MA477: Data Science
Lesson 17 Outline — 24 February 2026
United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Administrative

- Student review
- Work Problem Set 2 due tonight
- Classification comparison discussion

2 Decision Tree Regression: Worked Example with Explicit Split Selection

Algorithm and Loss Function

Two–Step Algorithm (Regression Tree)

1. Partition the predictor space into J disjoint regions
$$R_1, R_2, \ldots, R_J.$$
2. For any observation $x \in R_j$, predict using the region mean:
$$\hat{y}(x) = \hat{y}_{R_j} \quad \text{where} \quad \hat{y}_{R_j} = \frac{1}{n_j} \sum_{i:x_i \in R_j} y_i.$$

Loss Function (Residual Sum of Squares)

$$\text{RSS}(T) = \sum_{j=1}^{J} \sum_{i:x_i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2.$$

The tree is built by choosing splits that minimize RSS.

Simple Baseball Dataset

ID	Years	Hits	$y = \log(\text{Salary})$
A	2	80	5.1
B	3	60	5.0
C	4	120	5.2
D	6	90	6.0
E	7	130	6.6
F	8	160	6.8

Question 1: Determine the First Split Explicitly

We determine the best first split by evaluating all candidate splits for the variable **Years**.

Possible midpoints:

$$2.5,\ 3.5,\ 5,\ 6.5,\ 7.5$$

Evaluate Split at Years < 5

Left: A, B, C

$$\hat{y}_L = \frac{5.1 + 5.0 + 5.2}{3} = 5.1$$

$$\text{RSS}_L = 0.02$$

Right: D, E, F

$$\hat{y}_R = \frac{6.0 + 6.6 + 6.8}{3} = 6.4667$$

$$\text{RSS}_R \approx 0.3467$$

$$\text{Total RSS} = 0.3667$$

Evaluate Split at Years < 3.5

Left: A, B

$$\hat{y}_L = 5.05$$

$$\text{RSS}_L = (5.1 - 5.05)^2 + (5.0 - 5.05)^2 = 0.005$$

Right: C, D, E, F

$$\hat{y}_R = \frac{5.2 + 6.0 + 6.6 + 6.8}{4} = 6.15$$

$$\text{RSS}_R \approx 1.61$$

$$\text{Total RSS} \approx 1.615$$

The split at Years < 5 produces a much smaller RSS.

$$\boxed{\text{Best first split: Years} < 5}$$

Complete Depth-2 Tree

Now split the right node on Hits < 110.

Right-left: D

$$\hat{y} = 6.0, \quad \text{RSS} = 0$$

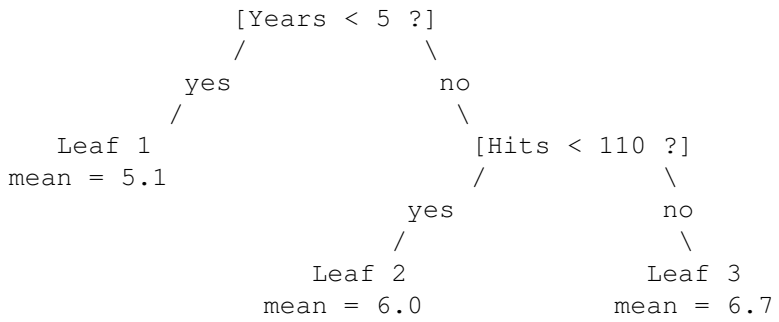
Right-right: E, F

$$\hat{y} = 6.7, \quad \text{RSS} = 0.02$$

Total RSS:

$$0.02 + 0 + 0.02 = 0.04$$

Tree Structure



Pruning: Cost-Complexity

We minimize

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

Question 2: Prune with $\alpha = 0.35$

Tree A (3 leaves):

$$C_{0.35}(A) = 0.04 + 0.35(3) = 1.09$$

Tree B (2 leaves):

$$C_{0.35}(B) = 0.3667 + 0.35(2) = 1.0667$$

Tree C (1 leaf):

$$C_{0.35}(C) = 3.1685 + 0.35(1) = 3.5185$$

Select Tree B

Question 3: New Dataset

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

Best first split:

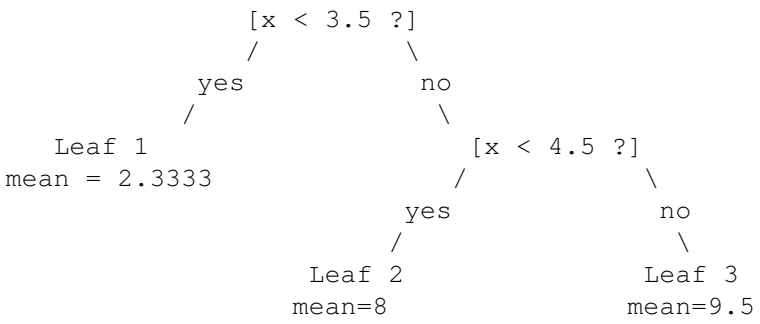
$$x < 3.5$$

Second split:

$$x < 4.5$$

Final RSS:

$$1.1666$$



Key Questions and Answers: Regression Trees

1. **Cost of a Split**

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

   Answer: After splitting the data into two regions R_1 and R_2, we compute the mean response in each region:

   $$\hat{y}_{R_j} = \frac{1}{n_j} \sum_{i:x_i \in R_j} y_i.$$

   The cost of the split is the total residual sum of squares (RSS):

   $$\text{RSS}_{\text{split}} = \sum_{i:x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2} (y_i - \hat{y}_{R_2})^2.$$

   To evaluate whether the split improves the model, we compare this value to the RSS of the parent node. The reduction in RSS measures the improvement created by the split.

2. **Determining Split Points**

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

   Answer: First, sort the observed values of the predictor. Then generate candidate split points at the midpoints between consecutive distinct sorted values:

   $$s_k = \frac{x_{(k)} + x_{(k+1)}}{2}.$$

   Each midpoint represents a distinct way of partitioning the data. For each candidate s_k, we compute the resulting RSS and select the split that minimizes it.

3. **Reducing Variance in Decision Trees**

   Question: 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?

   Answer: Variance can be reduced by averaging many trees. Methods such as **bagging** and **random forests** build multiple trees using bootstrapped samples of the data and average their predictions:

$$\hat{f}_{\text{avg}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{(b)}(x).$$

   Averaging reduces variance because fluctuations specific to individual trees tend to cancel out, resulting in a more stable and accurate predictor.