10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Overfitting
+
k-Nearest Neighbors

Matt Gormley
Lecture 4
Jan. 28, 2019

Q&A

Q: Why don't my entropy calculations match those on the slides?

A: H(Y) is conventionally reported in "bits" and computed using log base 2.
e.g., $H(Y) = - P(Y=0) \log_2 P(Y=0) - P(Y=1) \log_2 P(Y=1)$

Q: When and how do we decide to stop growing trees? What if the set of values an attribute could take was really large or even infinite?

A: We'll address this question for discrete attributes today. If an attribute is real-valued, there's a clever trick that only considers O(L) splits where L = # of values the attribute takes in the training set. Can you guess what it does?

Q: Why is entropy based on a sum of p(.) log p(.) terms?

A: We don't have time for a full treatment of why it *has* to be this, but we can develop the right intuition with a few examples…

Reminders

- **Homework 2: Decision Trees**
  - Out: Wed, Jan 23
  - Due: Wed, Feb 6 at 11:59pm
- **10601 Notation Crib Sheet**

INDUCTIVE BIAS
(FOR DECISION TREES)

Decision Tree Learning Example

Dataset:

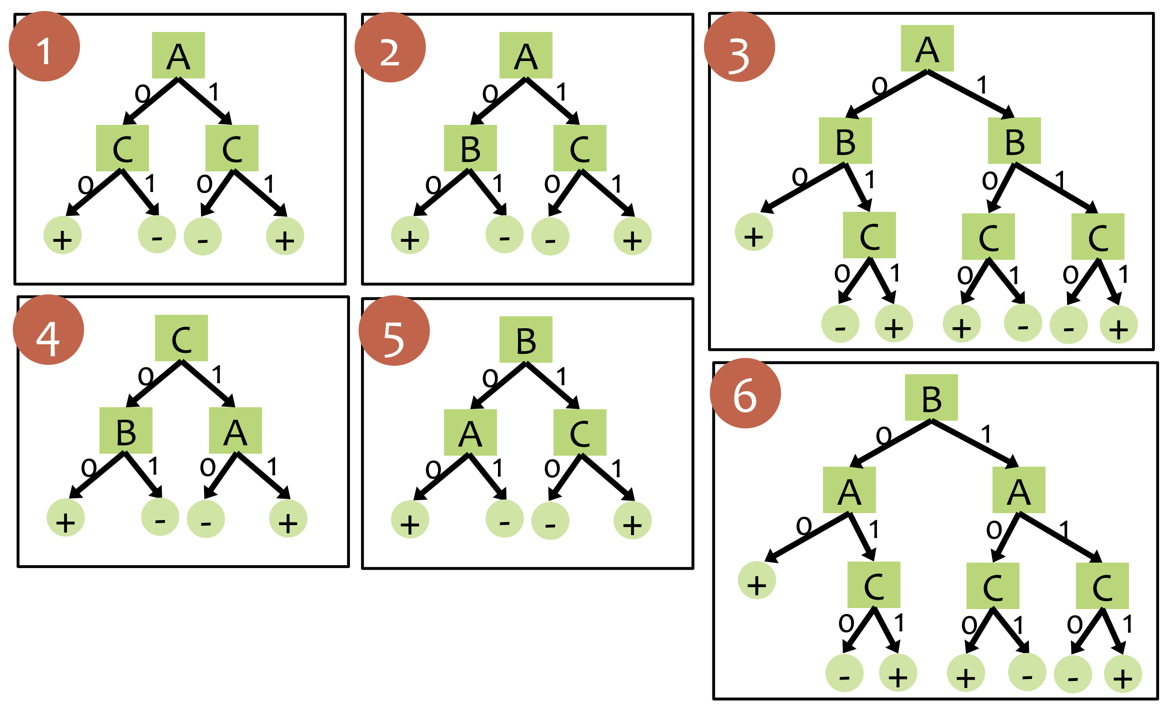
Output Y, Attributes A, B, C

Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

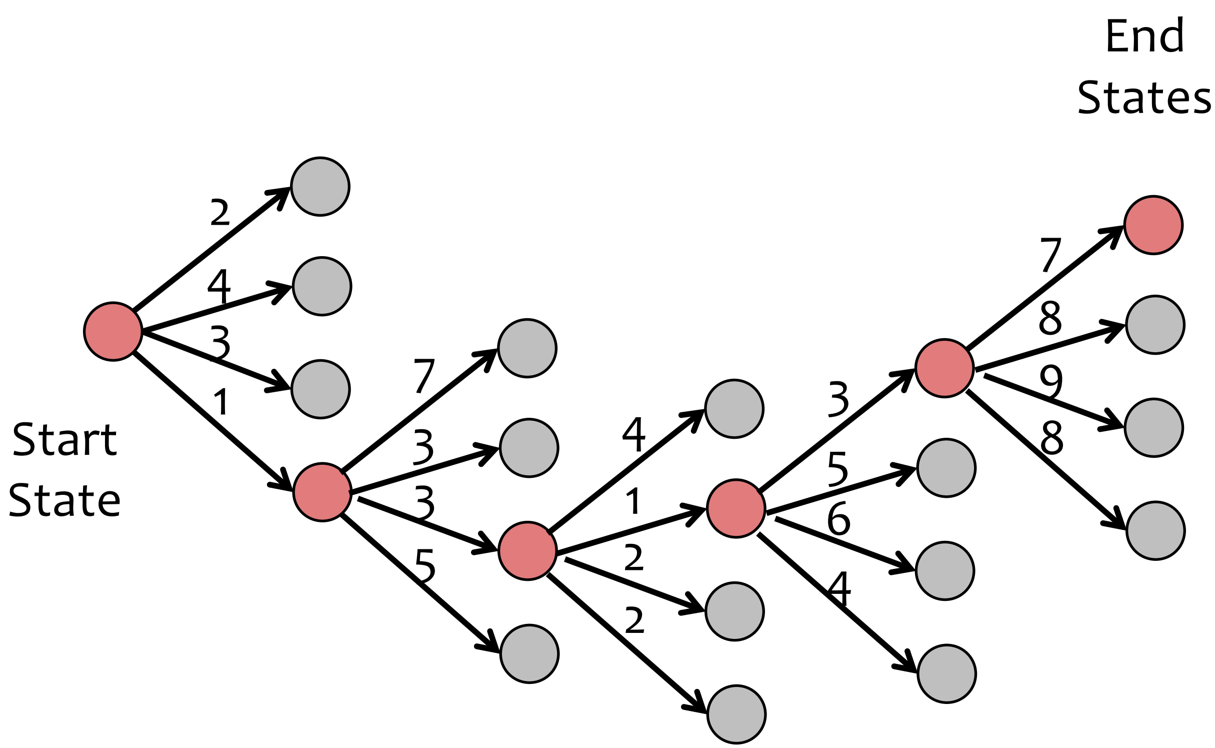
In-Class Exercise

Which of the following trees would be **learned by the ID3 algorithm** using “error rate” as the splitting criterion?

(Assume ties are broken alphabetically.)



Background: Greedy Search



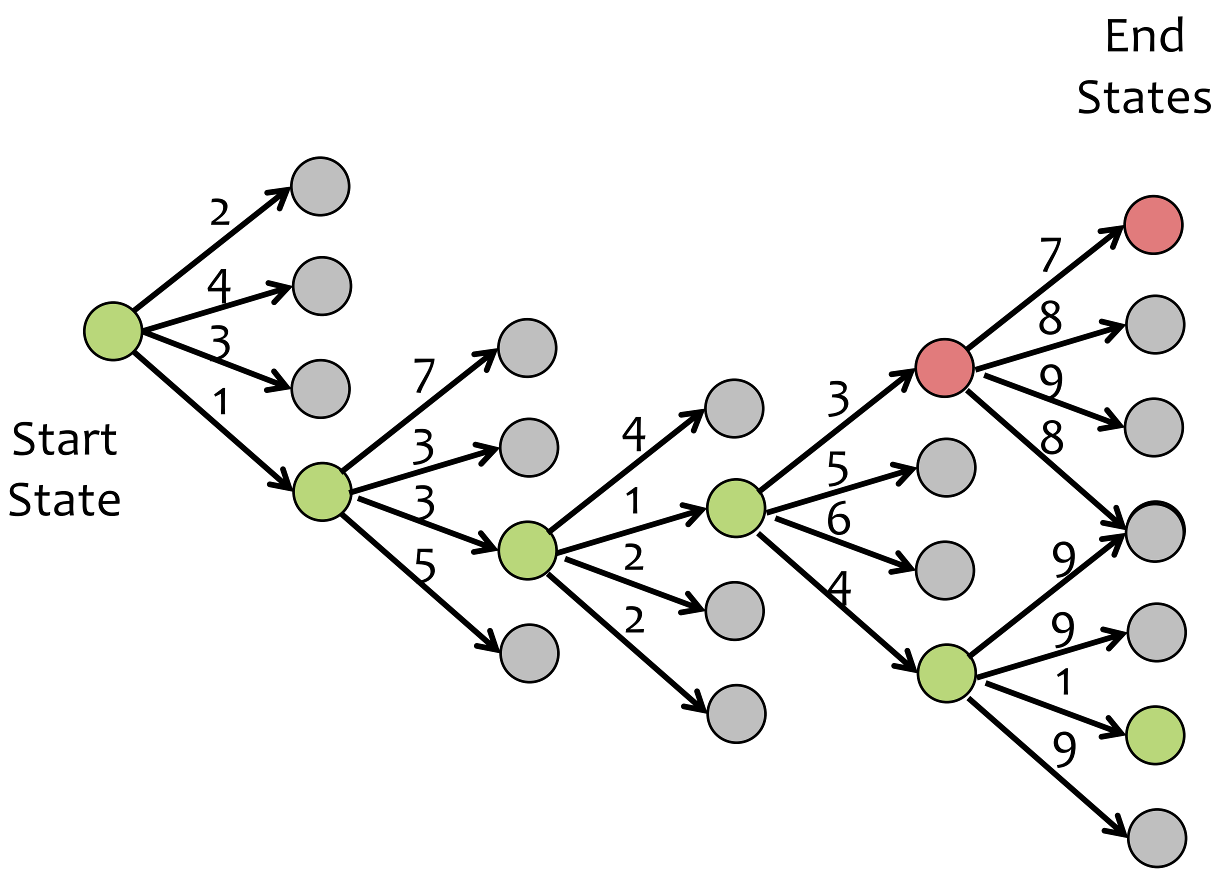
Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

Background: Greedy Search



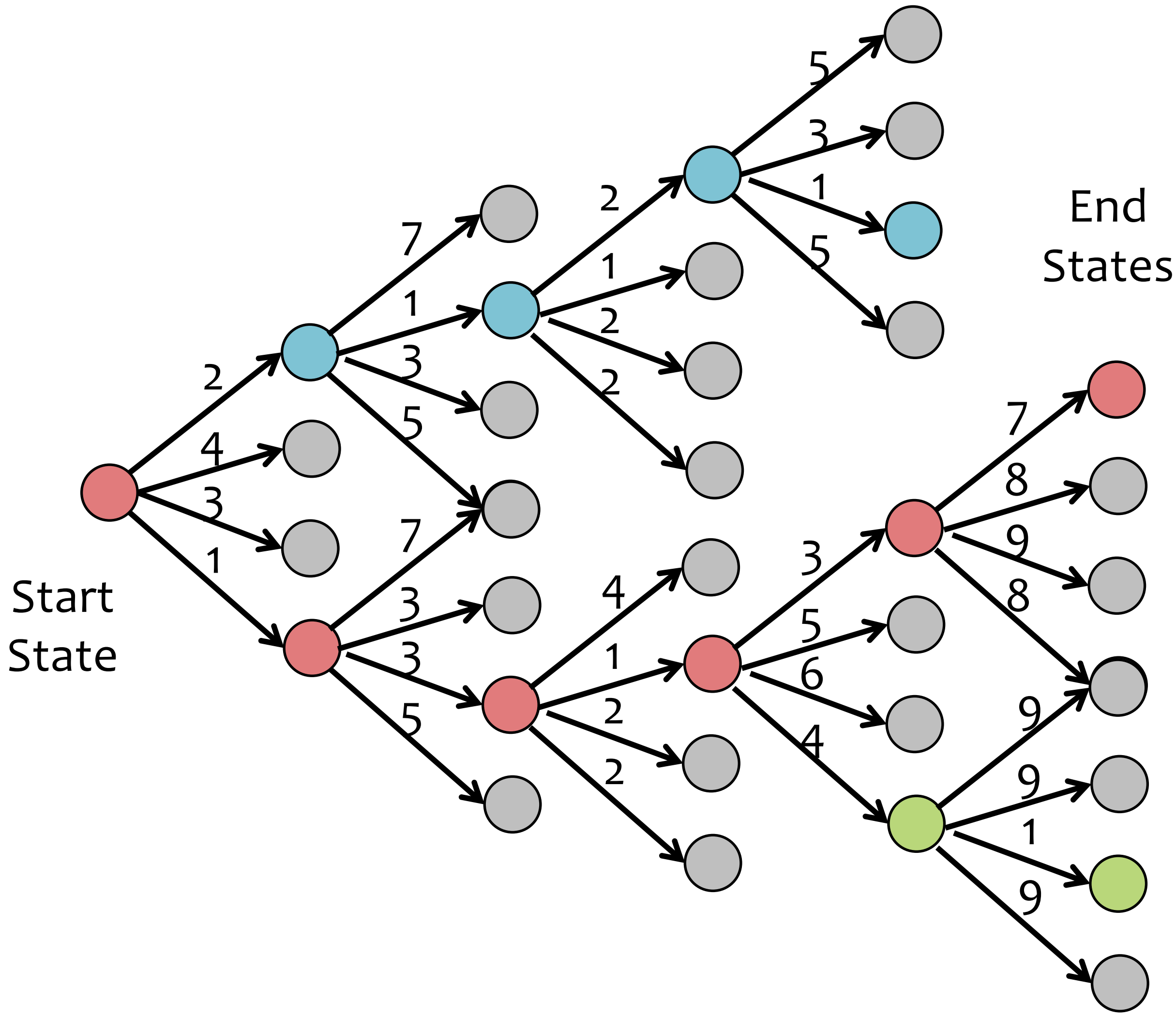
Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

Background: Greedy Search



Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

Decision Trees

Chalkboard

– ID3 as Search

DT: Remarks

Question: Which tree does ID3 find?

TODO: HIDE NEXT TWO SLIDES

DT: Remarks

Question: Which tree does ID3 find?

Definition:

We say that the **inductive bias** of a machine learning algorithm is the principal by which it generalizes to unseen examples

Inductive Bias of ID3:

Smallest tree that matches the data with high mutual information attributes near the top

Occam's Razor: (restated for ML)

Prefer the simplest hypothesis that explains the data

Decision Tree Learning Example

Dataset:

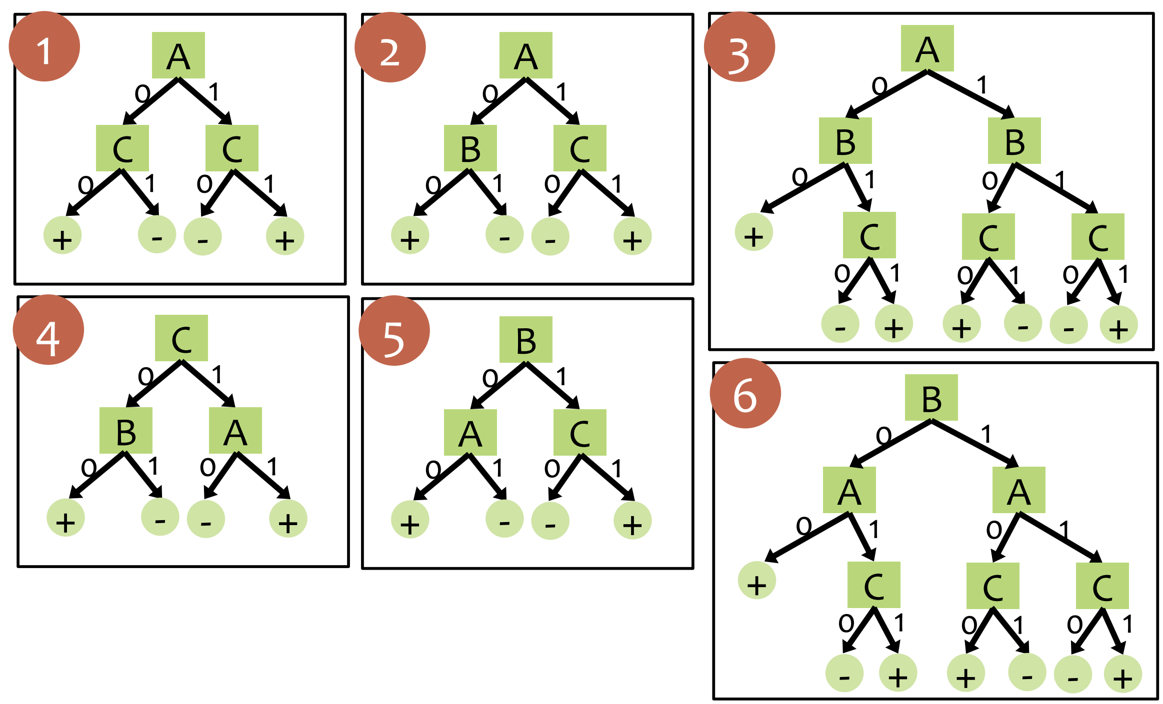
Output Y, Attributes A, B, C

Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

In-Class Exercise

Suppose you had an algorithm that found **the tree with lowest training error that was as small as possible,** which tree would it return?

(Assume ties are broken by choosing the smallest.)



OVERFITTING (FOR DECISION TREES)

Decision Tree Generalization

Question:

Which of the following would generalize best to unseen examples?

A. Small tree with low training accuracy
B. Large tree with low training accuracy
C. Small tree with high training accuracy
D. Large tree with high training accuracy

Answer:

Overfitting and Underfitting

Underfitting

- The model…
  - is too simple
  - is unable captures the trends in the data
  - exhibits too much bias
- *Example*: majority-vote classifier (i.e. depth-zero decision tree)
- *Example*: a toddler (that has **not** attended medical school) attempting to carry out medical diagnosis

Overfitting

- The model…
  - is too complex
  - is fitting the noise in the data
  - or fitting random statistical fluctuations inherent in the "sample" of training data
  - does not have enough bias
- *Example*: our "memorizer" algorithm responding to an "orange shirt" attribute
- *Example*: medical student who simply memorizes patient case studies, but does not understand how to apply knowledge to new patients

Overfitting

Consider a hypothesis h and its

- Error rate over training data: $error_{train}(h)$
- True error rate over all data: $error_{true}(h)$

We say h <u>overfits</u> the training data if

$$error_{true}(h) > error_{train}(h)$$

Amount of overfitting =

$$error_{true}(h) - error_{train}(h)$$

Slide from Tom Mitchell

Overfitting in Decision Tree Learning

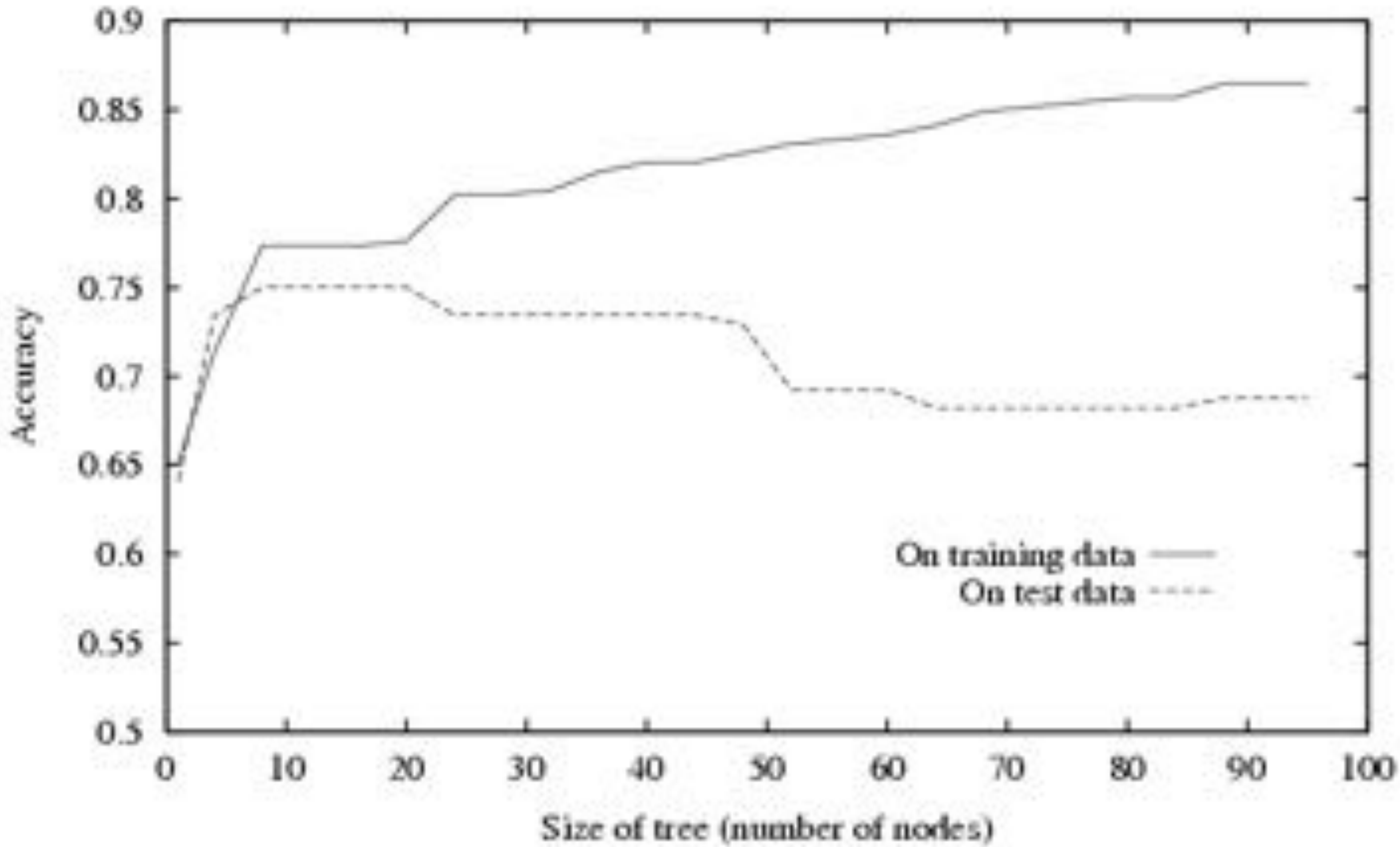


Figure from Tom Mitchell

How to Avoid Overfitting?

For Decision Trees…

1. Do not grow tree beyond some **maximum depth**
2. Do not split if splitting criterion (e.g. mutual information) is **below some threshold**
3. Stop growing when the split is **not statistically significant**
4. Grow the entire tree, then **prune**

Reduced-Error Pruning

Split data into *training* and *validation* set

Create tree that classifies *training* set correctly

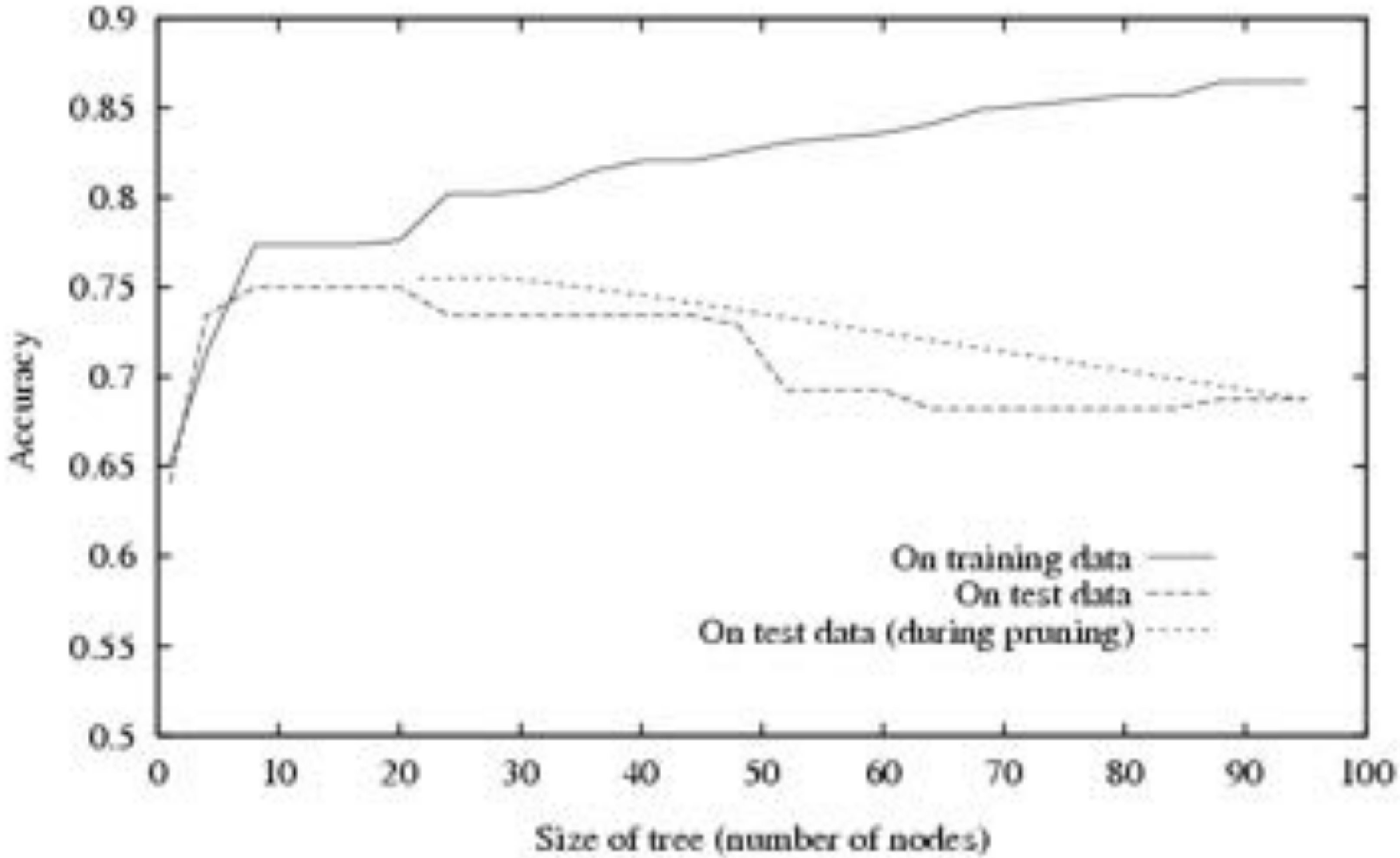
Do until further pruning is harmful:

1. Evaluate impact on *validation* set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves *validation* set accuracy

- produces smallest version of most accurate subtree
- What if data is limited?

Slide from Tom Mitchell

Effect of Reduced-Error Pruning



Slide from Tom Mitchell

Questions

- Will ID3 always include all the attributes in the tree?
- What if some attributes are real-valued? Can learning still be done efficiently?
- What if some attributes are missing?

Decision Trees (DTs) in the Wild

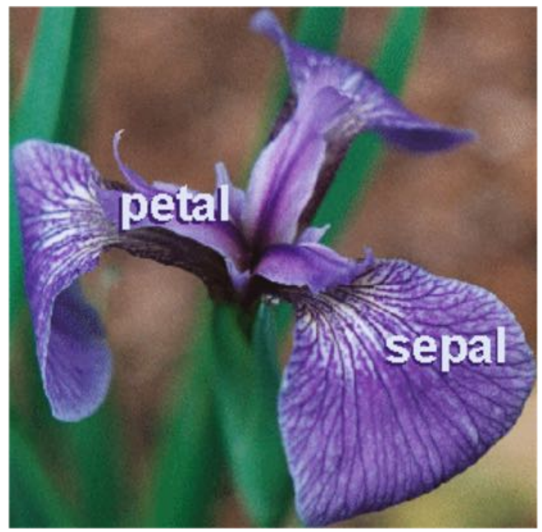
- DTs are one of the most popular classification methods for practical applications
  - Reason #1: The learned representation is **easy to explain** a non-ML person
  - Reason #2: They are **efficient** in both computation and memory
- DTs can be applied to a wide variety of problems including **classification, regression, density estimation,** etc.
- **Applications of DTs** include…
  - medicine, molecular biology, text classification, manufacturing, astronomy, agriculture, and many others
- **Decision Forests** learn many DTs from random subsets of features; the result is a very powerful example of an **ensemble method** (discussed later in the course)

DT Learning Objectives

You should be able to…

1. Implement Decision Tree training and prediction
2. Use effective splitting criteria for Decision Trees and be able to define entropy, conditional entropy, and mutual information / information gain
3. Explain the difference between memorization and generalization [CIML]
4. Describe the inductive bias of a decision tree
5. Formalize a learning problem by identifying the input space, output space, hypothesis space, and target function
6. Explain the difference between true error and training error
7. Judge whether a decision tree is "underfitting" or "overfitting"
8. Implement a pruning or early stopping method to combat overfitting in Decision Tree learning

CLASSIFICATION


petal
sepal

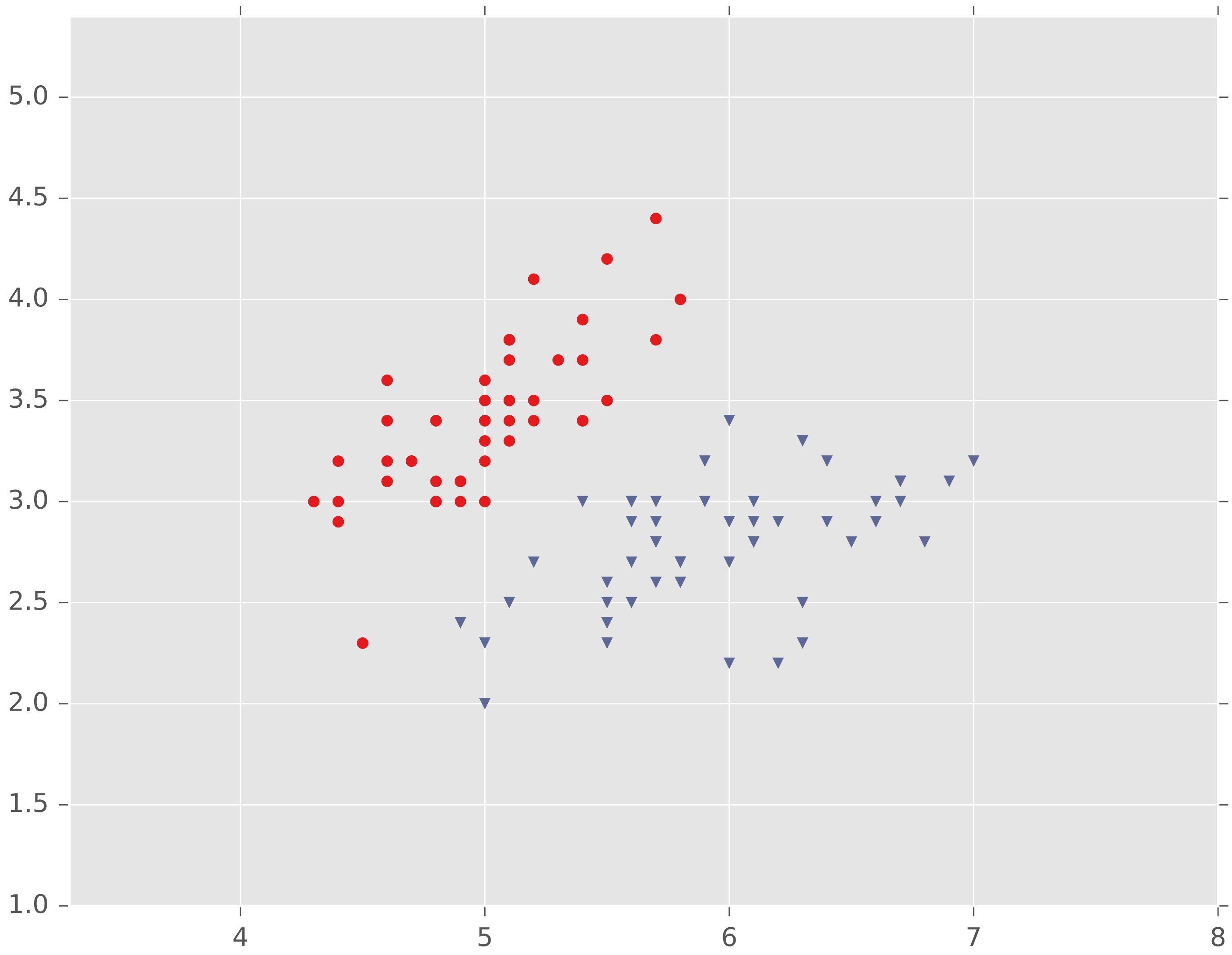
Fisher Iris Dataset

Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	6.7	3.0	5.0	1.7

Full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set

Fisher Iris Dataset



Classification

Chalkboard:

- Binary classification
- 2D examples
- Decision rules / hypotheses

K-NEAREST NEIGHBORS

k-Nearest Neighbors

Chalkboard:

- Nearest Neighbor classifier
- KNN for binary classification