Decision Trees

Stopping Criteria for Tree Induction

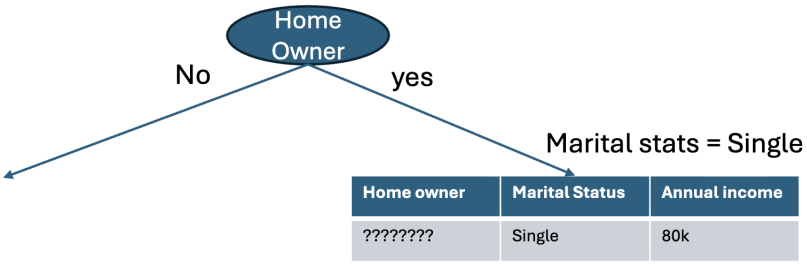
1. Stop expanding a node when all the records belong to the same class
2. Stop expanding a node when all the records have similar attribute values
3. Early termination - When the performance is "good enough" or the tree is N number of layers

Handling Missing Values

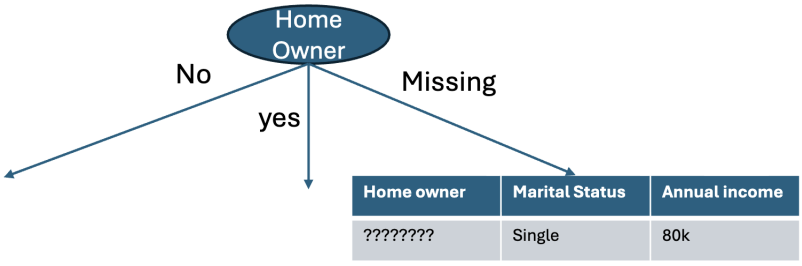
- What happens if we reach a node that asks for a home owner value, but the test instance is missing the value?

Home Owner	Marital Status	Annual Income
??????????	Single	80k

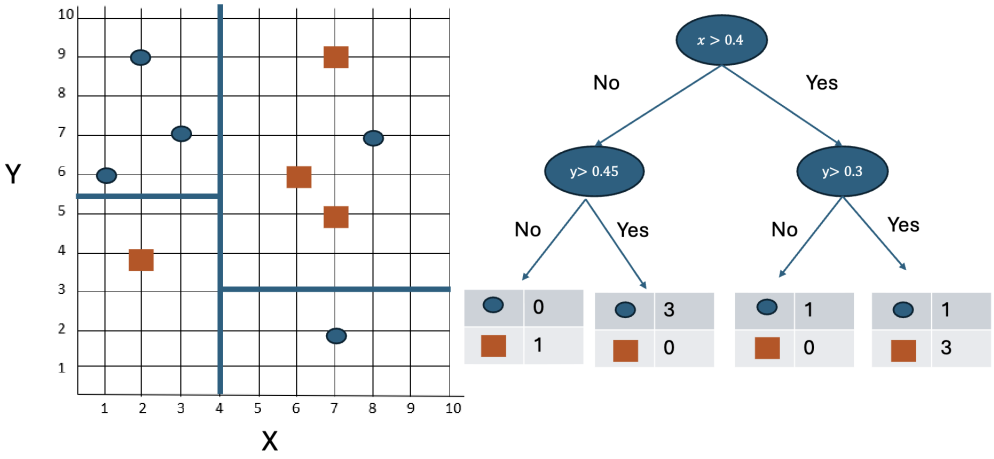
- Method 1: Surrogate Split
  - Use a different attribute to split



- Method 2: Separate Class
  - Have a class reserved for missing values



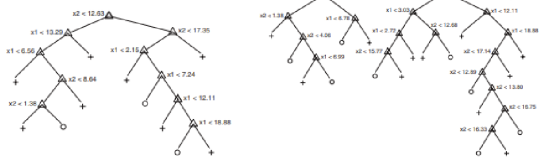
What is happening with the splits?



Classification Errors

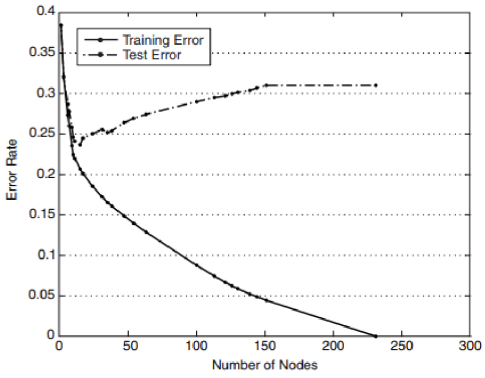
- Training errors (apparent errors)
  - Errors made on the training set
- Generalization errors
  - Expected error of a model over random selection of records from same distribution

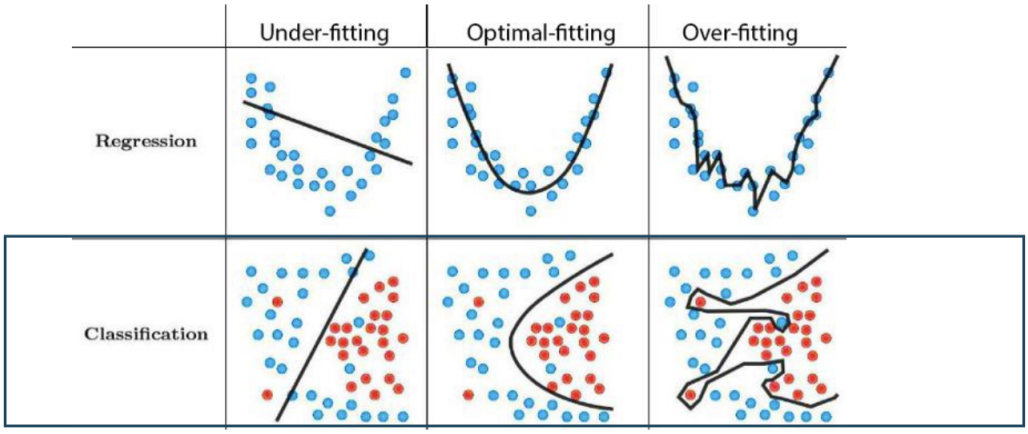
Underfitting and Overfitting (Example)



(a) Decision tree with 11 leaf nodes.

(b) Decision tree with 24 leaf nodes.

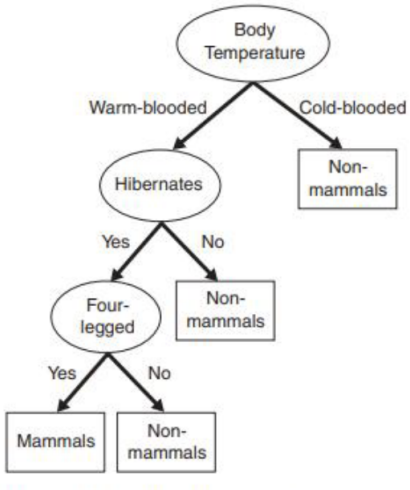




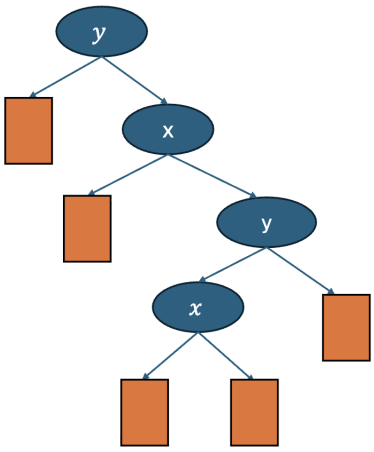
Overfitting due to Insufficient Examples

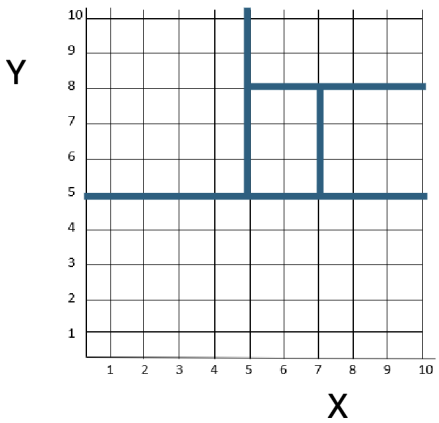
- Larger training set has a higher chance of more accurately representing the data and less likely to overfit

Name	Body Temp.	Gives Birth	Four Legged	Hibernates	Class Label Mammal / Not Mammal
Salamander	Cold-Blooded	No	Yes	Yes	No
Guppy	Cold-Blooded	Yes	No	No	No
Eagle	Warm-Blooded	No	No	No	No
Poorwill	Warm-Blooded	No	No	Yes	No
Platypus	Warm-Blooded	No	Yes	Yes	Yes

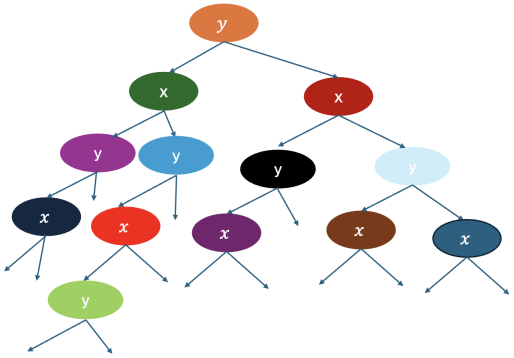


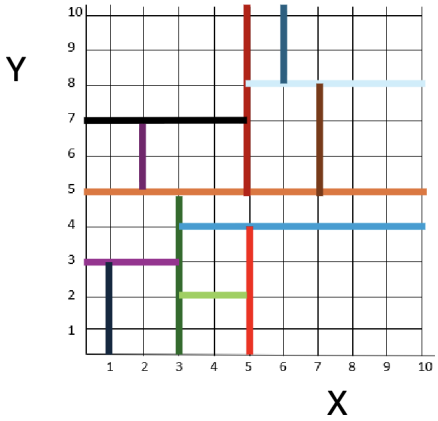
3 Leaf Nodes



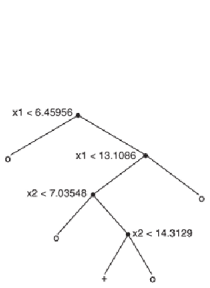


14 Leaf Nodes

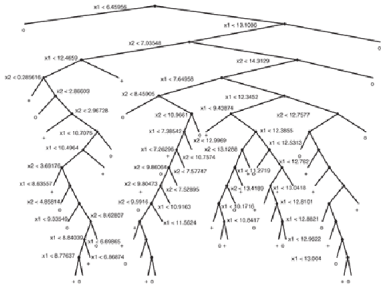




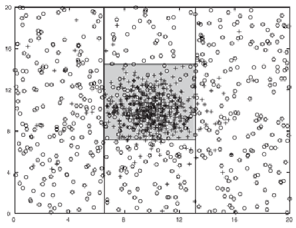
Overfitting due to Complexity



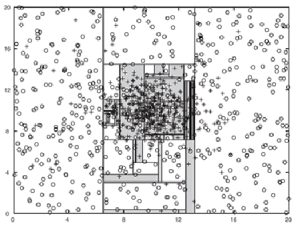
(a) Decision tree with 5 leaf nodes.



(b) Decision tree with 50 leaf nodes.



(c) Decision boundary for tree with 5 leaf nodes.



(d) Decision boundary for tree with 50 leaf nodes.

Notes on Overfitting

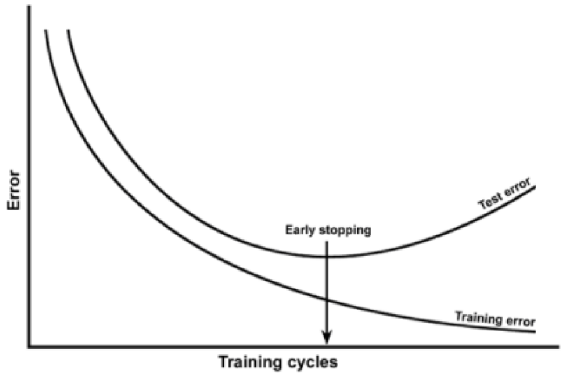
- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records
- Need new ways for estimating errors
- We need to estimate the generalization error such that we can evaluate a classification model for generalization
- A good model: Low training error as well as low error during testing / validation

When am I Overfitting?

- Observing the performance of a model on a validation / development set can help detect overfitting
- Using a validation set:

  - The original training data set is divided into two smaller subsets: training subset and validation subset
  - Training set is used to build a classification model
  - The validation set is used to estimate the generalization error
  - Typically, the training set is far larger than the validation set
- Ideally, we expect the performance on the validation set to be similar to the performance on the test set

Performance on Validation

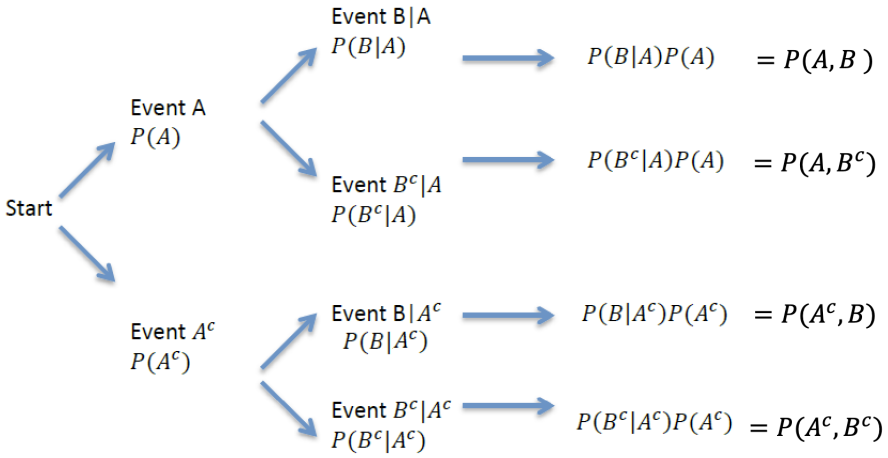


Bayes Classifier

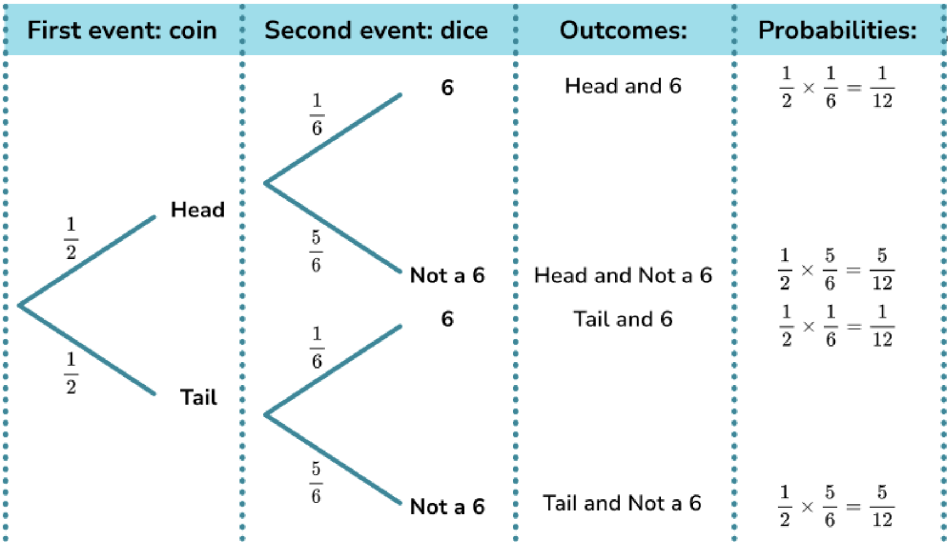
Conditional Probabilities

- $\Pr(\text{eaten by a shark}) = \text{Low}$
- $\Pr(\text{eaten by a shark} \mid \text{you are swimming with sharks}) = \text{A little more}$
- $\Pr(\text{eaten by a shark} \mid \text{swimming with sharks covered in chum}) = \text{High}$
- Marble Example
  - 5 blue and 5 red marbles in a bag
  - $\Pr(\text{selecting a blue marble}) = 0.5$
  - $\Pr(\text{selecting a blue marble} \mid \text{you removed all red marbles}) = 1$

Probability Tree Diagram

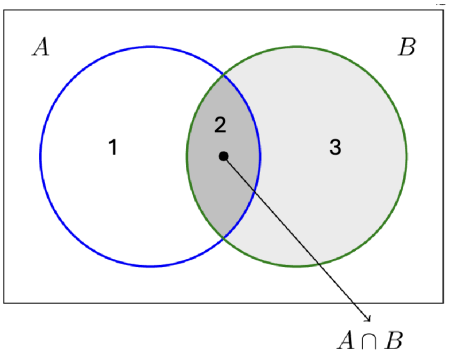


- Mutually exclusive events means they can't both occur. For example, we can't get both outcomes (heads and tails) on the coin flip at the same time.



Visualizing Conditional Probabilities

- $\Pr(A, B) = \Pr(A \cap B)$
- $\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$



- $\Pr(B) = 2 + 3$

- $\Pr(B) = \Pr(A, B) = \Pr\left(\bar{A}, B\right)$
- $\Pr(B) = \Pr(A) \cdot \Pr(B \mid A) = \Pr\left(\bar{A}\right) \cdot \Pr\left(B \mid \bar{A}\right)$