MA477: Data Science

Lesson 22 Outline — 10 March 2026

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

---

1 Administrative

- Calendar review
- Student review
- Tree Based Methods mentimeter
- Random Forest Bagging Discussion
- Coding Exercise

2 Random Forest Lesson Objectives

- Understand how the random forest algorithm differs from bagging.
- Use functions in Python's sklearn to fit random forest models.
- Understand how you can extract feature importance using Random Forests

3 Random Forests

Quick Review: Bagging

Recall that **bagging (bootstrap aggregating)** reduces variance by averaging predictions from many decision trees trained on bootstrap samples.

1. Draw B bootstrap samples from the training data.
2. Train a deep decision tree on each sample.
3. Aggregate predictions:

   Regression:

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

   Classification: majority vote among the B trees.

Bagging works well because deep trees have **high variance**, and averaging reduces this variance.

However, bagging has an important limitation.

If one predictor is very strong, most trees will choose it for their top split. This causes the trees to become **highly correlated**.

Averaging highly correlated models does *not* reduce variance as effectively.

Pasting

Pasting is a variation of bagging in which models are trained on random subsets of the training data drawn **without replacement**.

Suppose the training dataset contains n observations. For each model $b = 1, \ldots, B$, we draw a subset of size $k \leq n$ from the training data *without replacement* and train a model on that subset.

This produces predictions

$$Z_1(x), Z_2(x), \ldots, Z_B(x)$$

and the ensemble prediction is the average

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

Comparison with Bagging

- **Bagging:** Each model is trained on a bootstrap sample of size n drawn *with replacement*. Some observations may appear multiple times in the sample, while others are omitted.
- **Pasting:** Each model is trained on a random subset of the data drawn *without replacement*. Each observation can appear at most once in a given training subset.

Both approaches aim to generate multiple diverse models whose predictions can be averaged to reduce variance.

Practical Notes

- Bagging is typically preferred because bootstrap sampling creates more variation among training sets.
- Pasting is sometimes useful when datasets are very large, since each model can be trained on a smaller subset of the data.

4 Variance Reduction from Averaging

A key motivation behind bagging and random forests is that averaging many high-variance models can substantially reduce prediction variance.

Suppose we train B models that each produce a prediction for an input x. Denote these predictions by

$$Z_1(x), Z_2(x), \ldots, Z_B(x).$$

The ensemble prediction is the average

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

Variance Reduction under Independence

Assume that the predictions $Z_1, \ldots, Z_B$ are independent and each has variance σ^2.

Then the variance of the averaged predictor is

$$\mathrm{Var}\big(\bar{Z}(x)\big) = \mathrm{Var}\left(\frac{1}{B} \sum_{b=1}^{B} Z_b(x) \right).$$

Using properties of variance,

$$\mathrm{Var}\big(\bar{Z}(x)\big) = \frac{1}{B^2}\sum_{b=1}^{B}\mathrm{Var}(Z_b).$$

Since each model has variance σ^2,

$$\mathrm{Var}\big(\bar{Z}(x)\big) = \frac{1}{B^2}(B\sigma^2) = \frac{\sigma^2}{B}.$$

Thus averaging B independent models reduces variance by a factor of B.

Interpretation of B and n

It is important to distinguish two different quantities:

- n = number of observations in the training dataset.
- B = number of models (trees) in the ensemble.

The dataset size n determines how the model is trained. The ensemble size B determines how many models are averaged.

Variance reduction in bagging occurs because predictions from B models are averaged, not because the dataset size changes.

Assumptions Behind the $1/B$ Variance Reduction

The variance reduction result

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

relies on two key assumptions:

1. The predictions $Z_1, \ldots, Z_B$ are independent.
2. Each prediction has the same variance σ^2.

In practice, predictions from different trees are not fully independent because they are trained on related bootstrap samples.

If predictions are positively correlated with correlation ρ, the variance becomes

$$\mathrm{Var}(\bar{Z}) = \sigma^2\left(\rho + \frac{1-\rho}{B}\right).$$

This shows that averaging is most effective when the individual models are **weakly correlated**. Random forests improve upon bagging by reducing this correlation through random feature selection at each split.

Big Picture: Random Forest Idea

Random forests improve bagging by introducing additional randomness when building trees.

Instead of allowing every split to consider all p predictors, each split considers only a **random subset of predictors**.

If m predictors are sampled from the p available predictors, the split can only use those m candidates.

Typical choice:

$$m \approx \sqrt{p}$$

for classification problems.
This strategy helps:

- Decorrelate the trees
- Encourage different predictors to be used
- Improve the effectiveness of averaging

If $m = p$, the random forest becomes equivalent to bagging.

Random Forest Algorithm

1. Draw B bootstrap samples from the training data.
2. For each bootstrap sample, grow a deep decision tree.
3. At each split in the tree:
   (a) Randomly select m predictors from the full set of p predictors.
   (b) Determine the best split using only those m predictors.
4. Repeat until the tree is fully grown (or another stopping rule is met).
5. Aggregate predictions across all trees:
   Regression:
   $$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*(b)}(x)$$
   Classification: majority vote among the B trees.

Why Random Forests Work

Random forests reduce prediction variance through two mechanisms:

1. **Bootstrap sampling** (same as bagging)
2. **Random feature selection at each split**

The second mechanism reduces correlation among trees, making averaging more effective.

Discussion Questions

1. Why does bagging fail to substantially reduce variance when one predictor is much stronger than all others?
2. How does randomly selecting m predictors at each split help reduce correlation between trees?
3. What happens if we set $m = p$ in a random forest?
4. Why can we increase the number of trees B in a random forest without causing overfitting?