MA477: Data Science
Lesson 19 Outline — 27 February 2026
United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Administrative

- Student review
- Decision Tree for Classification Discussion
- Exercise on engineering models

2 Decision Tree Classification Lesson Objectives

- Understand how the algorithm for classification trees differs from regression trees.
- Understand and know how to interpret the Gini index.
- Understand and know how to interpret entropy.
- Use functions in Python's sklearn module to fit classification trees.

3 Entropy-Based Splitting and Comparison with Linear Models

Basic Setup

Suppose we have a binary classification problem with response

$$Y \in \{\text{Yes}, \text{No}\}.$$

For a node m, let

$$\hat{p}_m = \text{proportion of Yes observations in node } m.$$

The parent node entropy is

$$H(\text{parent}) = -\left(\hat{p}\log_2\hat{p} + (1-\hat{p})\log_2(1-\hat{p})\right).$$

For a candidate split that produces left and right nodes, the post-split entropy is

$$H(\text{split}) = \frac{n_L}{n}H_L + \frac{n_R}{n}H_R,$$

where H_L and H_R are the entropies of the child nodes.
The **information gain** is

$$IG = H(\text{parent}) - H(\text{split}).$$

The split that maximizes IG is selected.

Worked Example

Heart Disease Example Dataset

We consider the following small dataset with:

- **MaxHR_high** (categorical: 1 = high, 0 = low)
- **Age** (real-valued)
- **HD** $\in \{\text{Yes}, \text{No}\}$

ID	MaxHR_high	Age	HD
1	1	44.2	No
2	1	46.7	No
3	1	49.8	No
4	1	52.1	No
5	1	57.4	No
6	1	60.3	Yes
7	0	43.9	Yes
8	0	48.5	Yes
9	0	55.2	Yes
10	0	58.9	No

Summary

Total observations: $n = 10$

$$\text{Yes} = 4, \qquad \text{No} = 6$$

$$\hat{p} = \frac{4}{10} = 0.4$$

Parent entropy:

$$H(\text{parent}) = -\left(0.4 \log_2 0.4 + 0.6 \log_2 0.6\right) \approx 0.971.$$

Consider 10 patients with binary outcome HD (Yes/No).

Total counts:

$$\text{Yes} = 4, \quad \text{No} = 6.$$

Thus

$$\hat{p} = 0.4.$$

Parent entropy:

$$H(\text{parent}) = -\left(0.4 \log_2 0.4 + 0.6 \log_2 0.6\right) \approx 0.971.$$

Split 1: MaxHR_high (Categorical)

Left node (6 obs): Yes = 1, No = 5

$$H_L = -\left(\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}\right) \approx 0.650.$$

Right node (4 obs): Yes = 3, No = 1

$$H_R = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \approx 0.811.$$

Weighted entropy:

$$H(\text{split}) = \frac{6}{10}(0.650) + \frac{4}{10}(0.811) \approx 0.714.$$

Information gain:

$$IG = 0.971 - 0.714 = 0.257.$$

This split improves node purity.

Split 2: Age < 50 (Real-Valued)

Suppose both resulting nodes contain:

$$\text{Yes} = 2, \quad \text{No} = 3.$$

Each child has $\hat{p} = 0.4$, identical to the parent.
Thus

$$H_L = H_R = 0.971.$$

Weighted entropy remains

$$H(\text{split}) = 0.971,$$

so

$$IG = 0.$$

This split does **not** improve impurity.

Linear Model vs Decision Tree

Linear models assume a global functional form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j.$$

Decision trees assume a piecewise-constant model:

$$f(X) = \sum_{m=1}^{M} c_m \mathbf{1}(X \in R_m),$$

where R_m are disjoint regions of feature space.

	Linear Model	**Decision Tree**
Functional Form	Global linear	Piecewise constant
Assumption	Linear boundary	Axis-aligned splits
Interpretability	Moderate	Very high
Handles Nonlinearity	Poor (unless engineered)	Naturally
Handles Categories	Requires dummies	Directly
Variance	Low	High (unstable)
Prediction Smoothness	Smooth	Discontinuous

Summary

Entropy measures uncertainty in a node. A good split reduces entropy. Information gain quantifies the reduction.

Trees choose splits greedily to maximize information gain, while linear models estimate global parameters to minimize a loss function.

4 Practice Problem: Entropy with Two Categorical Predictors

Question

Consider the following small dataset for predicting Heart Disease (HD), where $Y \in \{\text{Yes}, \text{No}\}$.

There are two categorical predictors:

- Smoking $\in \{\text{Yes}, \text{No}\}$
- Exercise $\in \{\text{Low}, \text{High}\}$

ID	Smoking	Exercise	HD
1	Yes	Low	Yes
2	Yes	Low	Yes
3	Yes	High	No
4	Yes	High	No
5	No	Low	Yes
6	No	Low	No
7	No	High	No
8	No	High	No

Tasks:

1. Compute the entropy of the parent node.
2. Compute the entropy after splitting on Smoking.
3. Compute the entropy after splitting on Exercise.
4. Which split would a decision tree choose using information gain?

Answer

Step 1: Parent Entropy

Total observations: 8

$$\text{Yes} = 3, \quad \text{No} = 5$$

$$\hat{p} = 3/8 = 0.375$$

$$H(\text{parent}) = -\left(0.375 \log_2 0.375 + 0.625 \log_2 0.625\right) \approx 0.954$$

Step 2: Split on Smoking

Smoking = Yes (4 obs): Yes = 2, No = 2

$$H_{Yes} = -\left(0.5 \log_2 0.5 + 0.5 \log_2 0.5\right) = 1$$

Smoking = No (4 obs): Yes = 1, No = 3

$$H_{No} = -\left(0.25 \log_2 0.25 + 0.75 \log_2 0.75\right) \approx 0.811$$

Weighted entropy:

$$H(\text{split}) = \frac{4}{8}(1) + \frac{4}{8}(0.811) = 0.905$$

Information gain:

$$IG_{Smoking} = 0.954 - 0.905 = 0.049$$

Step 3: Split on Exercise

Exercise = Low (4 obs): Yes = 3, No = 1

$$H_{Low} = -\left(0.75 \log_2 0.75 + 0.25 \log_2 0.25\right) \approx 0.811$$

Exercise = High (4 obs): Yes = 0, No = 4

$$H_{High} = -\left(0 \log_2 0 + 1 \log_2 1\right) = 0$$

Weighted entropy:

$$H(\text{split}) = \frac{4}{8}(0.811) + \frac{4}{8}(0) = 0.406$$

Information gain:

$$IG_{Exercise} = 0.954 - 0.406 = 0.548$$

Conclusion

$$IG_{Exercise} > IG_{Smoking}$$

Therefore, the decision tree would split first on **Exercise**.