

# **Introduction to Machine Learning**

## **CART: Growing a Tree**

[compstat-lmu.github.io/lecture\\_i2ml](https://compstat-lmu.github.io/lecture_i2ml)

# TREE GROWING

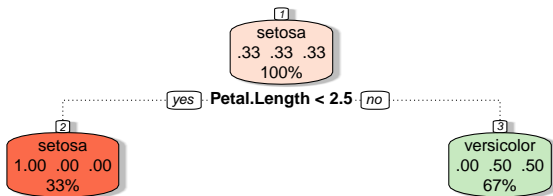
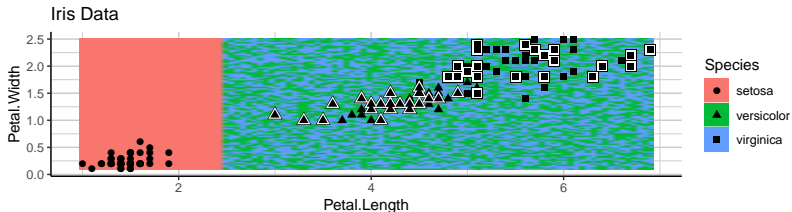
We start with an empty tree, a root node that contains all the data. Trees are then grown by recursively applying *greedy* optimization to each node  $\mathcal{N}$ .

Greedy means we do an **exhaustive search**: All possible splits of  $\mathcal{N}$  on all possible points  $t$  for all features  $x_j$  are compared in terms of their empirical risk  $\mathcal{R}(\mathcal{N}, j, t)$ .

The training data is then distributed to child nodes according to the optimal split and the procedure is repeated in the child nodes.

# TREE GROWING

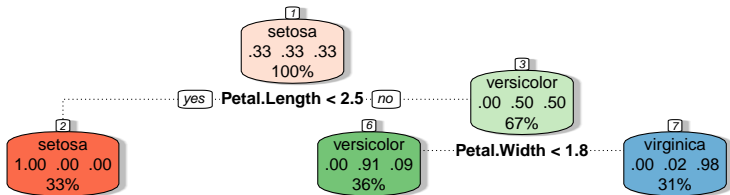
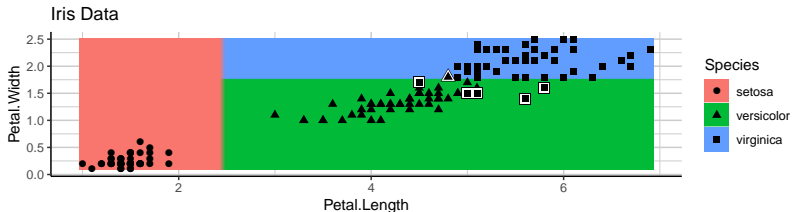
Start with a root node of all data, then search for a feature and split-point that minimizes the empirical risk in the child nodes.



Nodes display their current label distribution here for illustration.

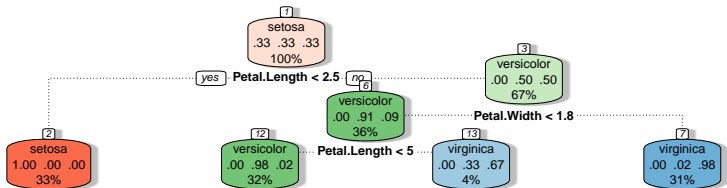
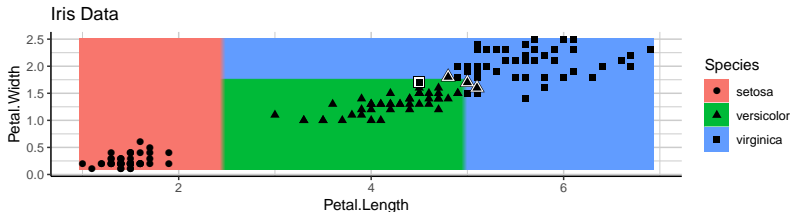
# TREE GROWING

We then proceed recursively for each child node: Iterate over all features, and for each feature over all possible split points. Select the best split and divide data in parent node into left and right child nodes:

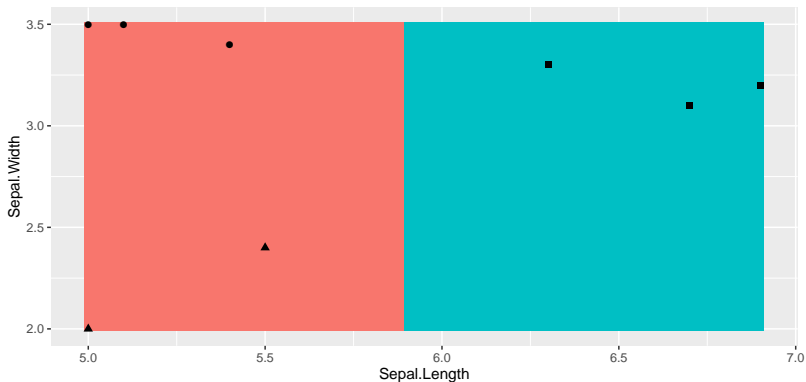


# TREE GROWING

We then proceed recursively for each child node: Iterate over all features, and for each feature over all split points. Select the “best” split and divide data in parent node into left and right child nodes:



# SPLIT PLACEMENT



Splits are usually placed at the mid-point of the observations they split: the large margin to the next closest observations makes better generalization on new, unseen data more likely.