Decision Trees

How to decide

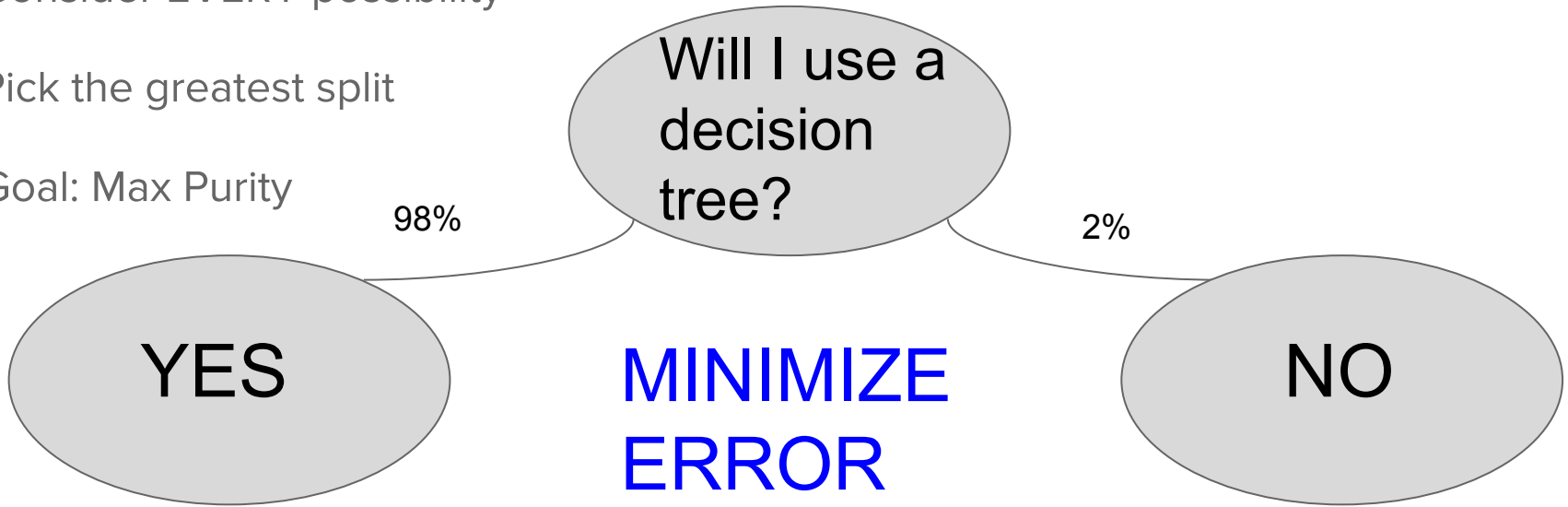
Why Decision Trees

They help us see into logical space

Consider EVERY possibility

Pick the greatest split

Goal: Max Purity



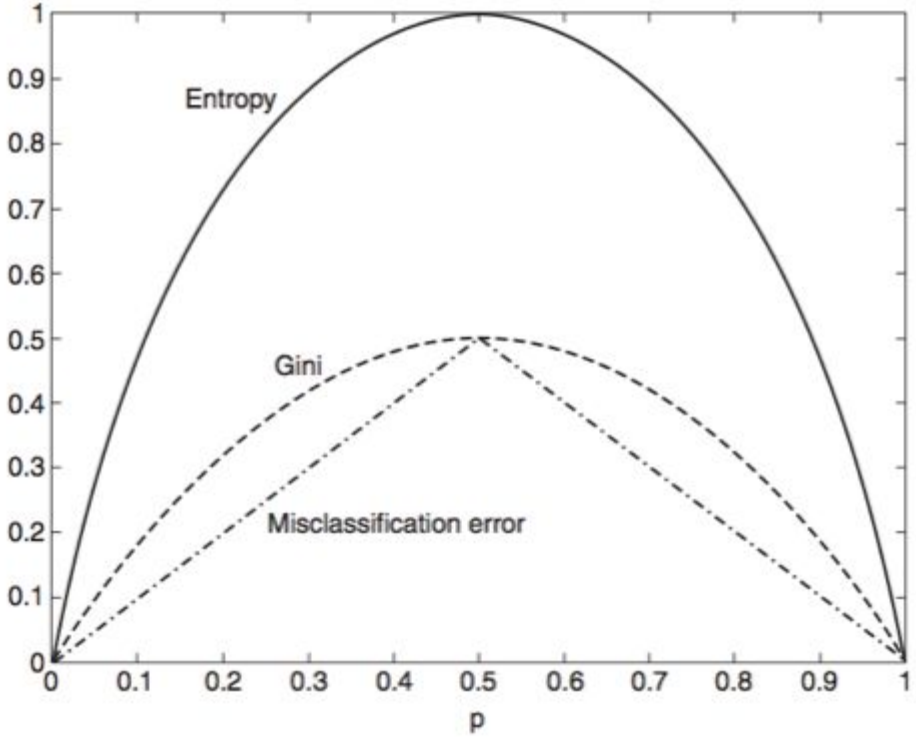
What is Purity

Gini is the percent chance that a randomly drawn element will be misclassified

Purity is information gain

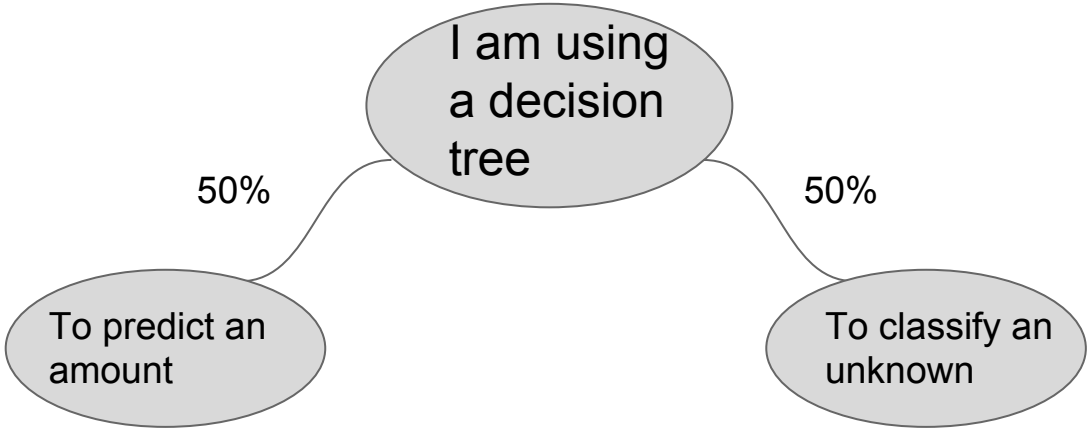
Information Gain =

Entropy(parent) - Weighted Sum of Entropy(Children)



What kind of decision?

The Classification And Regression Tree (CART)

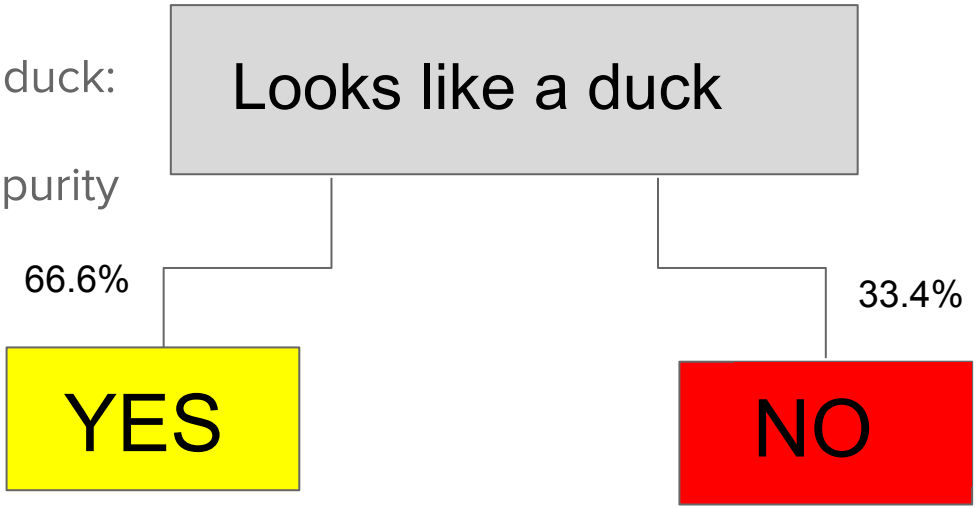


Classification Decisions

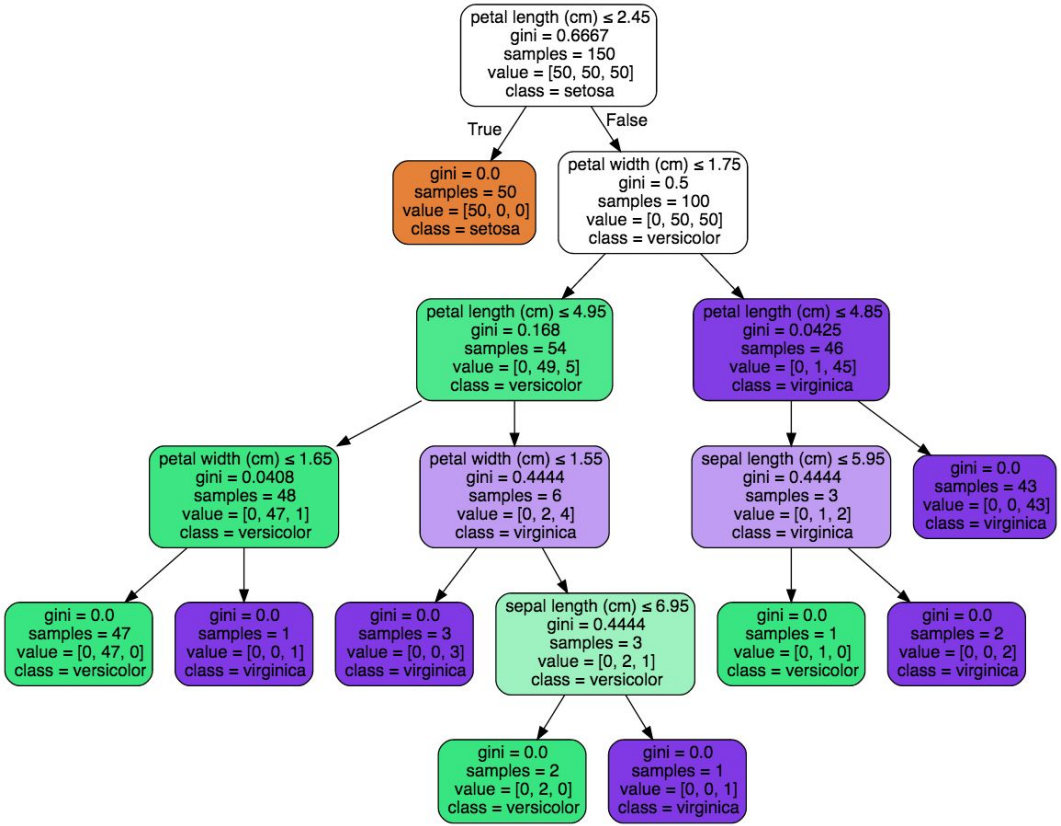
Is it a duck?

If it looks like a duck:

Big increase in purity



As seen in SciKitLearn



Regression Decisions

Useful for continuous variables

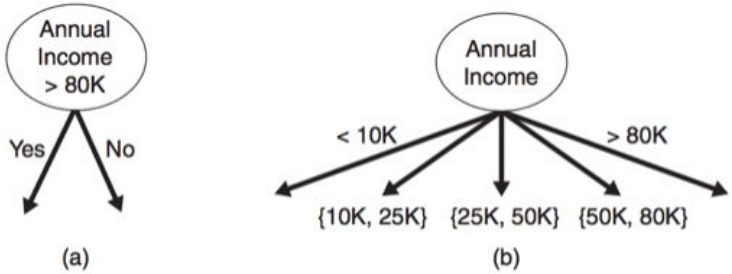
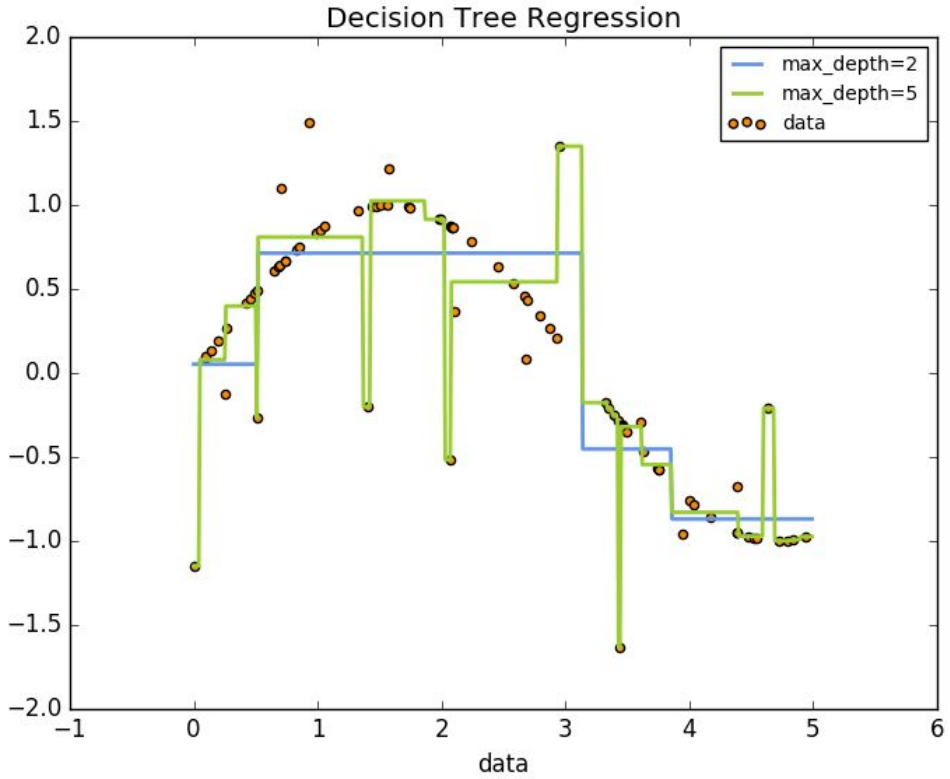
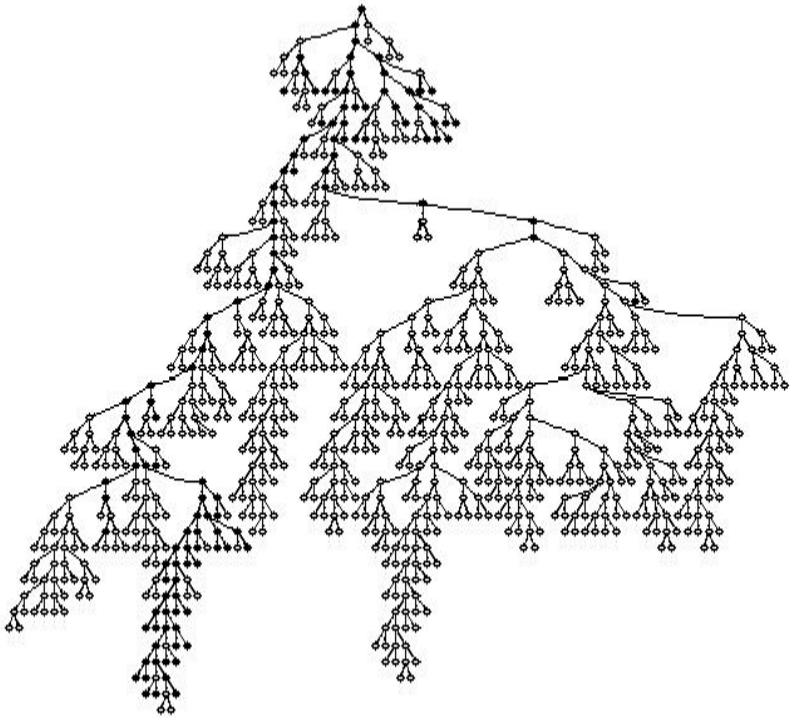
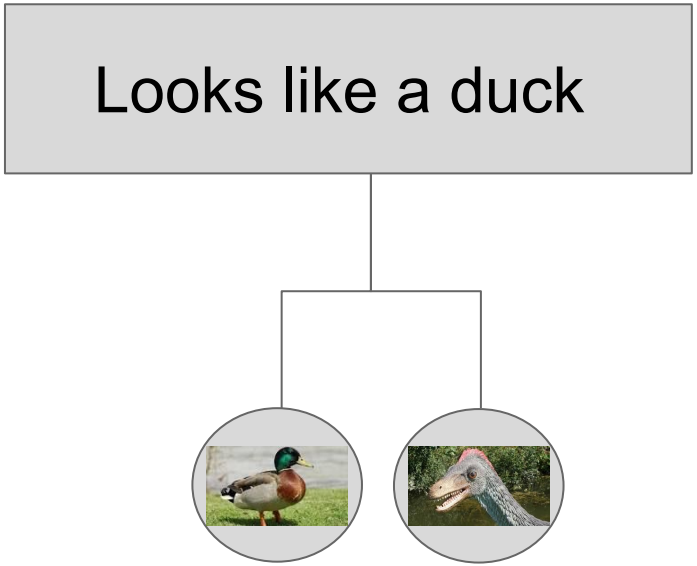


Figure 4.11. Test condition for continuous attributes.



Bias or Variance?





Methods to reduce bias

Pruning: Remove the nodes with the least explanatory power

Max depth: Specify number of nodes at outset

Reduce complexity, reduce overfitting

Pros and Cons of Decision Trees

Pros:

- Simple to understand and interpret
- Can handle categorical and numerical
- Little data prep
- Performs well with big data
- Mirrors human decision-making

Pros and Cons of Decision Trees

Cons:

- A change in the training data can dramatically affect predictions
- Prone to overfitting and complexity
- Computationally expensive