MA477: Data Science

Lesson 4 Board Sheet — 15 January 2026

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

1 Lesson Objectives

By the end of this lesson, cadets will be able to:

- Understand the steps in completing an end-to-end machine learning project.
- Understand how to implement data pipelines, clean/standardize data, visualize data, how to handle categorical/text
- Understand and know how to perform cross-validation

2 Loss Functions

1. **What is Error**
   How do we decide how "wrong" a single prediction is?
2. **Vectors of Error**
   If a model makes many predictions, how can we combine all those errors into one number?
3. **Most Common Error Metric**
   What is the most commonly used error metric in regression problems, and why is it so widely adopted in practice?
4. **Alternative Error Metric**
   What is another commonly used error metric, and in what situations might it be preferred over the most common choice?
5. **Tradeoffs Between Metrics**
   What are the primary advantages and disadvantages of these two error metrics, particularly with respect to sensitivity to large errors and robustness to outliers?
6. **Analytical Demonstration**
   How can the difference between these two error metrics be demonstrated analytically, for example by comparing their geometric interpretations or their level sets in two dimensions?

3 Limited data discussion

1. With limited data, what types of error can we reduce?
2. Given we are limited to the data available what can we do to improve our model?
3. What are the costs of this method?

Step	Student checklist for a prediction project
1	**Clarify the question:** Define the target, task type (regression vs. classification), and what "good" performance means in context.
2	**EDA (Exploratory Data Analysis):** Load the data, inspect data types, summary statistics, missingness, outliers, class balance, and simple plots/relationships to understand what you have.
3	**Baseline + preprocessing:** Build a simple baseline model; decide if features need **scaling/standardization** (especially for distance/gradient-based models) and encode categorical variables as needed.
4	**Evaluation + regularization:** Choose an appropriate metric/loss; use a validation plan (train/test split or cross-validation); apply **regularization** (e.g., L1/L2) and tune hyperparameters to reduce overfitting.
5	**Overfitting check:** Compare training vs. validation performance; look for instability across folds/splits; simplify model or increase regularization if needed.
6	**Interpret + sanity-check:** Confirm predictions make sense; check feature effects; verify there is no data leakage; summarize limitations and next steps.

A simple visual (5-fold CV)

Legend: **T** = training fold, **V** = validation fold