

Estimated time: 5 minutes

Overfit decision trees



Data scientists can correct this through a process called “pruning”, which just means adjusting the hyperparameters of the tree and preventing it from classifying every single data point. Trees can be pre-pruned, where we set limits before building the tree, or post-pruned, where we trim the tree back after it has been built by removing or collapsing nodes that don’t contain much information. Technically, when we talk about pruning decision trees, we’re talking only about post-pruning, but we’ll use the term here to refer to both methods.

Pre-pruning criteria can include:

- Maximum tree depth
 - limiting how deep the tree grows (how many levels of splitting)
- Minimum node size
 - requiring that each node have a minimum number of data points in order to split it further
- Minimum leaf size
 - Requiring at least a certain number of data points in each leaf
- Maximum features
 - maximum number of features to consider at each split
 - Introduces randomness