MA289 Mathematics of AI

Lesson 10 Outline — 17 March 2026

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

---

1 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

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

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.

2 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.

2.1 Why Averaging Reduces Variance

Suppose

$$Z_1, Z_2, \ldots, Z_B$$

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.

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.

4 Out-of-Bag (OOB) Error

One of the most useful properties of bagging is that it provides a built-in estimate of test error without requiring cross-validation.

Why OOB Observations Exist

Recall that bagging trains each tree on a **bootstrap sample**, which is formed by sampling n observations *with replacement* from the original training set of size n.

For a fixed observation i, the probability that it is *not* selected in a single bootstrap draw is

$$\left(1 - \frac{1}{n}\right)^n.$$

As n becomes large,

$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} \approx 0.37.$$

Therefore, on average, about 37% of the observations are **not** used in constructing any given bootstrap tree. These observations are called **out-of-bag (OOB)** observations.

Equivalently, each tree is trained on roughly 63% of the data and leaves out roughly 37%.

How OOB Error is Computed

Let $\hat{f}^{*(b)}$ denote the bth bootstrap tree.

For observation i:

1. Identify all trees that were trained *without* using observation i.
2. Use those trees to generate predictions for x_i.

This produces multiple predictions for observation i. We then aggregate them:

$$\hat{f}_{\text{OOB}}(x_i) = \frac{1}{|\mathcal{B}_i|}\sum_{b\in\mathcal{B}_i}\hat{f}^{*(b)}(x_i),$$

where $\mathcal{B}_i$ is the set of trees for which observation i was out-of-bag.

Regression Case

For regression, the OOB Mean Squared Error is

$$\text{OOB MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_{\text{OOB}}(x_i)\right)^2.$$

Classification Case

For classification, each OOB tree gives a class prediction. We take a majority vote among OOB trees to obtain $\hat{y}_{\text{OOB}}(x_i)$.

The OOB classification error is

$$\text{OOB Error} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{y_i \neq \hat{y}_{\text{OOB}}(x_i)\}.$$

Why OOB Error is Valid

For each observation i, the prediction $\hat{f}_{\text{OOB}}(x_i)$ is constructed only from trees that did not train on that observation.

Therefore, each OOB prediction is effectively made on "unseen" data.

When the number of trees B is large, the OOB error is approximately equivalent to leave-one-out cross-validation, but it is computationally much more efficient.

Key Insight

Bagging not only reduces variance through averaging,

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

but it also automatically provides a reliable estimate of generalization error through the OOB procedure.

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:

$$5, \quad 9, \quad 7, \quad 11.$$

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:

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

Thus variance decreases from 4 to 1.

Extension (Classification Version):

Four trees predict

$$A, \quad B, \quad A, \quad A.$$

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