

DATA 202 — Cross-Validation and Grid Search · Week 7 Wednesday

Same children as today's slides: ASD vs. NEU by two blood measurements, 111 training children and 48 test children. Your professor will tell you when to do each part.

1 · Five splits

Test accuracy of a tree at each `max_depth`, for five different random splits.

random_state	depth 1	depth 2	depth 3	depth 4	depth 5
1	0.750	0.938	0.938	0.896	0.833
7 (Monday's)	0.792	0.875	0.875	0.771	0.729
13	0.812	0.896	0.896	0.875	0.875
21	0.583	0.771	0.729	0.708	0.688
30	0.750	0.875	0.875	0.896	0.854

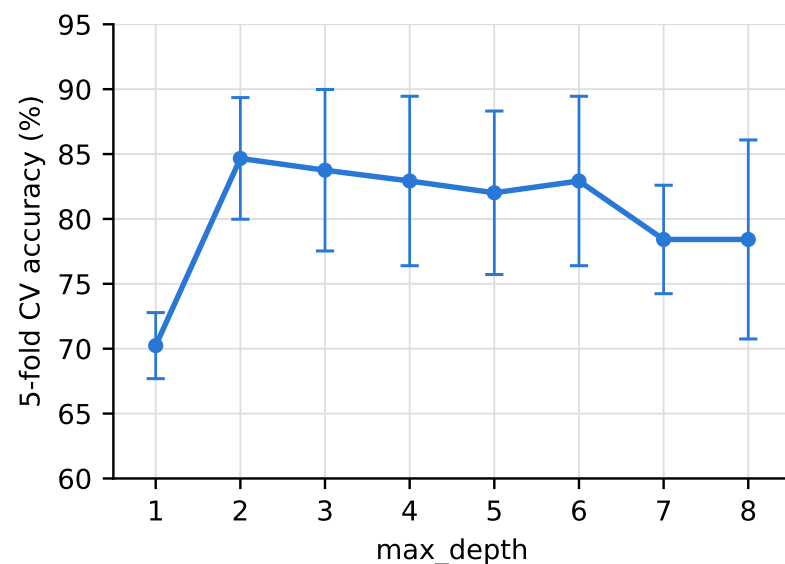
1. In each row, **circle** the best score.
2. Depth 2: lowest _____, highest _____.

2 · Five folds

	training set: 111 children						test set: 48
	fold 1	fold 2	fold 3	fold 4	fold 5		
round 1						→ 0.870	locked
round 2						→ 0.818	locked
round 3						→ 0.864	locked
round 4						→ 0.909	locked
round 5						→ 0.773	locked

1. The 111 training children are cut into 5 folds. In each round, **shade** the fold that gets scored.
2. Cross-validation score = mean of the 5 rounds = _____ ÷ 5 = _____

3 · Pick the depth



1. Mean accuracy over the 5 rounds, for each depth. The bars show how much the rounds disagreed.
2. **Circle** the best depth.
3. **Label** the underfitting side and the overfitting side.

4 · The grid

5-fold cross-validation accuracy for every combination of `max_depth` and `min_samples_leaf`.

min_samples_leaf	depth 1	depth 2	depth 3	depth 4	depth 5	depth 6	no limit
1	0.702	0.847	0.838	0.829	0.820	0.829	0.784
2	0.702	0.847	0.828	0.820	0.829	0.829	0.829
5	0.702	0.847	0.856	0.847	0.847	0.856	0.856
10	0.702	0.828	0.828	0.828	0.828	0.828	0.828
20	0.702	0.774	0.774	0.774	0.774	0.774	0.774

1. **Circle** the best score.
2. **Box** every cell within 0.01 of the best. (One child out of 111 is about 0.009.)
3. Trees fitted: _____ combinations $\times$ _____ folds = _____
4. Which boxed cell would you choose, and why?