MA477: Data Science

Lesson 18 Board Sheet — 24 February 2026

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

1 Regression Tree Lesson Objectives

- Understand the basic algorithm for building a regression tree.
- Use functions in Python's sklearn module to fit regression trees.
- Know how a regression tree splits the decision space.
- Understand the advantages and disadvantages of tree-based methods.

2 Discussion Questions: Regression Trees and Pruning

1. **Cost of a Split**

   Once a candidate split (for example, Years < 5) has been chosen, how do we compute error?

2. **Determining Split Points**

   For a continuous predictor variable in a regression tree, how are the candidate split points generated from the observed data?

3. **Reducing Variance in Decision Trees**

   Because a single decision tree can have high variance and be sensitive to small changes in the data, how can we reduce the variance of a tree-based model?

4. **New Dataset: Guided Construction by Hand**

   Consider the dataset:

ID	x	y
1	1	2
2	2	2
3	3	3
4	4	8
5	5	9
6	6	10

   Answer the following step-by-step:

   (a) List all possible split points for x.

   (b) For each candidate split, compute the mean of the left and right regions.

   (c) Compute the RSS for each candidate split.

   (d) Identify the split that minimizes RSS.

   (e) After choosing the best first split, which region would you split next? Why?

   (f) Compute the new leaf means after the second split (use Right Hand Side).

   (g) Compare the RSS before and after the second split. How much did it decrease?

   (h) If $\alpha = 1$, would you keep the second split? Justify using

$$C_\alpha(T) = \mathrm{RSS}(T) + \alpha|T|.$$