



Variations in  
Decision  
Trees

B. Juliano

Introduction

Decision  
Trees

Decision Tree  
Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

# VARIATIONS IN DECISION TREES

## A Brief Introduction to Data Mining

**Ben Juliano**

**Department of Computer Science**  
California State University, Chico  
Chico, CA 95929-0410

<http://csci.ecst.csuchico.edu>

MATH Colloquium

February 25, 2011



# Introduction

## Variations in Decision Trees

B. Juliano

### Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Very Large Databases Abound**

According to a Focus.com study in February 2010

- Library of Congress  
5 million digital documents, 20 TB of text data
- Central Intelligence Agency  
comprehensive statistics on more than 250 countries and entities
- Amazon.com  
more than 42 TB of data
- YouTube.com  
at least 45 TB of data
- World Data Centre for Climate  
220 TB of web data, 6 PB of additional data

What about NASA? Google?



# Introduction

## Variations in Decision Trees

B. Juliano

### Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

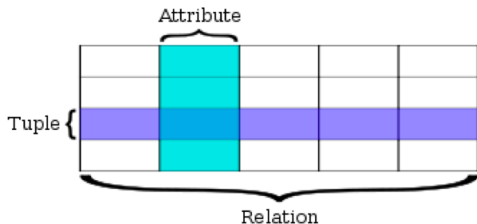
Summary

References

Resources

## • Traditional Database Operations

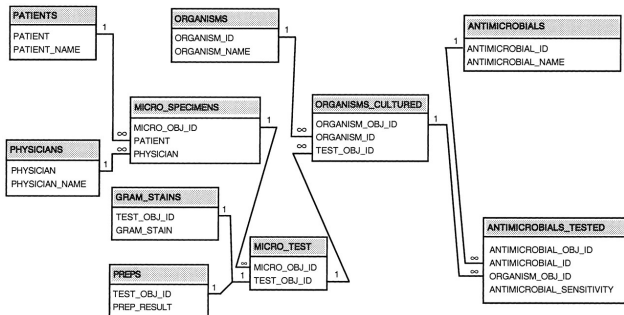
- Example: Relational databases (Codd, 1970)
  - Queries made against a relational database are expressed in a *relational calculus* or a *relational algebra*.





## • Traditional Database Operations

- Example: Relational databases (Codd, 1970)
  - Queries made against a relational database are expressed in a *relational calculus* or a *relational algebra*.





# Introduction

## Variations in Decision Trees

B. Juliano

### Introduction

### Decision Trees

### Decision Tree Induction

### Example

### Sample Run 1

### Improvements

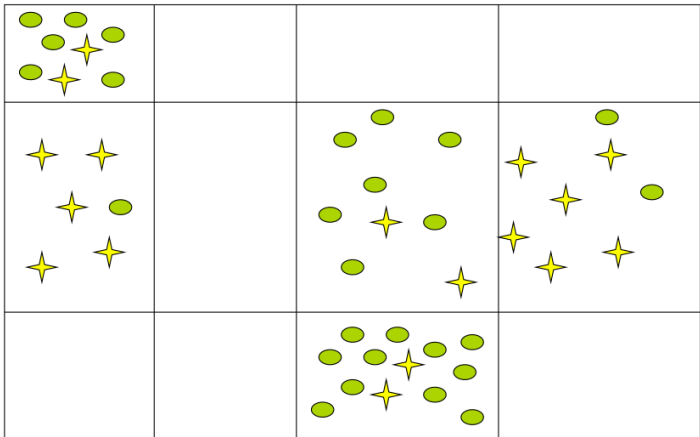
### Sample Run 2

### Summary

### References

### Resources

- Consider the following problem ...





# Introduction

## Variations in Decision Trees

B. Juliano

### Introduction

Decision  
Trees

Decision Tree  
Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Supervised vs. Unsupervised Learning**

- Supervised learning (**classification**)
  - Supervision: The training data are accompanied by labels indicating the class they belong to.
  - New data is classified based on what was learned from the training set.
- Unsupervised learning (**clustering**)
  - The training data has no class labels.
  - Given training data, the goal is to establish the existence of classes or clusters in the data.



# Introduction

## Variations in Decision Trees

B. Juliano

### Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Classification**

- Given a collection of records (*training set*); each record contains a set of attributes, and one of the attributes is the **class** (to be predicted).
- Find a model for the **class** attribute as a function of the values of the other attributes.
- Candidate models can be validated using a separate collection of records (*test set*).
- Goal: Previously unseen records should be assigned a **class** as accurately as possible.



# Introduction

## Variations in Decision Trees

B. Juliano

## Introduction

## Decision Trees

## Decision Tree Induction

## Example

## Sample Run 1

## Improvements

## Sample Run 2

## Summary

## References

## Resources

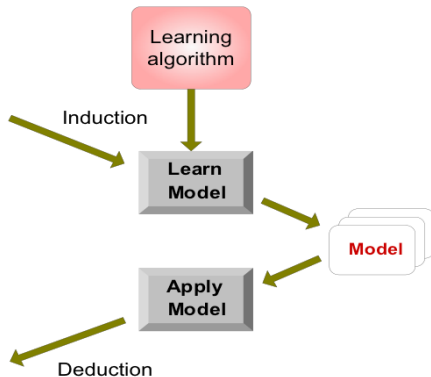
## • Illustrating a Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set





# Introduction

## Variations in Decision Trees

B. Juliano

### Introduction

Decision  
Trees

Decision Tree  
Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Classification Techniques**

- **decision tree**
- rule-based
- memory-based
- neural networks
- naive Bayes and Bayesian belief networks
- support vector machines



# Decision Trees

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

## Example: The Weather Data Set

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



# Decision Trees

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

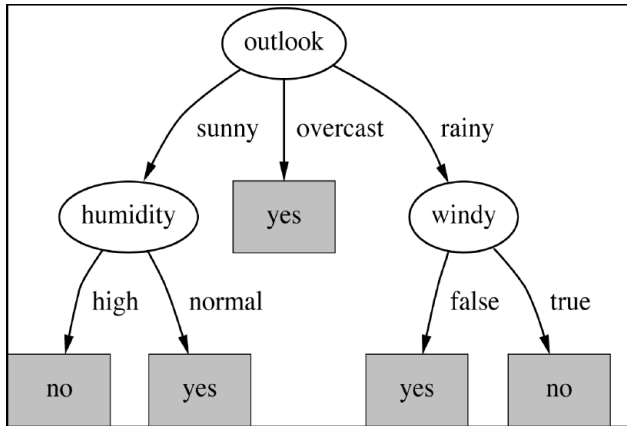
Sample Run 2

Summary

References

Resources

## Example: A Decision Tree





# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision  
Trees

**Decision Tree  
Induction**

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Top-down algorithm**

- First: Select attribute for root node  
Create branch for each possible attribute value.
- Then: Split instances into subsets  
One subset for each branch extending from the root node.
- Finally: Repeat recursively for each branch, using only instances in the subset that reaches a branch.
- Stop if all instances have the same class.



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

**Decision Tree Induction**

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

**Decision Tree Induction**

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **How to specify the attribute test condition?**
  - depends on attribute type
    - **nominal** – categorical  
e.g. a set of countries or a set of colors
    - **ordinal** – ordered categorical  
e.g. finishers in a race or pay scales
    - **continuous** – numeric  
e.g. temperature or weight
  - depends on number of ways to split
    - multi-way split
    - 2-way (binary) split



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

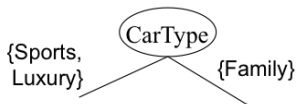
Resources

- Splitting **nominal** attributes

- **multi-way**: use as many partitions as distinct values



- **binary**: find optimal partitioning into two subsets





# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

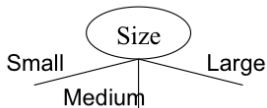
Summary

References

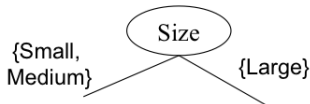
Resources

- Splitting **ordinal** attributes

- **multi-way**: use as many partitions as distinct values



- **binary**: find optimal partitioning into two subsets





# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

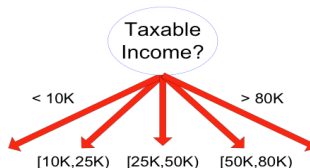
Summary

References

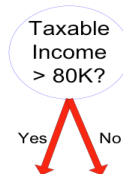
Resources

- Splitting **continuous** attributes

- **multi-way**: discretize to form ordinal attribute



- **binary**: consider all possible splits and select best cut





# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

**Decision Tree Induction**

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- Greedy strategy
  - Split the records based on an attribute test that optimizes certain criterion.
- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - **How to determine the best split?**
  - Determine when to stop splitting



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

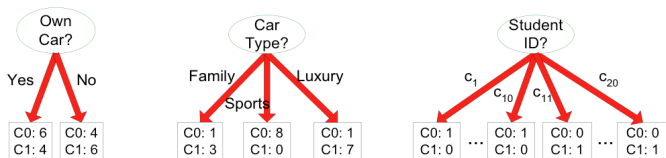
Summary

References

Resources

- **How to determine the best split?**

- Given: Before splitting, we have 10 records of Class 0 (C0) and 10 records of Class 1 (C1)



- Which split is the best?



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

**Decision Tree Induction**

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **How to determine the best split?**

- Greedy approach: prefer nodes with **homogeneous** class distribution
- Need a measure of node impurity

C0: 5
C1: 5

**Non-homogeneous,  
High degree of impurity**

C0: 9
C1: 1

**Homogeneous,  
Low degree of impurity**



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

**Decision Tree Induction**

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **ID3**, *Iterative Dichotomizer 3* (Quinlan, 1986)
  - attribute selection
    - choose attribute that produces “purest” nodes to generate the smallest tree
  - *impurity criterion*
    - *information gain* (a.k.a. Kullback–Leibler divergence) increases with the average purity of the subsets; so choose the attribute that gives the greatest information gain



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Computing Entropy** (Shannon, 1948)

- Given a probability distribution, *entropy* gives the amount of information (in bits) required to decide the classification of an arbitrary example

$$\begin{aligned} \text{entropy}(p_1, p_2, \dots, p_n) &= \sum_{i=1}^n -p_i \log_2 p_i \\ &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 \cdots \\ &\quad -p_n \log_2 p_n \end{aligned}$$



# Decision Tree Induction

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

### • Examples for computing entropy

- If  $p_1 = |C_1| = 0$  and  $p_2 = |C_2| = 6$   
$$\text{entropy}\left(\frac{0}{6}, \frac{6}{6}\right) = -\frac{0}{6} \log_2 \frac{0}{6} - \frac{6}{6} \log_2 \frac{6}{6} = \mathbf{0.00}$$
- If  $p_1 = |C_1| = 1$  and  $p_2 = |C_2| = 5$   
$$\text{entropy}\left(\frac{1}{6}, \frac{5}{6}\right) = -\frac{1}{6} \log_2 \frac{1}{6} - \frac{5}{6} \log_2 \frac{5}{6} = \mathbf{0.65}$$
- If  $p_1 = |C_1| = 2$  and  $p_2 = |C_2| = 4$   
$$\text{entropy}\left(\frac{2}{6}, \frac{4}{6}\right) = -\frac{2}{6} \log_2 \frac{2}{6} - \frac{4}{6} \log_2 \frac{4}{6} = \mathbf{0.92}$$
- If  $p_1 = |C_1| = 3$  and  $p_2 = |C_2| = 3$   
$$\text{entropy}\left(\frac{3}{6}, \frac{3}{6}\right) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = \mathbf{1.00}$$



# Example: *Outlook* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- From our weather dataset ...

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



# Example: *Outlook* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

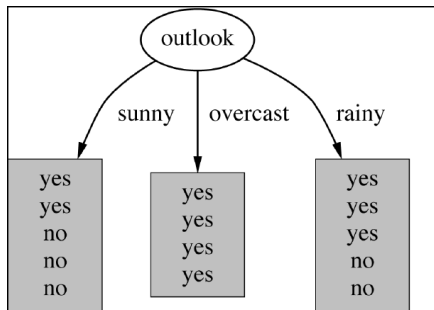
Sample Run 2

Summary

References

Resources

- consider splitting on *Outlook* attribute ...





## Example: *Outlook* attribute

### Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- *Outlook = Sunny*

$$\text{info}([2, 3]) = \text{entropy}\left(\frac{2}{5}, \frac{3}{5}\right) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.971 \text{ bits}$$

- *Outlook = Overcast*

$$\text{info}([4, 0]) = \text{entropy}\left(\frac{4}{4}, 0\right) = -1 \log_2 1 - 0 \log_2 0 = 0 \text{ bits}$$

- *Outlook = Rainy*

$$\text{info}([3, 2]) = \text{entropy}\left(\frac{3}{5}, \frac{2}{5}\right) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.971 \text{ bits}$$

- So, the expected information needed to classify objects in all subtrees of the *Outlook* attribute is

$$\begin{aligned} \text{info}([2, 3], [4, 0], [3, 2]) &= \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 \\ &= 0.693 \text{ bits} \end{aligned}$$



# Example: *Temperature* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- From our weather dataset ...

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



# Example: *Temperature* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

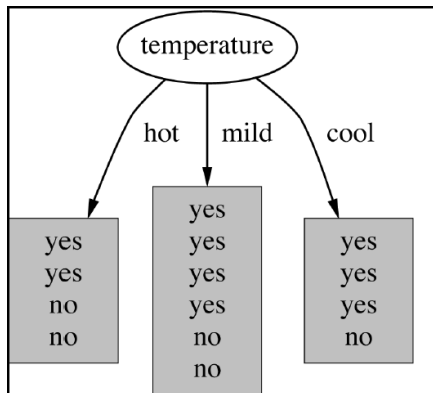
Sample Run 2

Summary

References

Resources

- consider splitting on *Temperature* attribute ...





## Example: *Temperature* attribute

### Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- *Temperature = Hot*

$$\text{info}([2, 2]) = \text{entropy}\left(\frac{2}{4}, \frac{2}{4}\right) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1.0 \text{ bits}$$

- *Temperature = Mild*

$$\text{info}([4, 2]) = \text{entropy}\left(\frac{4}{6}, \frac{2}{6}\right) = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.918 \text{ bits}$$

- *Temperature = Cool*

$$\text{info}([3, 1]) = \text{entropy}\left(\frac{3}{4}, \frac{1}{4}\right) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.811 \text{ bits}$$

- So, the expected information needed to classify objects in all subtrees of the *Temperature* attribute is

$$\begin{aligned} \text{info}([2, 2], [4, 2], [3, 1]) &= \frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811 \\ &= 0.911 \text{ bits} \end{aligned}$$



# Example: *Humidity* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- From our weather dataset ...

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



# Example: *Humidity* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

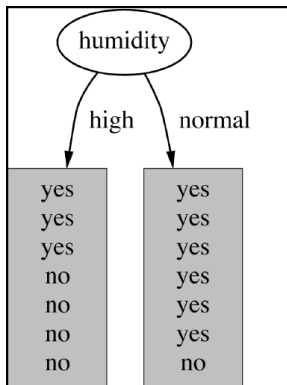
Sample Run 2

Summary

References

Resources

- consider splitting on *Humidity* attribute ...





# Example: *Humidity* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- *Humidity = High*

$$\text{info}([3, 4]) = \text{entropy}\left(\frac{3}{7}, \frac{4}{7}\right) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.985 \text{ bits}$$

- *Humidity = Normal*

$$\text{info}([6, 1]) = \text{entropy}\left(\frac{6}{7}, \frac{1}{7}\right) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.592 \text{ bits}$$

- So, the expected information needed to classify objects in all subtrees of the *Humidity* attribute is

$$\begin{aligned} \text{info}([3, 4], [6, 1]) &= \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.592 \\ &= 0.788 \text{ bits} \end{aligned}$$



# Example: *Windy* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

**Example**

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

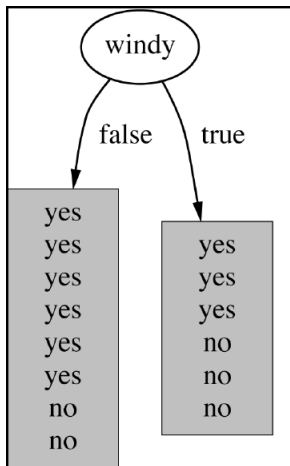
- From our weather dataset ...

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



# Example: *Windy* attribute

- consider splitting on *Windy* attribute ...





# Example: *Windy* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- *Windy* = *False*

$$\text{info}([6, 2]) = \text{entropy}\left(\frac{6}{8}, \frac{2}{8}\right) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.811 \text{ bits}$$

- *Windy* = *True*

$$\text{info}([3, 3]) = \text{entropy}\left(\frac{3}{6}, \frac{3}{6}\right) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} = 1.0 \text{ bits}$$

- So, the expected information needed to classify objects in all subtrees of the *Windy* attribute is

$$\begin{aligned} \text{info}([6, 2], [3, 3]) &= \frac{8}{14} \times 0.811 + \frac{6}{14} \times 1 \\ &= 0.892 \text{ bits} \end{aligned}$$



# Information Gain

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

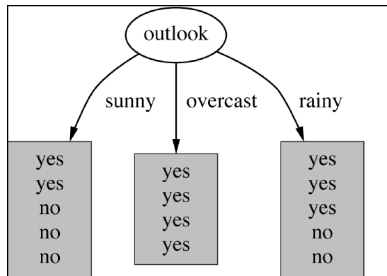
Sample Run 2

Summary

References

Resources

## • Computing Information Gain



- information gain = info *before* split – info *after* split

$$\begin{aligned} \text{gain}(\text{Outlook}) &= \text{info}([9, 5]) - \text{info}([2, 3], [4, 0], [3, 2]) \\ &= 0.940 - 0.693 = 0.247 \text{ bits} \end{aligned}$$



# Information Gain

## Variations in Decision Trees

B. Juliano

Introduction

Decision  
Trees

Decision Tree  
Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- **Computing Information Gain**

- So, the encoding information that would be gained by branching on each of the attributes under consideration is

$$\text{gain}(\textit{Outlook}) = 0.247 \text{ bits}$$

$$\text{gain}(\textit{Temperature}) = 0.029 \text{ bits}$$

$$\text{gain}(\textit{Humidity}) = 0.152 \text{ bits}$$

$$\text{gain}(\textit{Windy}) = 0.048 \text{ bits}$$

- Which attribute should we select for the root?



# We have a root node ... what's next?

## Variations in Decision Trees

B. Juliano

Introduction

Decision  
Trees

Decision Tree  
Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- Recall our top-down algorithm
  - First: Select attribute for root node  
Create branch for each possible attribute value.
  - Then: Split instances into subsets  
One subset for each branch extending from the root node.
  - **Finally: Repeat recursively for each branch, using only instances in the subset that reaches a branch.**
  - Stop if all instances have the same class.



# Splitting on *Temperature* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

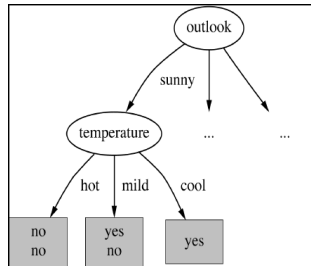
Sample Run 2

Summary

References

Resources

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



Applying the formulae to this subset of records gives us ...

$$\text{gain}(\text{Temperature}) = 0.571 \text{ bits}$$



# Splitting on *Humidity* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

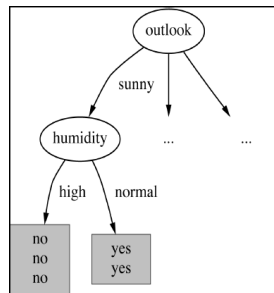
Sample Run 2

Summary

References

Resources

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



Applying the formulae to this subset of records gives us ...

$$\text{gain}(\textit{Humidity}) = 0.971 \text{ bits}$$



# Splitting on *Windy* attribute

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

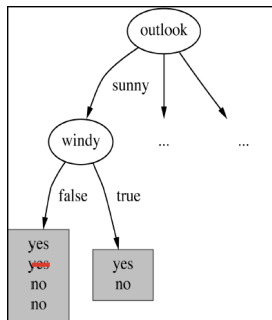
Sample Run 2

Summary

References

Resources

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



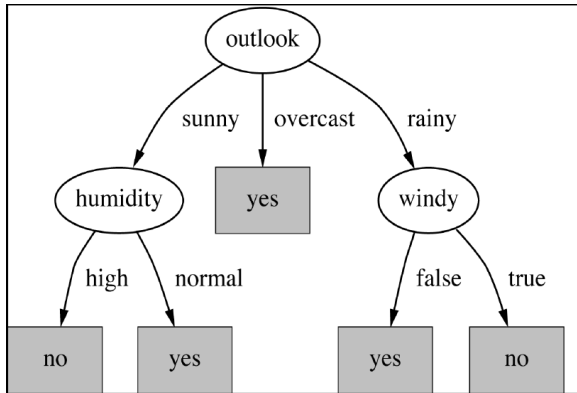
Applying the formulae to this subset of records gives us ...

$$gain(Windy) = 0.020 \text{ bits}$$



# Final Decision Tree

So, *Humidity* becomes the root at *Outlook = Sunny*.  
Repeating this process results in the following:





# Sample Run 1

## Variations in Decision Trees

B. Juliano

Introduction

Decision  
Trees

Decision Tree  
Induction

Example

**Sample Run 1**

Improvements

Sample Run 2

Summary

References

Resources



<http://www.cs.waikato.ac.nz/ml/weka>



<http://orange.biolab.si>



<http://rapid-i.com>



# Dealing with highly-branching attributes

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- Attributes with a large number of values (e.g. ID Code or Student ID) can be problematic
- Subsets are more likely to be pure if there is a large number of values
  - information gain tends to prefer splits that result in a large number of partitions, each being small but pure
  - *overfitting* – selection of an attribute that is non-optimal for prediction



# Dealing with highly-branching attributes

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

## • Computing Gain Ratio

$$\text{gainRatio}(\textit{attribute}) = \frac{\text{gain}(\textit{attribute})}{\text{info after split}}$$

- adjusts Information Gain by the entropy of the partitioning
- higher entropy partitioning (large number of small partitions) is penalized
- used in **ID3**'s successor: **C4.5** (Quinlan, 1993)



# Dealing with highly-branching attributes

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

Example: The weather data set with an *IDcode* attribute

ID code	Outlook	Temp.	Humidity	Windy	Play
A	Sunny	Hot	High	False	No
B	Sunny	Hot	High	True	No
C	Overcast	Hot	High	False	Yes
D	Rainy	Mild	High	False	Yes
E	Rainy	Cool	Normal	False	Yes
F	Rainy	Cool	Normal	True	No
G	Overcast	Cool	Normal	True	Yes
H	Sunny	Mild	High	False	No
I	Sunny	Cool	Normal	False	Yes
J	Rainy	Mild	Normal	False	Yes
K	Sunny	Mild	Normal	True	Yes
L	Overcast	Mild	High	True	Yes
M	Overcast	Hot	Normal	False	Yes
N	Rainy	Mild	High	True	No



# Dealing with highly-branching attributes

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

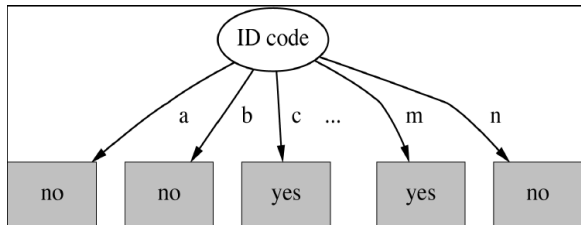
Sample Run 2

Summary

References

Resources

## • Computing Information Gain



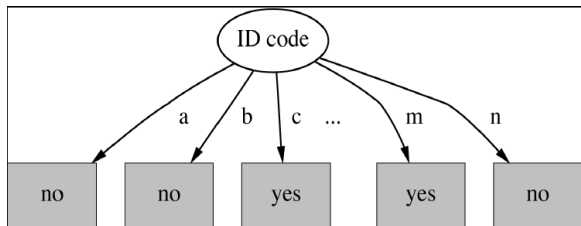
$$\begin{aligned} \text{gain}(IDcode) &= \text{info}([9, 5]) - \left( \text{info}([0, 1]) + \dots + \text{info}([0, 1]) \right) \\ &= 0.940 - 0 = 0.940 \end{aligned}$$

which means *IDcode* beats *Outlook* to be the root of the decision tree.



# Dealing with highly-branching attributes

- Computing Gain Ratio



$$\begin{aligned} \text{gainRatio}(\text{IDcode}) &= \frac{\text{gain}(\text{IDcode})}{\text{info}([1, 1, \dots, 1])} \\ &= \frac{0.940}{14 \times \left(-\frac{1}{14} \log_2 \frac{1}{14}\right)} = 0.246 \end{aligned}$$

but *IDcode* is a useless attribute (has zero average information value) and cannot be the root.



# Sample Run 2

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

**Sample Run 2**

Summary

References

Resources



<http://www.cs.waikato.ac.nz/ml/weka>



<http://orange.biolab.si>



<http://rapid-i.com>



# Summary

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

**Summary**

References

Resources

- **ID3 versus C4.5**
  - ID3 uses *information gain*
  - C4.5 can use either *information gain* or *gain ratio*
  - C4.5 can deal with
    - numeric/continuous attributes
    - missing values
    - noisy data
  - Alternate method: classification and regression trees
- **Decision trees ...**
  - requires little data preparation
  - are able to handle both categorical and numerical data
  - are simple to understand and interpret
  - generate models that can be statistically validated
  - perform well with large data in a short time



# References

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole, Monterey.
- E.F. Codd (1970). "A relational model of data for large shared data banks," *Communications of the ACM*, 13 (6): 377–387. doi:10.1145/362384.362685
- Focus.com Research (2010). "Top 10 largest databases in the world," <http://www.focus.com/fyi/operations/10-largest-databases-in-the-world>
- S. Kullback, R.A. Leibler (1951). "On information and sufficiency," *Annals of Mathematical Statistics*, 22 (1): 79–86. doi:10.1214/aoms/1177729694
- J.R. Quinlan (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco.
- J.R. Quinlan (1986). "Induction of decision trees," *Machine Learning*, 1 (1): 81–106.
- C.E. Shannon (1948). "A mathematical theory of communication," *Bell System Technical Journal*, 27: 379-423 (July), 623-656 (October).
- I.H. Witten, E. Frank, M.A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3e. Morgan Kaufmann, San Francisco.



# Resources

## Variations in Decision Trees

B. Juliano

Introduction

Decision Trees

Decision Tree Induction

Example

Sample Run 1

Improvements

Sample Run 2

Summary

References

Resources

- Rapid-I. *RapidMiner* open-source data mining system. <http://rapid-i.com>
- The R Foundation. *The R Project for Statistical Computing*. <http://www.r-project.org> (Note: See also R packages `rattle` data mining GUI and `mlbench` collection of machine learning benchmark problems.)
- University of Ljubljani Bioinformatics Laboratory. *Orange* open-source data visualization and analysis tool. <http://orange.biolab.si>
- University of Waikato Machine Learning Group. *Weka 3: Data Mining Software in Java*, <http://www.cs.waikato.ac.nz/ml/weka>
- Amazon Web Services. *AWS Public Data Sets*. <http://aws.amazon.com/datasets>
- Association for Computing Machinery. *ACM KDD Cup Center*. <http://www.sigkdd.org/kddcup>
- University of California – Irvine. *UCI Machine Learning Repository*. <http://archive.ics.uci.edu/ml>