

MA477: Data Science
Lesson 20 Outline — 03 March 2026
 United States Military Academy, West Point
 Instructor: MAJ Patrick Kuiper

1 Administrative

- Student review
- Decision Tree Bagging Discussion

2 Decision Tree Bagging Classification Lesson Objectives

- Understand how bagging improves basic decision trees by reducing variance.
- Define and interpret the out-of-bag error.
- Use functions in Python's sklearn module to implement bagging.

3 Algorithmic Steps: Decision Trees and Bagging

Decision Tree Algorithm (Regression Version)

Goal: Partition the feature space into regions and predict by averaging within each region.

1. For each predictor X_j and each possible split point s , consider the partition

$$R_1(j, s) = \{X \mid X_j < s\}, \quad R_2(j, s) = \{X \mid X_j \geq s\}.$$

2. Choose (j, s) to minimize the Residual Sum of Squares:

$$\sum_{i \in R_1(j, s)} (y_i - \bar{y}_{R_1})^2 + \sum_{i \in R_2(j, s)} (y_i - \bar{y}_{R_2})^2.$$

3. Repeat this splitting process recursively on each resulting region until a stopping rule is met (e.g., minimum node size).
4. For prediction at x_0 , locate the region R_m containing x_0 and predict

$$\hat{f}(x_0) = \bar{y}_{R_m}.$$

Bagging Algorithm

Goal: Reduce variance by averaging many high-variance trees.

1. Draw B bootstrap samples from the training data (sample with replacement).
2. Fit a deep, unpruned decision tree to each bootstrap sample, producing models

$$\hat{f}^{*(1)}, \dots, \hat{f}^{*(B)}.$$

3. For regression, compute the average prediction:

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

4. For classification, predict by majority vote among the B trees.

Key Contrast:

A decision tree builds *one* hierarchical partition of the feature space.

Bagging builds *many* trees and stabilizes them by averaging.

4 Bagging and Variance Reduction

Instructor Worked Example

Suppose three deep regression trees are trained on three bootstrap samples. For a single test observation x_0 , the trees produce the predictions:

$$\hat{f}^{*(1)}(x_0) = 8, \quad \hat{f}^{*(2)}(x_0) = 10, \quad \hat{f}^{*(3)}(x_0) = 12.$$

The bagged prediction is the average:

$$\hat{f}_{\text{bag}}(x_0) = \frac{1}{3}(8 + 10 + 12) = \frac{30}{3} = 10.$$

Now suppose each tree prediction has variance 9 and the trees are independent. The variance of the averaged predictor is

$$\text{Var}\left(\frac{1}{3}\sum_{b=1}^3 Z_b\right) = \frac{1}{3^2}\sum_{b=1}^3 \text{Var}(Z_b) = \frac{1}{9}(9 + 9 + 9) = \frac{27}{9} = 3.$$

So averaging reduced variance from 9 to 3.

Key insight: Averaging reduces variance by a factor of B when models are independent.

4.1 Why Averaging Reduces Variance

Suppose

$$Z_1, Z_2, \dots, Z_B$$

are independent with variance σ^2 .

Define

$$\bar{Z} = \frac{1}{B}\sum_{b=1}^B Z_b.$$

Then

$$\text{Var}(\bar{Z}) = \frac{\sigma^2}{B}.$$

Bagging applies this principle to high-variance models like deep decision trees.

The bagged estimator is

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

For classification, we use majority vote instead of averaging.

Discussion Questions

1. Why do deep decision trees tend to have high variance?
2. Why does increasing B in bagging not increase overfitting?
3. What assumption about the individual trees is required in order for variance to decrease by a factor of B ?

Student Exercise: Guided Bagging Example

Four bootstrap trees produce the following regression predictions for x_0 :

5, 9, 7, 11.

Question 1: Compute the bagged prediction.

Answer:

$$\hat{f}_{\text{bag}}(x_0) = \frac{1}{4}(5 + 9 + 7 + 11) = \frac{32}{4} = 8.$$

Assume each tree prediction has variance 4 and the trees are independent.

Question 2: Compute the variance of the averaged predictor.

Answer:

$$\text{Var}\left(\frac{1}{4}\sum_{b=1}^4 Z_b\right) = \frac{1}{4^2}\sum_{b=1}^4 4 = \frac{1}{16}(16) = 1.$$

Thus variance decreases from 4 to 1.

Extension (Classification Version):

Four trees predict

$A, B, A, A.$

The bagged prediction is class A because it receives three out of four votes.