MA289 Mathematics of AI
Lesson Outline — 6 January 2026
United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Introductions

2 Demo of board briefing

- First name or Appropriate Nick name
- Hometown - Houston, TX
- Academic History - PhD: Duke, Masters: Harvard, Bachelors: USMA
- Service History - Armor Officer, last Army job I was an Infantry BN XO, Deployed 4x to Iraq / Afghanistan
- Family - 5 kids: 4-19 years old, eldest just started college, we live in Cold Spring, NY across the river

Figure 1: My Family

2.1 Cadet Introductions

Modified Name Game - Write On the board, brief yours and everyone before you, starting from your right:

- Name and company

- Academic major
- Personal Interest or Hobby

3 Understanding Statistical Learning

3.1 What Is Statistical Learning?

Go to the board and tell us what Statistical Learning means for you.

3.2 Roles of Data Science in the Army

Discuss how you have heard, experienced, or guessed Data Science is used in the Army or somewhere else in life.

4 Understanding Statistical Learning / Machine Learning

Record your thoughts on these questions:

- Describe some of the key ways we distinguish between statistical learning techniques
- What are some challenges with statistical learning
- What previous concepts which you have learned at USMA will be used?

5 Linear Algebra Warm-Up (By Hand)

Purpose: Linear algebra is the language of data science. Today we focus on intuition, structure, and meaning — not proofs or computation speed.

5.1 Matrix Dimensions

Given the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

- How many rows does A have?
- How many columns does A have?
- What is the dimension of A?

Check: What dimensions must another matrix have so that multiplication with A is valid?

5.2 Matrix Multiplication: 3×2 times 2×2

Let

$$A = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ -1 & 3 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 1 \\ 0 & -1 \end{bmatrix}$$

- What are the dimensions of A and B?
- Is the product AB defined?
- What will the dimensions of AB be?
- Compute AB by hand.

Board Discussion: Each row of A is transformed by B. What does that suggest about matrix multiplication as a transformation?

5.3 Transpose Example

Given

$$C = \begin{bmatrix} 1 & 4 & -2 \\ 0 & 3 & 5 \end{bmatrix}$$

- What is the dimension of C?
- Compute C^T.
- What are the dimensions of C^T?

Interpretation:

- Rows become columns
- Transpose changes how we interpret inputs vs outputs
- This will matter later for dot products and projections

5.4 Solving for Variables Using Matrix Multiplication

Consider the system

$$\underbrace{\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 5 \\ 3 \end{bmatrix}}_{\mathbf{b}}$$

- What does this matrix equation represent?
- Rewrite it as two scalar equations.

Now suppose we multiply both sides by the inverse of the coefficient matrix:

$$A^{-1}A\mathbf{x} = A^{-1}\mathbf{b}$$

$$\mathbf{x} = A^{-1}\mathbf{b}$$

- Why does this isolate the unknown vector?
- What assumptions must be true for A^{-1} to exist?

Optional (By Hand): Compute A^{-1} and solve for (x, y):

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \quad A^{-1} = \frac{1}{(2)(1) - (1)(1)} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

5.5 Linear Independence: Dependent vs Independent

Example 1: Linearly Dependent Rows

$$D = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

- Is one row a scalar multiple of the other?
- What does this imply about the information content?

Example 2: Linearly Independent Rows

$$E = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}$$

- Can either row be written as a multiple of the other?
- How many independent directions exist?

6 Course Structure and Administration

6.1 Topics Covered

See course power point slide.

- Regression
- Classification
- clustering techniques
- Unsupervised Learning and NLP
- Regularization

6.2 Books, Software, and Tools

- ISLP
- Google Colab and Drive
- Overleaf

6.3 Grades.

Table 1: Course Evaluation and Point Distribution

Graded Event	**Points per Event**	**Number of Events**	**Total Points**
AI Certification Assignment	10	1	10
Lesson Assignments	40	2	80
Projects	300	3	900
Total			**1000**

7 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

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

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.