DATA 202 — Ensembles · Week 11 Wednesday

Pumpkin seeds of two varieties, measured from photos. Ç = Çerçevelik, Ü = Ürgüp Sivrisi. Your professor will tell you when to do each part.

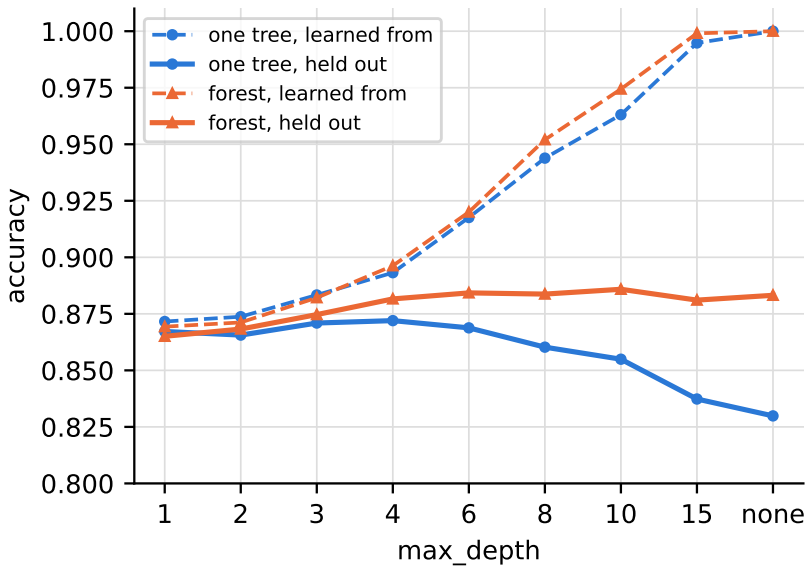
1 · Five trees vote

A random forest of five deep trees, each grown on its own bootstrap sample of the training seeds. Their votes on six new seeds:

seed	tree 1	tree 2	tree 3	tree 4	tree 5	forest	actual	right?
A	Ü	Ü	Ü	Ü	Ü	_____	Ü	_____
B	Ü	Ç	Ç	Ç	Ç	_____	Ç	_____
C	Ç	Ç	Ç	Ü	Ü	_____	Ç	_____
D	Ç	Ç	Ü	Ç	Ç	_____	Ç	_____
E	Ü	Ç	Ü	Ü	Ç	_____	Ü	_____
F	Ü	Ü	Ü	Ü	Ü	_____	Ç	_____

1. **Fill in** the forest's answer (the majority vote), then mark it right or wrong.
2. **Circle** every wrong vote of a single tree. Is there a tree with no circles? _____
3. **Pair talk:** which single tree would you trust alone? Would 100 more trees fix seed F?

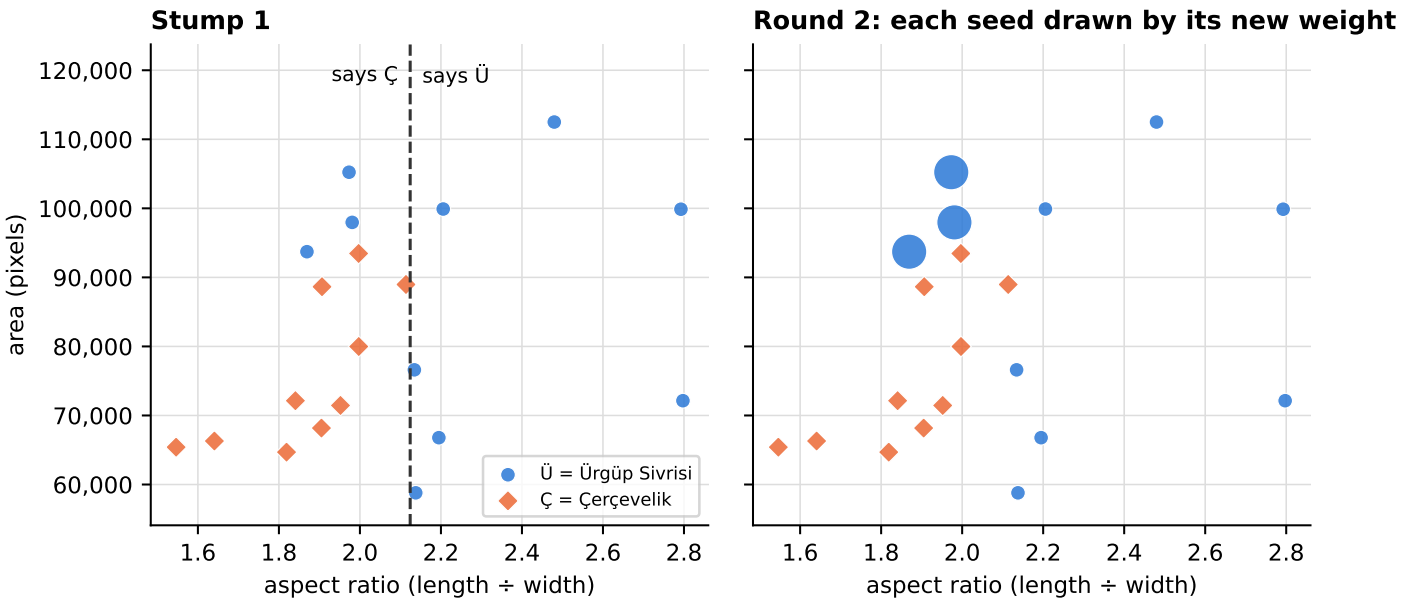
2 · Tree vs. forest



Five-fold cross-validation on the training seeds, as the trees grow deeper.

1. **Label** the side where the single tree overfits.
2. **Pair talk:** both reach 100% on the seeds they learned from. Why doesn't the forest's held-out accuracy fall?

3 · AdaBoost by hand



Twenty training seeds, two columns. **Left:** stump 1 asks one question, "aspect ratio ≤ 2.12?": left of the line it says Ç, right of it Ü. **Right:** the same seeds after reweighting, each drawn by its new weight.

1. **Left: circle** the seeds stump 1 gets wrong.
2. **Right: draw** the single cut (one vertical or one horizontal line) that stump 2 should make.
3. **Pair talk:** each stump alone misses several seeds, yet the weighted vote of three stumps gets all 20 right. How?