

Introduction to Machine Learning

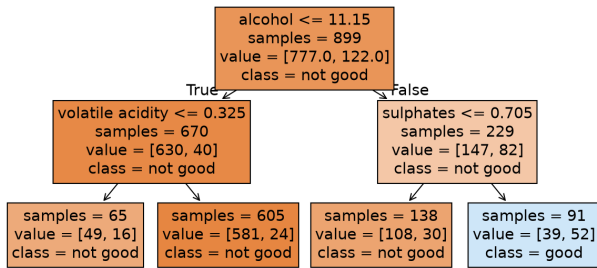
Day 6: Bagging and Random Forests

Joanna Bieri DATA301

Get out your handwritten notes!

Reading a Tree

PREDICT



A new wine: alcohol 12.1, volatile acidity 0.40, sulphates 0.80.

Which leaf does it land in? What does the tree predict?

Hint. True goes left. value is [not good, good].

Reading a Tree

ANSWER

- Alcohol 12.1 is above 11.15, so it goes **right**.
- Sulphates 0.80 is above 0.705, so **right** again.
- Far right leaf, value [39, 52]: the tree predicts **good**.
- Its probability of good is $52 / 91 = 0.571$. Q1a in your notes.

No Scaler Today

ASK

Logistic regression needed StandardScaler.

Our trees did not.

Why not?

Hint. What does every question in a tree compare?

No Scaler Today

ANSWER

- Every question compares **one** feature to a cutoff.
- Change the units and the cutoff moves, but every wine gives the same answer.
- Logistic regression adds up weighted features, so big numbers can swamp small ones. Q1c.

The Big Tree

ASK

A tree gets every training wine right.

Good news or bad news?

Hint. Remember the degree 15 polynomial from Day 2.

The Big Tree

ANSWER

- **Bad news.** It memorized the training wines, noise and all.
- Training accuracy 1.0. Validation average precision 0.299, against 0.532 for logistic regression.
- Low bias, very high variance. Q2a.

The Crowd

PREDICT

Three voters. Each is right 70 percent of the time, independently.

How often is the majority right?

Hint. The majority is right when at least 2 of the 3 are right.

Add two cases: all 3 right, or exactly 2 right (3 ways to pick who is wrong).

The Crowd

ANSWER

all 3 right:	$0.7 * 0.7 * 0.7$	= 0.343
exactly 2 right:	$3 * 0.7 * 0.7 * 0.3$	= 0.441
	total	0.784

The crowd, 0.784, beats every single voter, 0.7. That only works because their mistakes are independent.

A Bootstrap Sample

ASK

wine	alcohol
A	9.4
B	12.8
C	10.0
D	11.2
E	9.8

```
wines.sample(frac=1, replace=True)
```

It gave back the wines E, A, D, D, D.

Which wines are missing? Why is D in there three times?

Hint. `replace=True` means a wine goes back in the bag after it is picked.

A Bootstrap Sample

ANSWER

- **B and C are missing.** Only 3 different wines made it in.
- With replacement, a wine can be drawn again after it is picked.
- On our 899 training wines, one bootstrap sample kept 552 different wines, about 61 percent. Every tree gets a different mix.

Bagging

ASK

One overfit tree: 0.299.

200 overfit trees, averaged: 0.641.

Where did the improvement come from?

Hint. Each tree saw a different bootstrap sample. Did they memorize the same noise?

Bagging

ANSWER

- Each tree saw different wines, so it memorized **different** noise.
- The real pattern shows up in most trees and survives the average.
The noise cancels.
- Bagging fixes variance, not bias. Q4b and Q4c.

Random Forests

ASK

A random forest hides most features
from every question.

Why make each tree worse on purpose?

Hint. All five resampled trees asked about alcohol first. What does that do to a crowd?

Random Forests

ANSWER

- With every feature available, every tree leans on alcohol, so the trees look alike.
- Hiding features forces trees to find different clues, so their mistakes differ.
- Forest 0.667 against bagging 0.641 on validation. Q5a.

Random Forests

REVERSE

You want 200 trees on the training wines, the same result every run.

What is the code?

Hint. Make the object first, then fit it. No scaler needed.

Random Forests

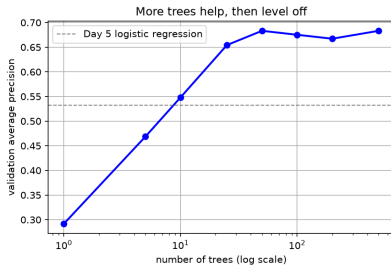
ANSWER

```
forest = RandomForestClassifier(  
    n_estimators=200, random_state=42)  
forest.fit(X_train, y_train)
```

`random_state` fixes the bootstrap samples and the feature choices.

How Many Trees?

ASK



Should we use 500 trees because 0.683 beats 0.667?

Hint. The validation set has 300 wines. How much would you expect a score to wobble?

How Many Trees?

ANSWER

- **No.** The curve levels off around 50 trees.
- After that, the wiggles are noise in 300 wines.
- More trees never hurt. They only cost time. Q5c.

Out-of-Bag

ASK

The forest reports an out-of-bag
accuracy of 0.899.

Is that good?

Hint. What does a model that always says "not good" score?

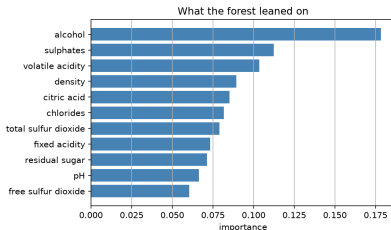
Out-of-Bag

ANSWER

- Only somewhat. Always saying not good scores **0.864**.
- `oob_score_` is accuracy. Out-of-bag average precision: 0.664.
- That matches validation's 0.667, by a different route. Q6b.

Feature Importance

ASK



A winemaker says "add more alcohol." What do you say?

Hint. Does the forest know why alcohol helps, or only that it helps?

Feature Importance

ANSWER

- Importance says what helped the forest **predict**, not what **causes** a good wine.
- Related features split the credit: alcohol and density.
- Q7a in your notes.

Questions?

Questions?