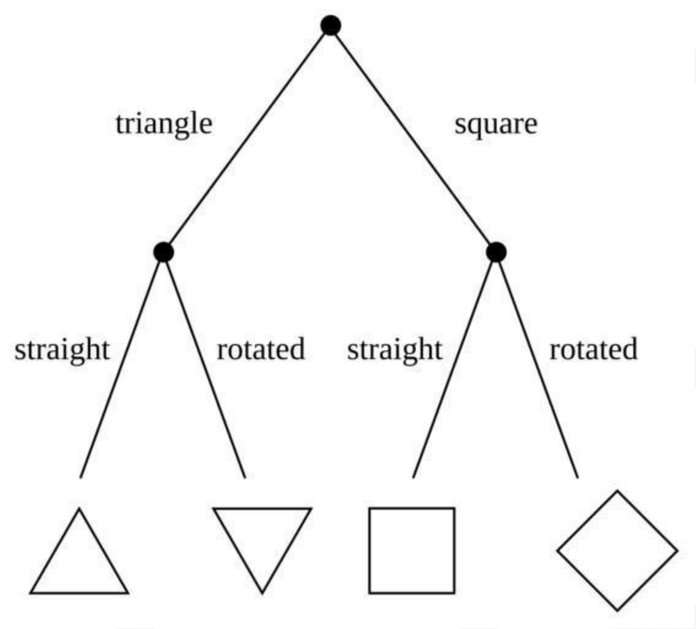
Decision Trees

Intro to Decision Trees

Decision trees are a way to split the predictor space into several simple regions.

Decision trees can be used for both **regression** and **classification**.



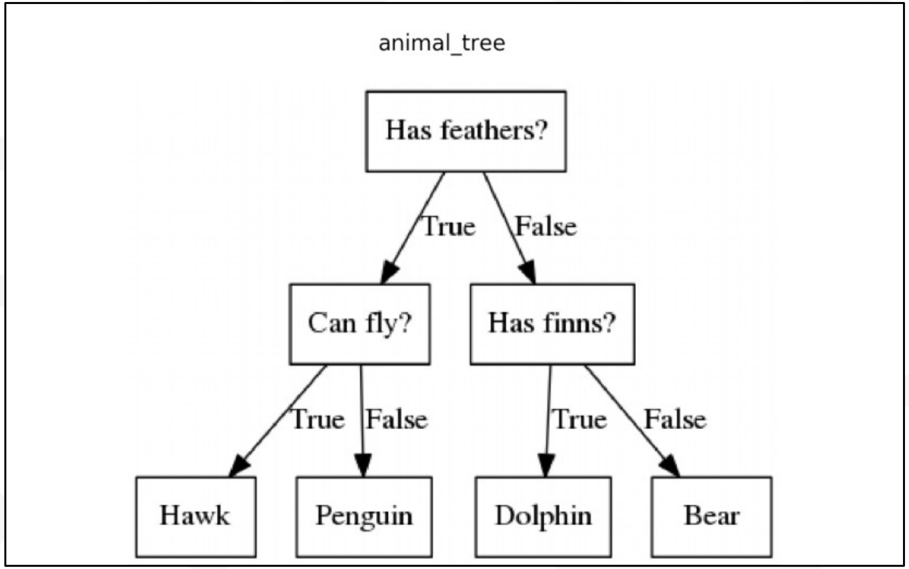
Decision Tree Format

- Decision tree is a flow chart-type structure

- Each node represents a “test” which decide which direction the flow takes for a given data instance

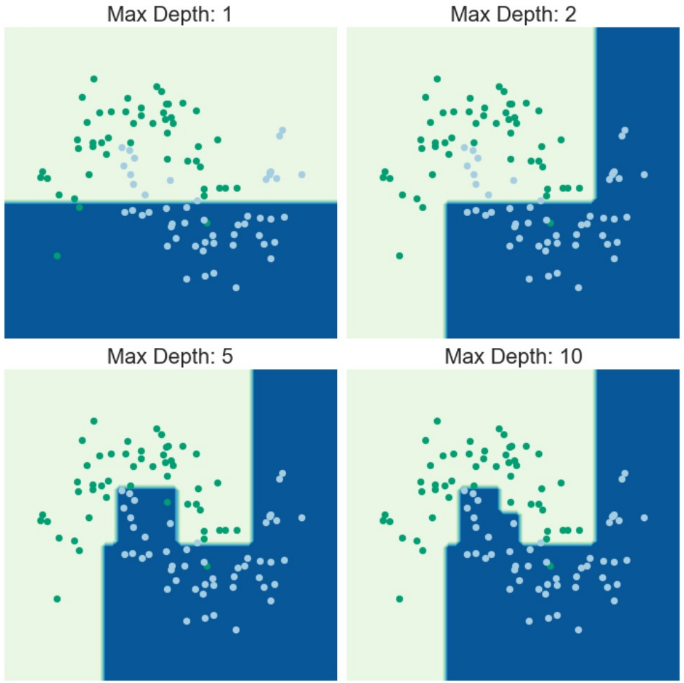
- Can think of it as a sort of a “20 questions” type of flow

- When you hit the terminal node at bottom – that is the predicted result.



How Do Decision Trees Work?

- Iterative splitting of the data into regions in a top-down greedy fashion
  - *Top down* means starts at the top of the tree (when all observations belong to the same region) and successively splits the predictor space.
  - *Greedy* means at each step, the best split is made at that step instead of looking ahead and trying to choose a split that would work better in the future.



Basic Example

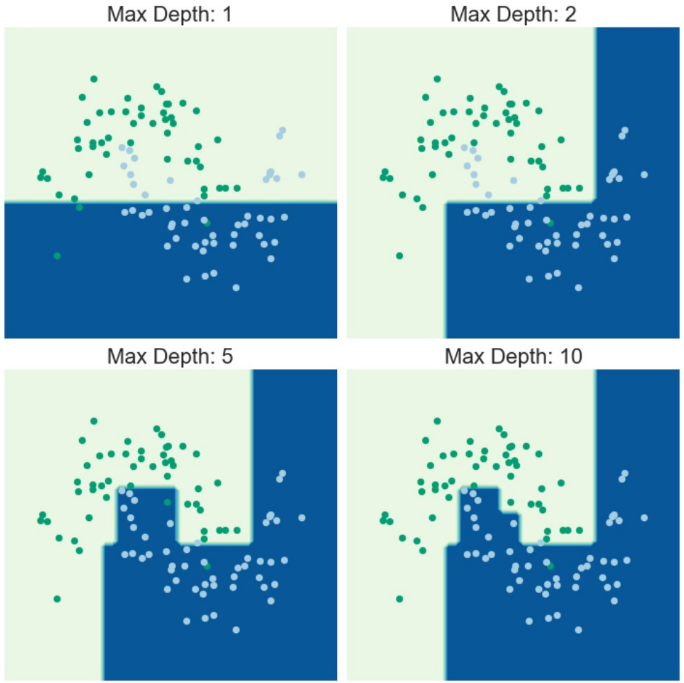
https://www.jeremyjordan.me/decision-trees/

How Do Decision Trees Work?

- How does it choose the decision points?
  - Chosen to minimize classification error in classification trees
  - Chosen to minimize Mean Squared Error in regression trees

- The resulting classifier separates the feature space into distinct subsets

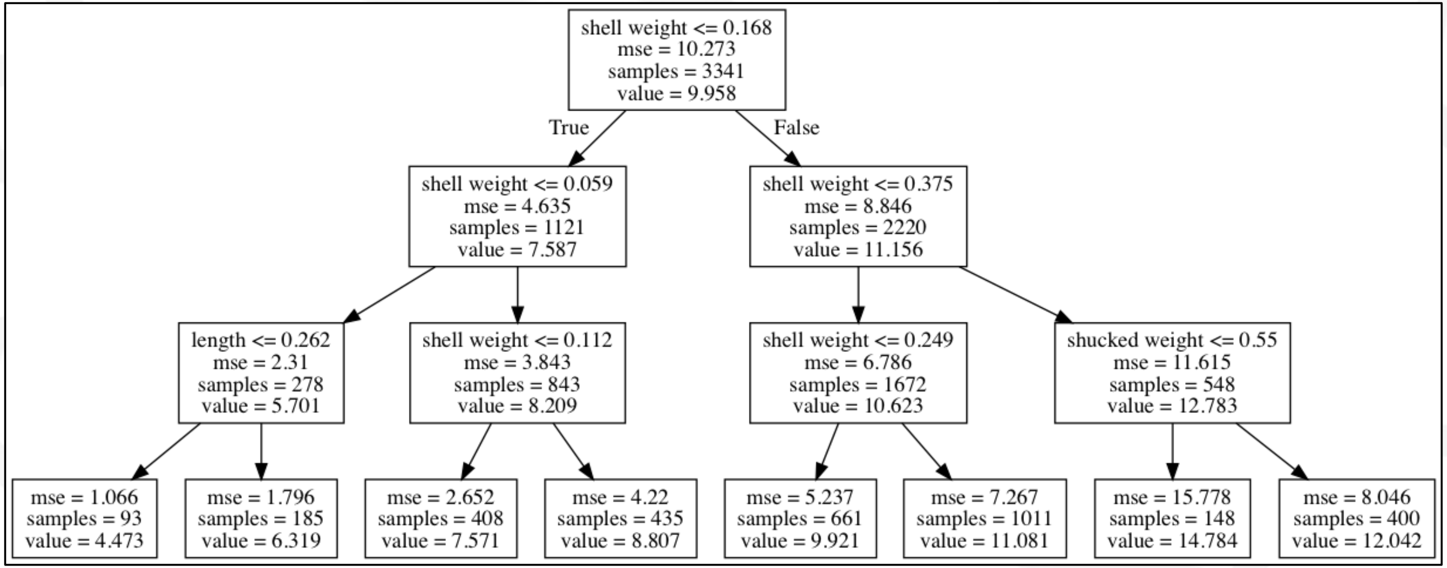
- Max Depth = Number of levels allowed in tree



Decision Tree Regressor

Here’s a tree to predict the number of rings on abalone based on variables such as shell weight, length, diameter, etc.

At the terminal node the “value” is your predicted number of rings. It is simply the mean value of all the samples in that split.



Pros & Cons of Decision Trees

Pros

- Easy to explain and understand the reasoning behind predictions.
- Can be displayed graphically and are easily interpreted.
- Can be used with categorical and continuous predictors.
- Can be used for regression or classification.

Cons

- Generally do not have as high predictive accuracy as other ML approaches.
- Subject to overfitting.
- Non-robust (i.e., small changes in the data can have a large change in the resulting tree).

Note: there are ways to address these cons that we will discuss in future lectures.

Which is Better - Linear Regression or Decision Trees?

It depends!

- If the relationship between the predictors and response is approximately linear, linear regression will outperform decision trees.
- If there is non-linearity (complex relationships between predictors and response), decision trees will outperform linear regression.
- Decision trees can be easier to interpret than linear regression. So if you need something highly interpretable, decision trees might be the best choice.