
- bias (underfitting) vs variance (overfitting)
- flexibility vs. Interpretability
- parametric vs. non-parametric

Table 1: Conceptual Comparison of Parametric and Non-Parametric Models

Characteristic	**Parametric Models**	**Non-Parametric Models**
Model form	Fixed functional form	No fixed functional form
Number of parameters	Fixed; does not grow with data	Grows with data or model complexity
Assumptions	Strong assumptions about structure or distribution	Minimal assumptions about data
Flexibility	Limited	High
Data requirements	Lower	Higher
Bias–variance tendency	Higher bias, lower variance	Lower bias, higher variance
Interpretability	Often high	Varies; often lower
Computational cost	Typically low	Often higher
Typical modeling question	"Does a simple relationship explain the data?"	"What structure does the data suggest?"

Table 2: Examples of Parametric and Non-Parametric Models

Model	**Category**	**Reason for Classification**
Linear Regression	Parametric	Assumes a linear relationship with a fixed number of coefficients
Logistic Regression	Parametric	Uses a fixed sigmoid function with a fixed parameter set
Linear Discriminant Analysis (LDA)	Parametric	Assumes class-conditional normal distributions
Naive Bayes	Parametric	Assumes specific probability distributions and conditional independence
k-Nearest Neighbors (k-NN)	Non-Parametric	No explicit model; predictions depend directly on stored data
Decision Trees	Non-Parametric	Tree structure adapts to data with no fixed size
Random Forests	Non-Parametric	Ensemble of adaptive decision trees
Kernel Density Estimation (KDE)	Non-Parametric	Estimates density directly from data without distributional assumptions
Gaussian Processes	Non-Parametric	Places a distribution over functions rather than parameters
LOESS / LOWESS	Non-Parametric	Fits local models whose form depends on nearby data
Neural Networks	Semi-Parametric	Fixed architecture but expressive power grows with data and training

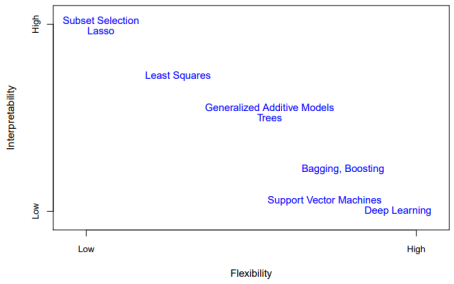

FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Figure 1

4 Bias–Variance–Noise Decomposition

This section presents a step-by-step algebraic derivation of the bias–variance–noise decomposition for squared prediction error. The goal is to understand how the expected error of a learning algorithm separates into distinct, interpretable components.

Step 1: Data-Generating Process and Prediction Error

Assume the observed response is generated according to

$$y = f(x) + \varepsilon,$$

where $f(x)$ is the true (unknown) function and ε is a random noise term. We assume

$$\mathbb{E}[\varepsilon \mid x] = 0, \qquad \mathrm{Var}(\varepsilon \mid x) = \sigma^2.$$

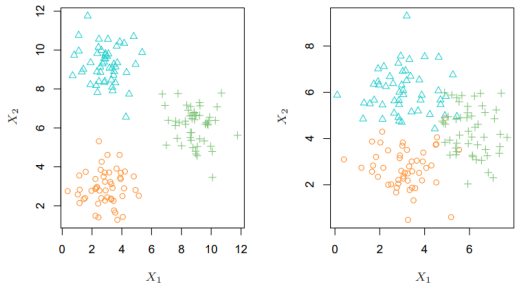

FIGURE 2.8. *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

Figure 2

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon). \tag{2.7}$$

Figure 3

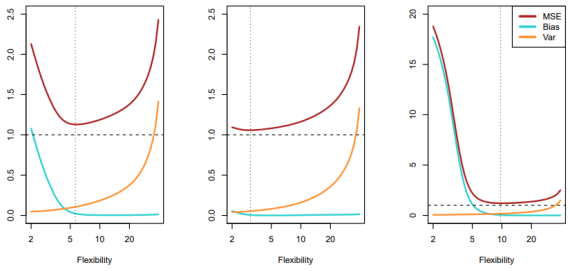

FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

Figure 4

Let $\hat{f}(x)$ denote the prediction produced by a learning algorithm trained on a random dataset. Because the training data are random, $\hat{f}(x)$ is itself a random variable.

The quantity of interest is the expected squared prediction error at a fixed input x:

$$\mathbb{E}\big[(y - \hat{f}(x))^2 \mid x\big].$$

—

Step 2: Substitute the Data Model and Expand the Square

Substituting $y = f(x) + \varepsilon$ into the error expression gives

$$\mathbb{E}\big[(f(x) + \varepsilon - \hat{f}(x))^2 \mid x\big].$$

Rewriting the terms,

$$\mathbb{E}\big[(f(x) - \hat{f}(x) + \varepsilon)^2 \mid x\big].$$

Expanding the square yields

$$\mathbb{E}\big[(f(x) - \hat{f}(x))^2 + 2\varepsilon(f(x) - \hat{f}(x)) + \varepsilon^2 \mid x\big].$$

By linearity of expectation, this separates into three terms:

$$\mathbb{E}\big[(f(x) - \hat{f}(x))^2 \mid x\big] + 2\mathbb{E}\big[\varepsilon(f(x) - \hat{f}(x)) \mid x\big] + \mathbb{E}\big[\varepsilon^2 \mid x\big].$$

—

Step 3: Eliminate the Cross Term and Isolate Noise

The middle term vanishes because the noise has mean zero and is independent of the trained model:

$$\mathbb{E}[\varepsilon(f(x) - \hat{f}(x)) \mid x] = 0.$$

The final term is the noise variance:

$$\mathbb{E}[\varepsilon^2 \mid x] = \sigma^2.$$

Thus, the expected error reduces to

$$\mathbb{E}\big[(y - \hat{f}(x))^2 \mid x\big] = \mathbb{E}\big[(f(x) - \hat{f}(x))^2 \mid x\big] + \sigma^2.$$

The remaining term represents model-dependent error.

—

Step 4: Decompose Model Error into Bias and Variance

Define the average prediction across training datasets as

$$\bar{f}(x) = \mathbb{E}_D[\hat{f}(x)].$$

Add and subtract $\bar{f}(x)$ inside the squared term:

$$f(x) - \hat{f}(x) = \big(f(x) - \bar{f}(x)\big) + \big(\bar{f}(x) - \hat{f}(x)\big).$$

Squaring and expanding,

$$(f(x) - \hat{f}(x))^2 = (f(x) - \bar{f}(x))^2 + 2(f(x) - \bar{f}(x))(\bar{f}(x) - \hat{f}(x)) + (\bar{f}(x) - \hat{f}(x))^2.$$

Taking expectation over datasets, the cross term vanishes because

$$\mathbb{E}_D[\hat{f}(x)] = \bar{f}(x).$$

Therefore,

$$\mathbb{E}_D[(f(x) - \hat{f}(x))^2] = (f(x) - \bar{f}(x))^2 + \mathbb{E}_D[(\hat{f}(x) - \bar{f}(x))^2].$$

The first term is the squared bias, and the second term is the variance of the model.

—

Step 5: Final Bias–Variance–Noise Decomposition

Substituting this result back into the expression from Step 3 gives

$$\mathbb{E}\big[(y - \hat{f}(x))^2 \mid x\big] = \underbrace{(\bar{f}(x) - f(x))^2}_{\text{Bias}^2} + \underbrace{\mathrm{Var}_D(\hat{f}(x))}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Noise}}.$$

This decomposition shows that the expected prediction error is the sum of:

- **Bias squared**, measuring systematic error due to model assumptions;
- **Variance**, measuring sensitivity to the training data;
- **Irreducible noise**, representing randomness inherent in the data.