10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Decision Trees

+

k-Nearest Neighbors

Matt Gormley
Lecture 3
January 24, 2018

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 1: Background**
  - **Out: Wed, Jan 17**
  - **Due: Wed, Jan 24 at 11:59pm**
  - unique policy for this assignment: we will grant (essentially) any and all extension requests
- **Homework 2: Decision Trees**
  - **Out: Wed, Jan 24**
  - **Due: Mon, Feb 5 at 11:59pm**

DECISION TREES

Tennis Example

Dataset:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure from Tom Mitchell

Tennis Example

Which attribute yields the best classifier?

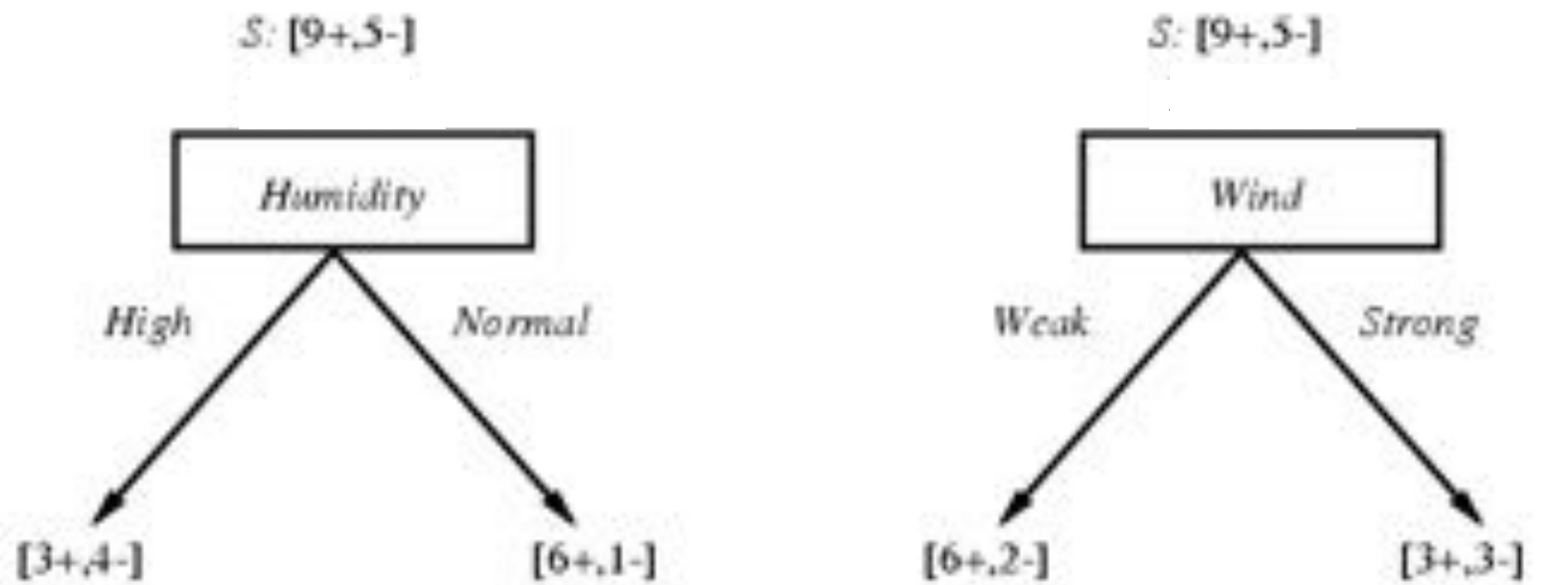


Figure from Tom Mitchell

Tennis Example

Which attribute yields the best classifier?

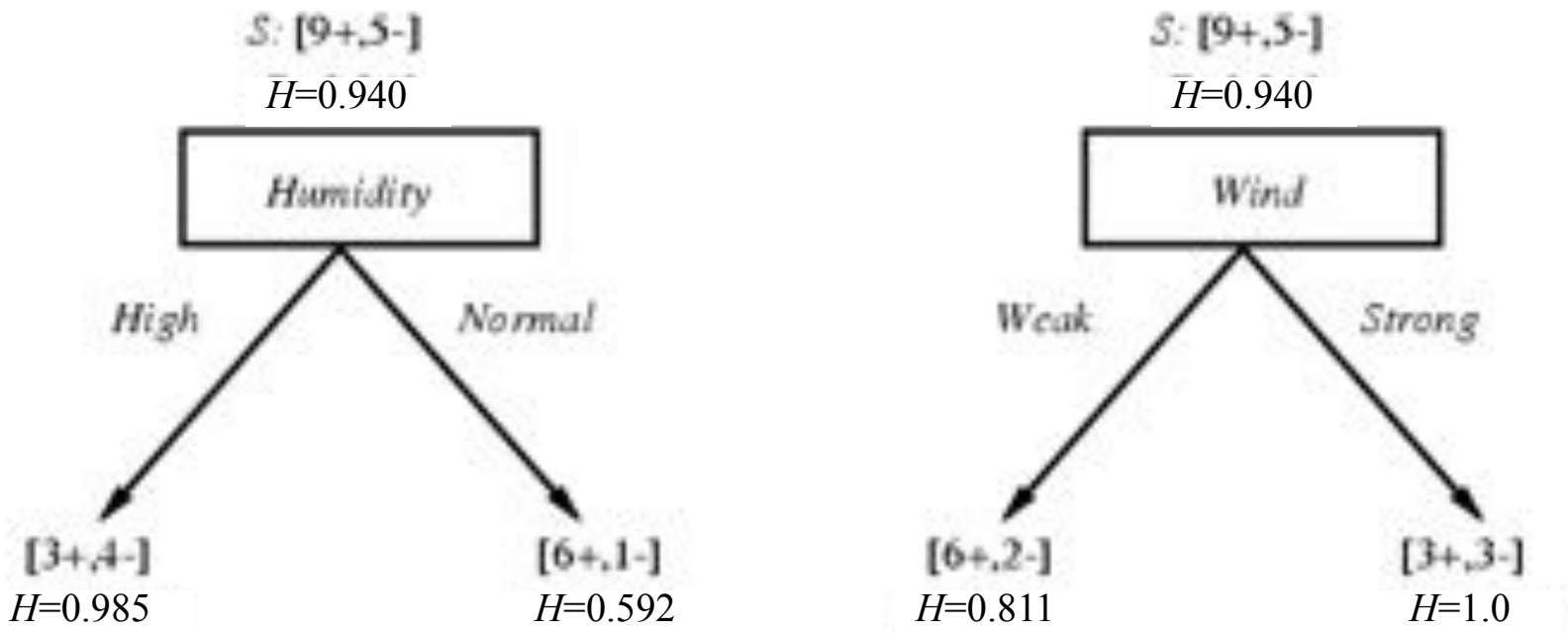


Figure from Tom Mitchell

Tennis Example

Which attribute yields the best classifier?

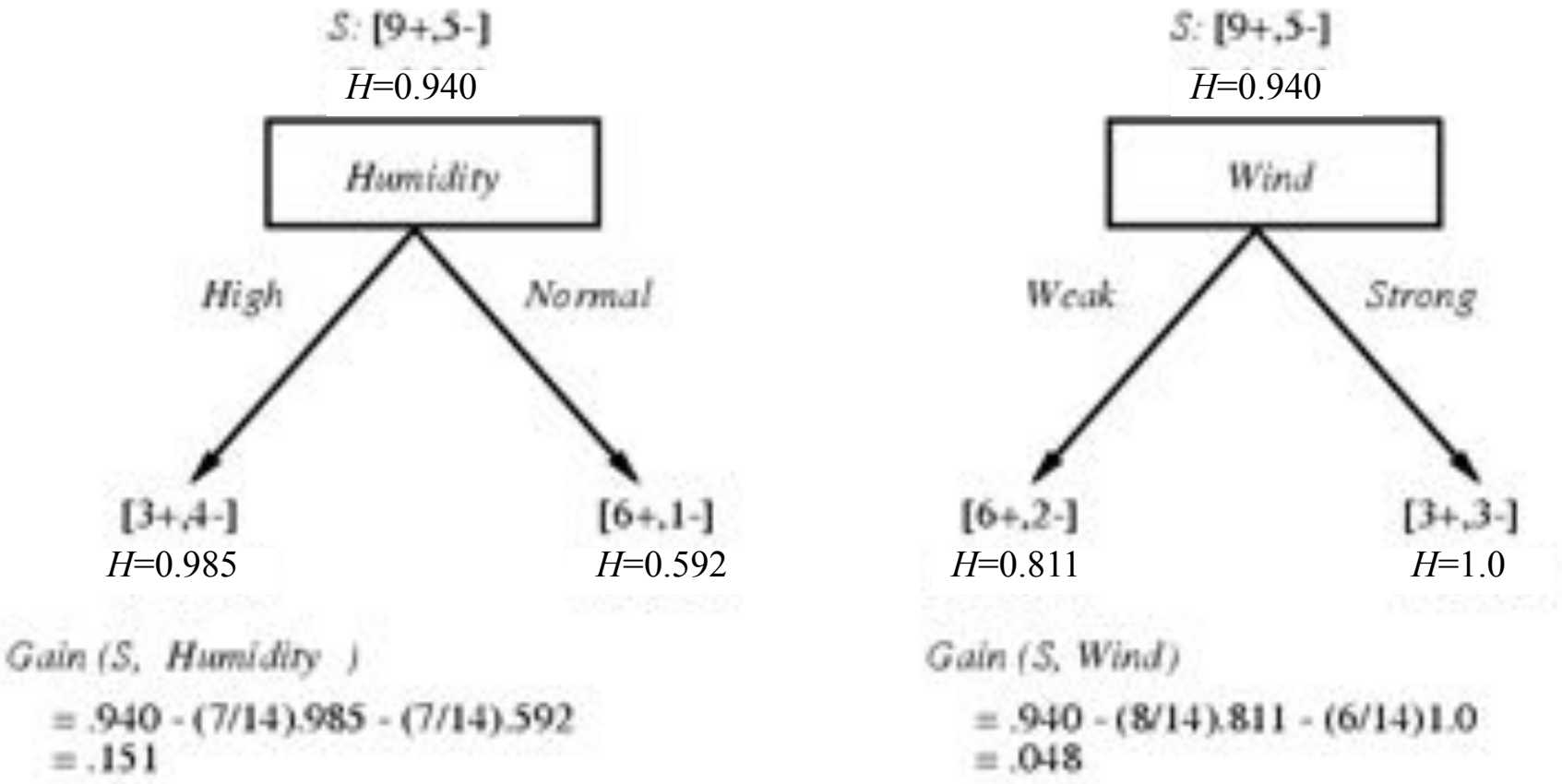
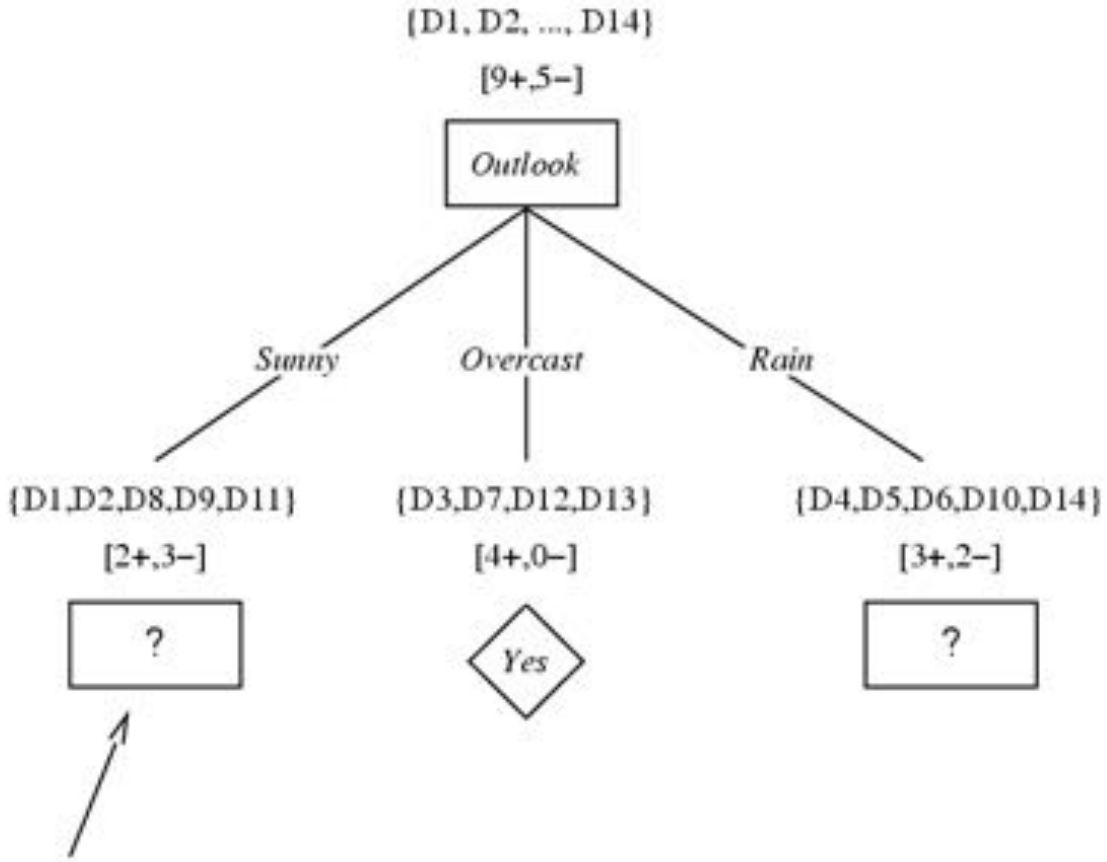


Figure from Tom Mitchell

Tennis Example



S_{sunny} = {D1,D2,D8,D9,D11}

Gain (S_{sunny}, *Humidity*) = .970 − (3/5) 0.0 − (2/5) 0.0 = .970

Gain (S_{sunny}, *Temperature*) = .970 − (2/5) 0.0 − (2/5) 1.0 − (1/5) 0.0 = .570

Gain (S_{sunny}, *Wind*) = .970 − (2/5) 1.0 − (3/5) .918 = .019

Figure from Tom Mitchell

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
0	1	0
0	1	0
1	1	0
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1

In-Class Exercise

1. Which attribute would **misclassification rate** select for the next split?
2. Which attribute would **information gain** select for the next split?
3. *Justify your answers.*

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
0	1	0
0	1	0
1	1	0
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1

Decision Trees

Chalkboard

- ID3 as Search
- Inductive Bias of Decision Trees
- Occam's Razor

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 overfits 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

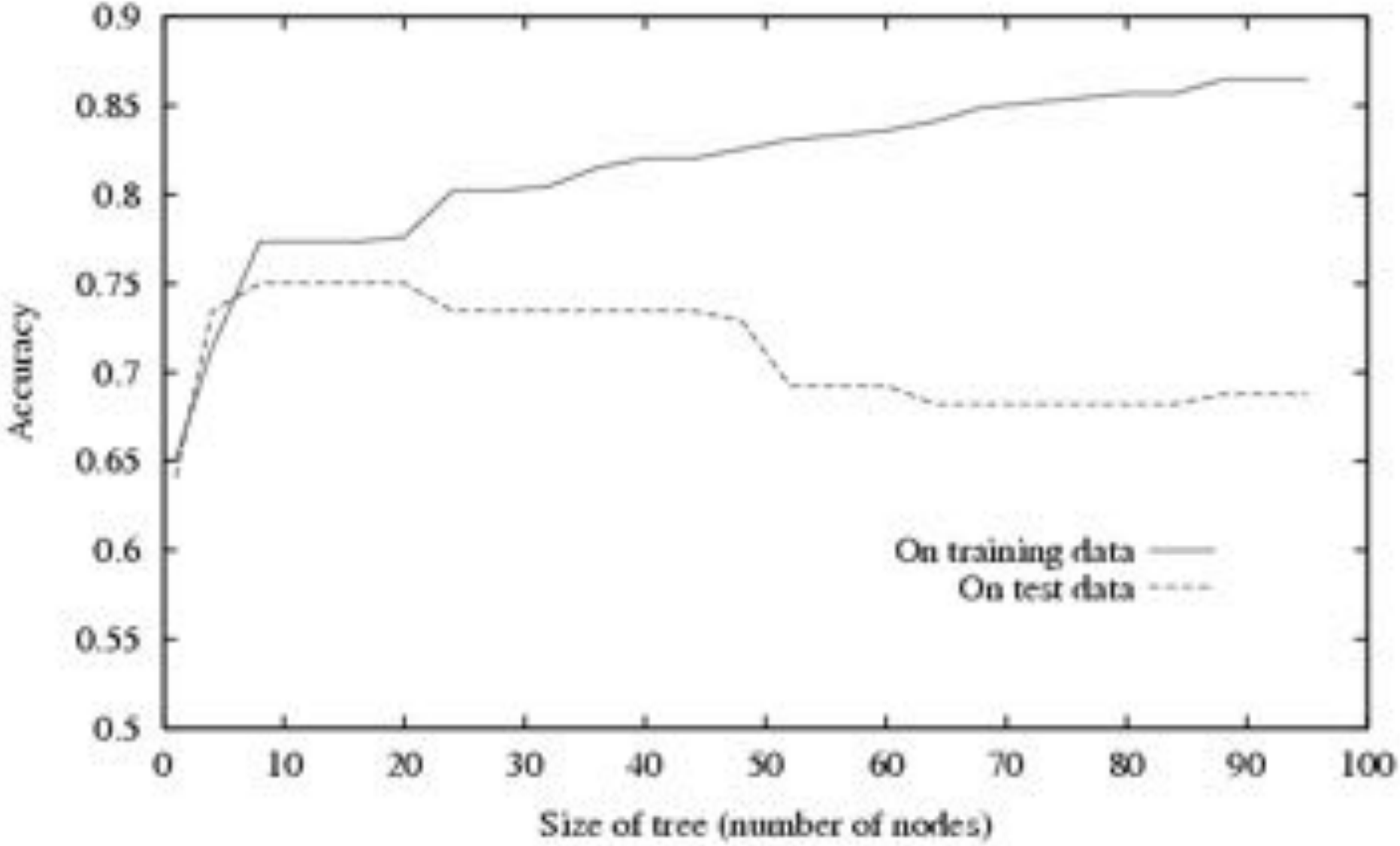


Figure from Tom Mitchell

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. Info. Gain) 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

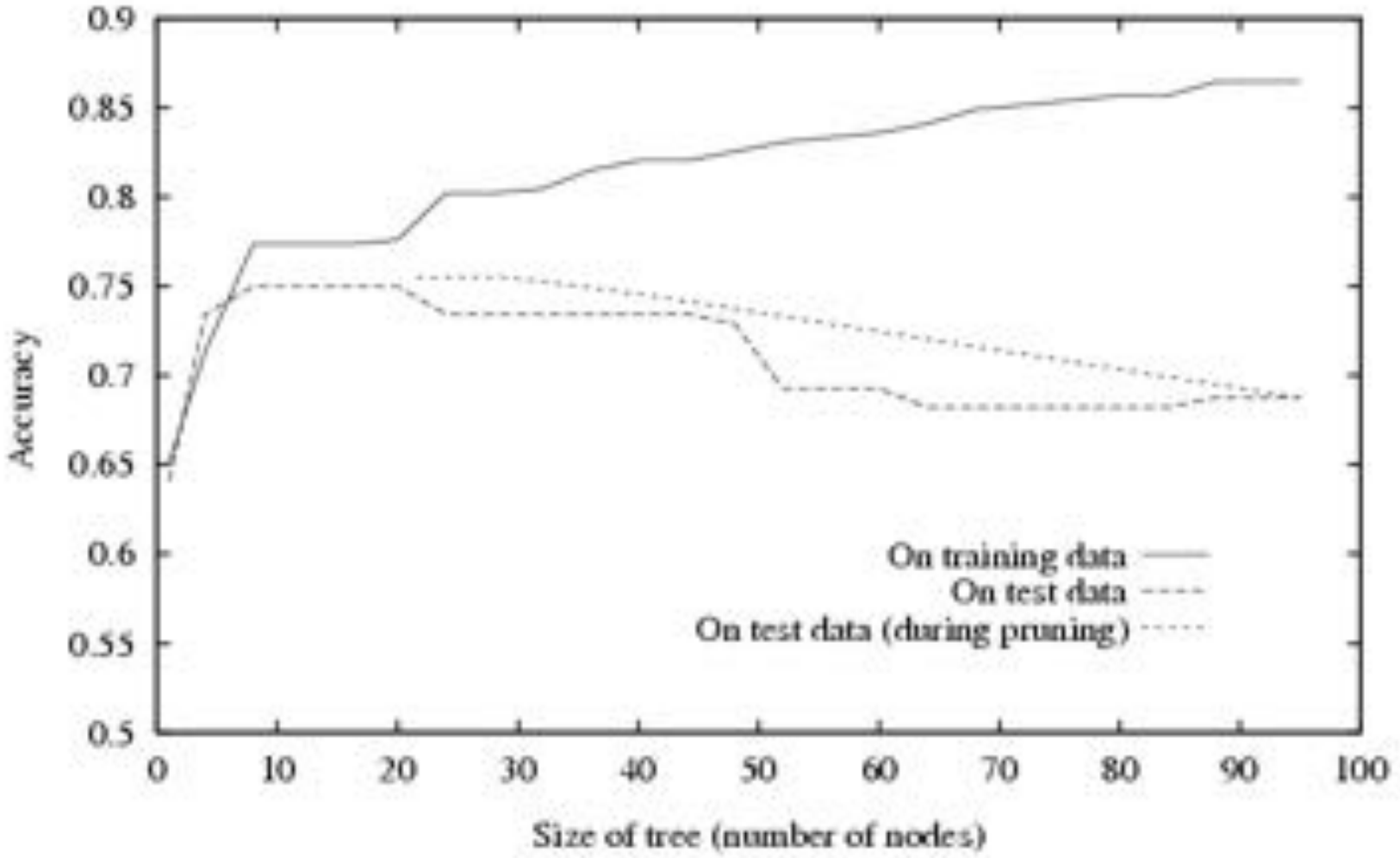
Do until further pruning is harmful:

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

KNN Outline

- **Classification**
  - Binary classification
  - 2D examples
  - Decision rules / hypotheses
- **k-Nearest Neighbors (KNN)**
  - Nearest Neighbor classification
  - k-Nearest Neighbor classification
  - Distance functions
  - Case Study: KNN on Fisher Iris Data
  - Case Study: KNN on 2D Gaussian Data
  - Special cases
  - Choosing k
- **Experimental Design**
  - Train error vs. test error
  - Train / validation / test splits
  - Cross-validation

CLASSIFICATION


petal
sepal

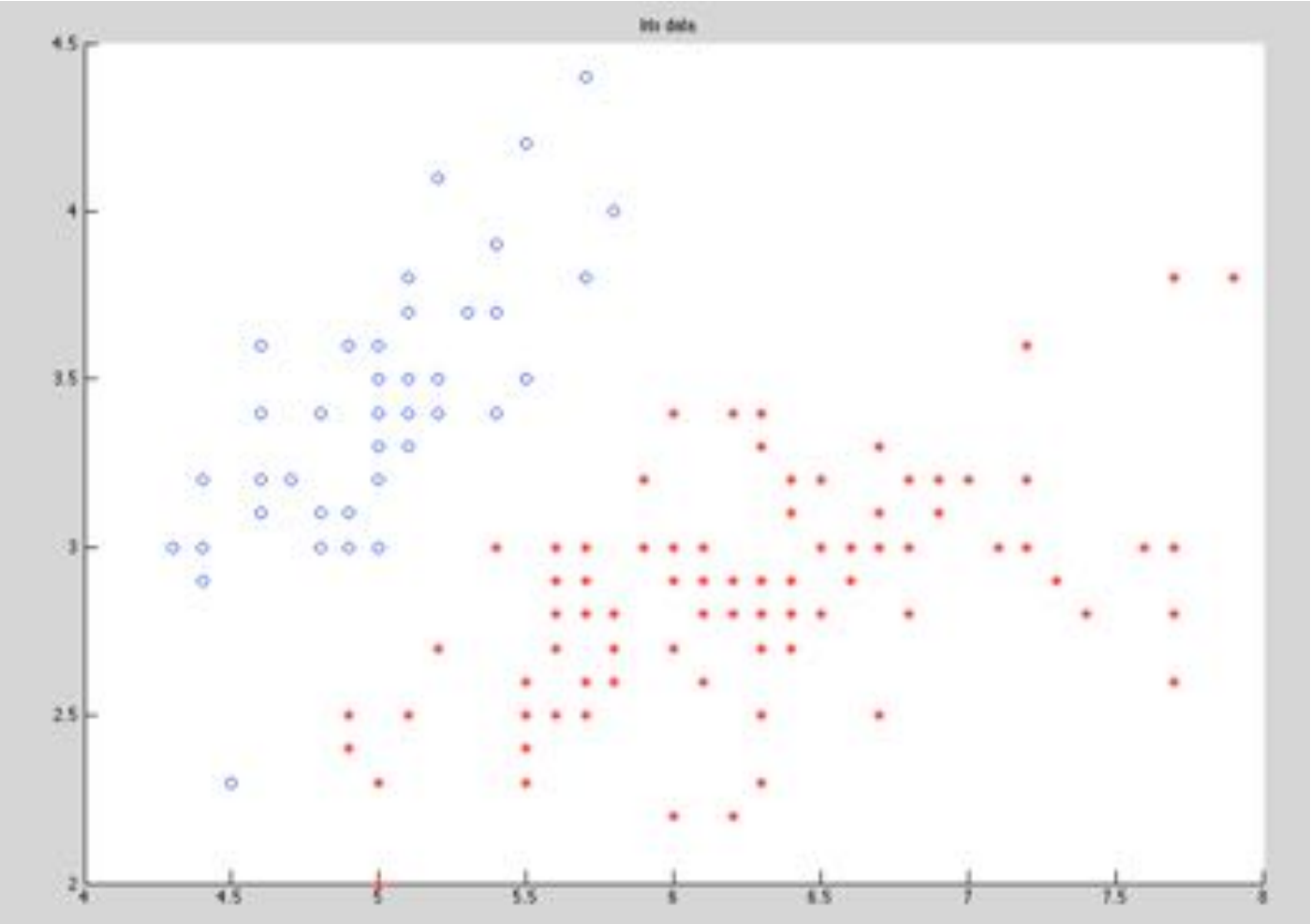
Fisher Iris Dataset

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:

- KNN for binary classification
- Distance functions
- Efficiency of KNN
- Inductive bias of KNN
- KNN Properties