Data Mining Concepts

Chapter 27

Outline

- Introduction
- Overview
- Association Rules
- Classification
- Clustering
- Approaches to Other Data Mining Problems
- Applications of Data Mining
- Commercial Data Mining Tools

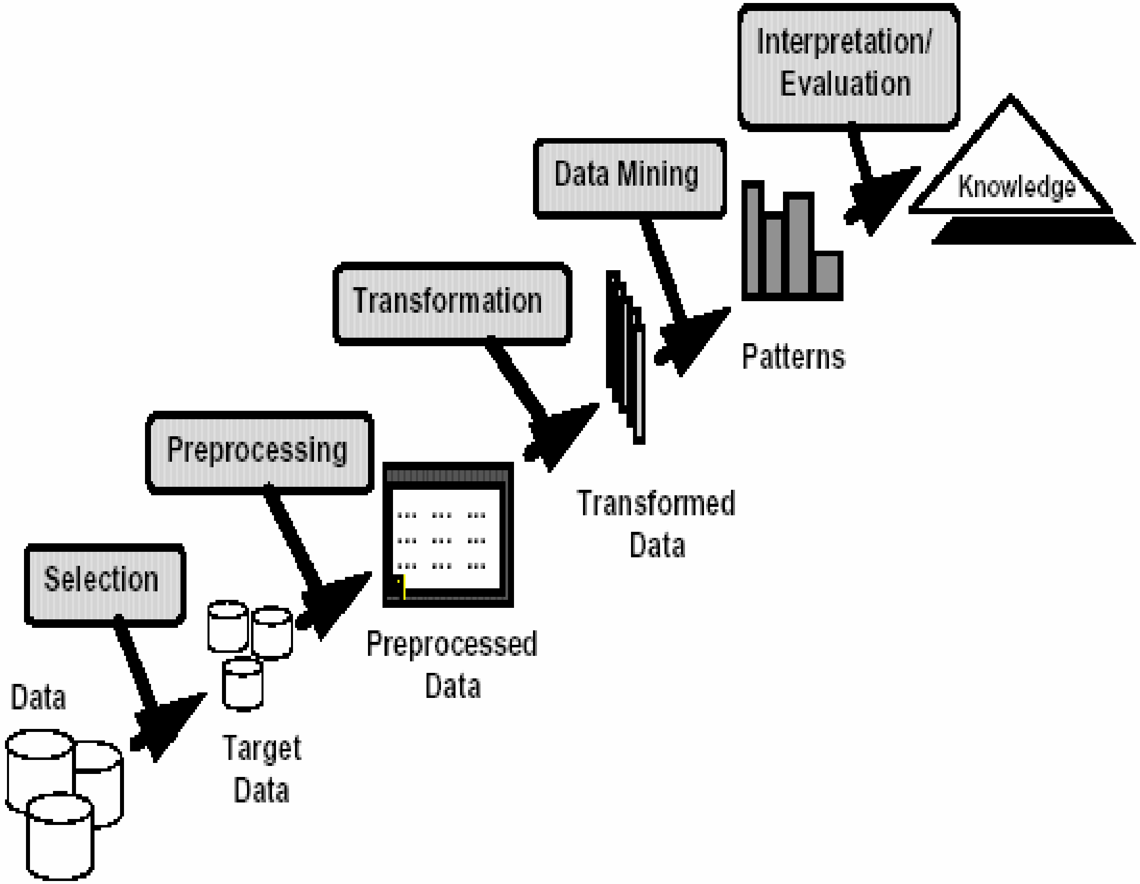
- Introduction

- Data Mining (DM)is the discovery of new information in terms of patterns or rules from vast amount of data.
- It is part of knowledge discovery in a database (KDD). (KDD will be briefly explained later).
- It uses techniques from such areas as:
  - Machine learning
  - Statistics
  - Neural networks
  - Genetic algorithms

- Overview ...

- knowledge discovery in a database goes through 6 phases

1. Data selection
2. Data cleansing
3. Enrichment
4. Data transformation
5. Data mining
6. Reporting



… - Overview …

- Goals of data mining and knowledge discover
  - **Prediction**: DM can show how certain attributes within the data will behave in the future.
  - **Identification**: DM can be used to identify the existence of an item or event or an activity
  - **Classification**: DM can partition the data so that the different partitions can be identified based on combinations of parameters.
  - **Optimization**: DM can be used to optimize the use of limited resources such as time or space or money or material.

… – Overview

- The knowledge discovered during DM can be described as

1. **Association rules**: correlate the presence of a set of items with another range of values for another set of variables
2. **Classification hierarchies**: create hierarchies of classes
3. **Sequential patterns**: sequence of actions or events
4. **Patterns within time series**: similarities detected within positions of the time series
5. **Categorization and segmentation**: partition a given population of events or items into sets of “similar” elements

- Association Rules ...

- An **association rule** is of the form $X \Rightarrow Y$ where $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_m\}$ are sets of distinct items
  The rule states that if a customer buys X, he is also likely to buy Y

- **Support** for the rule LHS ⇒ RHS is the percentage of transactions that hold all the items in the union, the set LHS ∪ RHS.

- **Confidence** for the rule LHS ⇒ RHS is the percentage (fraction) of all transactions that include items in LHS and out of these the ones that include items of RHS.

... - Association Rules ...

- Example:

Transaction id	Time	items bought
101	6:35	milk, bread, cookies, juice
792	7:38	milk, juice
1130	8:05	milk, eggs
1735	8:40	bread, cookies, coffee

Milk ➔ Juice, 50% support, 66.7% confidence
Bread ➔ Juice, 25% support, 50% confidence

… - Association Rules …

- The goal of mining association rules is to generate all possible rules that exceed some minimum user-specified support and confidence thresholds.

- The problem of mining association rules is thus decomposed into two sub-problems:
  - Generate all item sets that have a support that exceeds the threshold. These sets of items are called **large itemsets.**
  - For each large item set, all the rules that have a minimum confidence are generated as follows:
    for a large itemset X and $Y \subset X$, let $Z = X - Y$;
    then if support (X)/support (Z) $\Rightarrow$ minimum confidence, the rule $Z \Rightarrow Y$ (i.e., $X - Y \Rightarrow Y$) is a valid rule.

... - Association Rules ...

Basic Algorithms for Finding Association Rules

- The current algorithms (Apriori Algorithm) that find large itemsets are designed to work as follows:
  - Test the support for itemsets of length 1, called 1-itemsets, by scanning the database. Discard those that do not meet minimum required support.
  - Extend the large 1-itemsets into 2-itemsets by appending one item each time, to generate all candidate itemsets of length two. Test the support for all candidate itemsets by scanning the database and eliminate those 2-itemsets that do not meet the minimum support.
  - Repeat the above steps; at step *k,* the previously found (*k* - 1) itemsets are extended into *k*-itemsets and tested for minimum support.
  - The process is repeated until no large itemsets can be found.

... - Association Rules

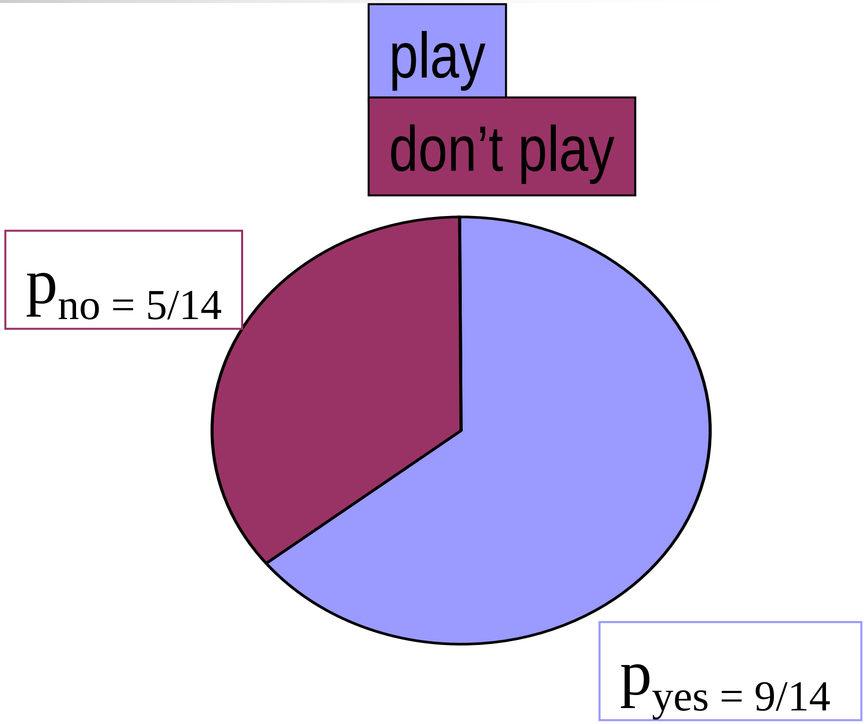
- **Apriori Algorithm** is based on the following 2 properties:
  1. **Antimonotonicity**: A subset of a large itemset must also be large.
  2. **Downward closure**: A superset of a small itemset is also small.
- Several other algorithms have been proposed to mine association rules:
  - Sampling algorithms
  - Frequent-pattern tree algorithm
  - Partition algorithm

- Classification

- Learn a function that assigns a record to one of several predefined classes
  - A.k.a. supervised learning
  - Given a set of training data set with a group of attributes and a target
  - To predict the value of target
- Techniques of classification
  - Decision tree
  - Neural networks

-- Decision trees: Attribute selection ...

outlook	temperature	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



- maximal gain of information
- maximal reduction of Entropy = $- p_{yes} \log_2 p_{yes} - p_{no} \log_2 p_{no}$

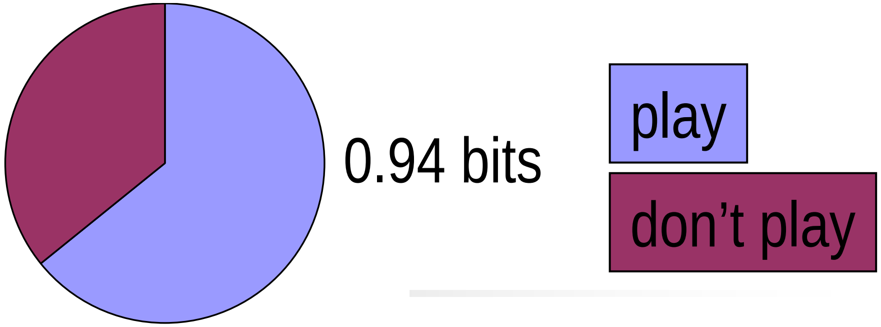
$$= - 9/14 \log_2 9/14 - 5/14 \log_2 5/14$$

$$= \mathbf{0.94\ bits}$$

http://www-lmmb.ncifcrf.gov/~toms/paper/primer/latex/index.html

http://directory.google.com/Top/Science/Math/Applications/Information_Theory/Papers/

... - Decision trees
Attribute selection



maximal information gain

	play	don't play
[illegible]	[illegible]	3
[illegible]	[illegible]	0
[illegible]	[illegible]	2

outlook

	play	don't play
high	3	4
normal	6	1

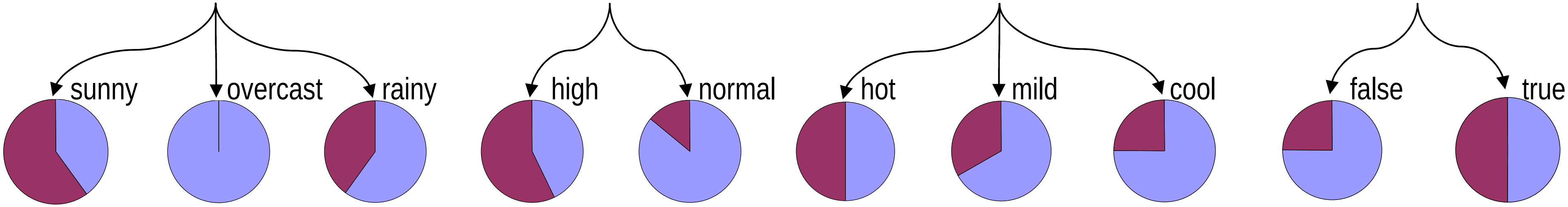
humidity

	play	don't play
hot	2	2
mild	4	2
cool	3	1

temperature

	play	don't play
FALSE	6	2
TRUE	3	3

windy



amount of information required to specify class of an example given that it reaches node

0.97 bits * 5/14 | 0.0 bits * 4/14 | 0.97 bits * 5/14 → + = 0.69 bits

gain: 0.25 bits

0.98 bits * 7/14 | 0.59 bits * 7/14 → + = 0.79 bits

gain: 0.15 bits

1.0 bits * 4/14 | 0.92 bits * 6/14 | 0.81 bits * 4/14 → + = 0.91 bits

gain: 0.03 bits

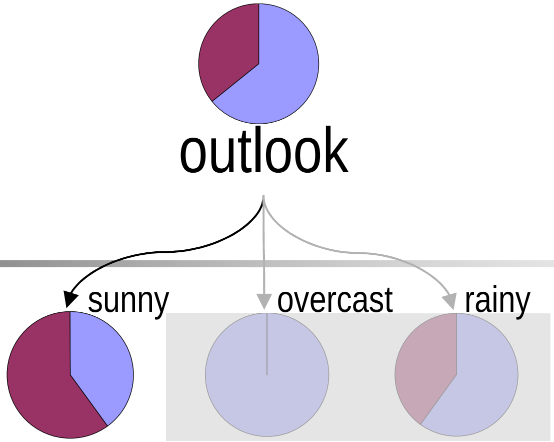
0.81 bits * 8/14 | 1.0 bits * 6/14 → + = 0.89 bits

gain: 0.05 bits

Decision trees
Building

play

don't play



outlook

sunny

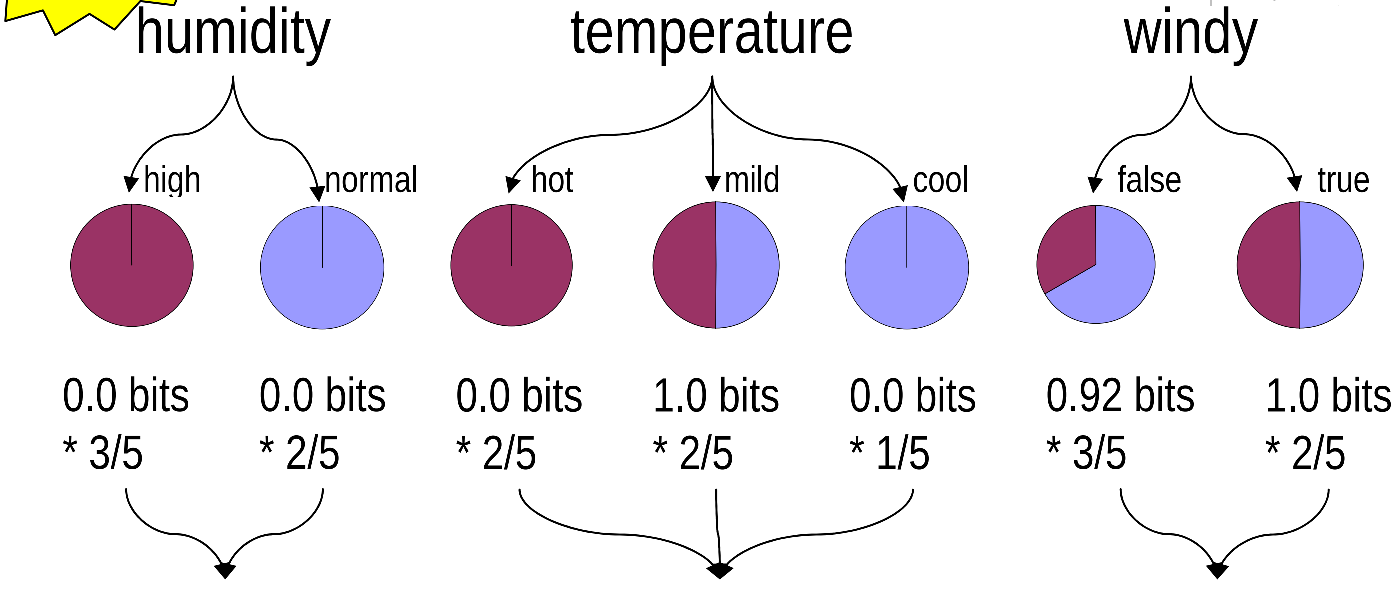
overcast

rainy

0.97 bits

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

maximal information gain



humidity

high — 0.0 bits * 3/5

normal — 0.0 bits * 2/5

+

= 0.0 bits

gain: 0.97 bits

temperature

hot — 0.0 bits * 2/5

mild — 1.0 bits * 2/5

cool — 0.0 bits * 1/5

+

= 0.40 bits

gain: 0.57 bits

windy

false — 0.92 bits * 3/5

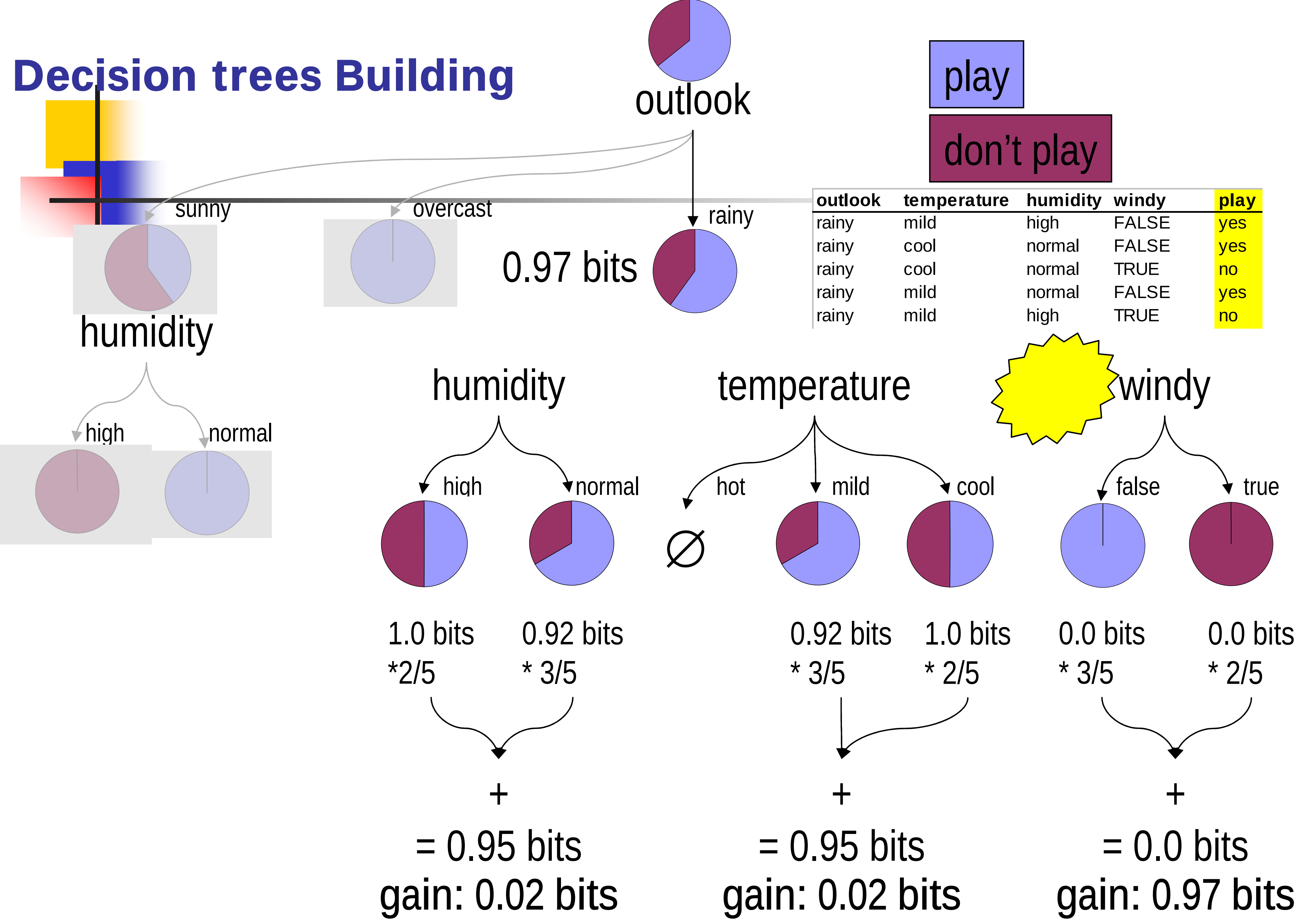
true — 1.0 bits * 2/5

+

= 0.95 bits

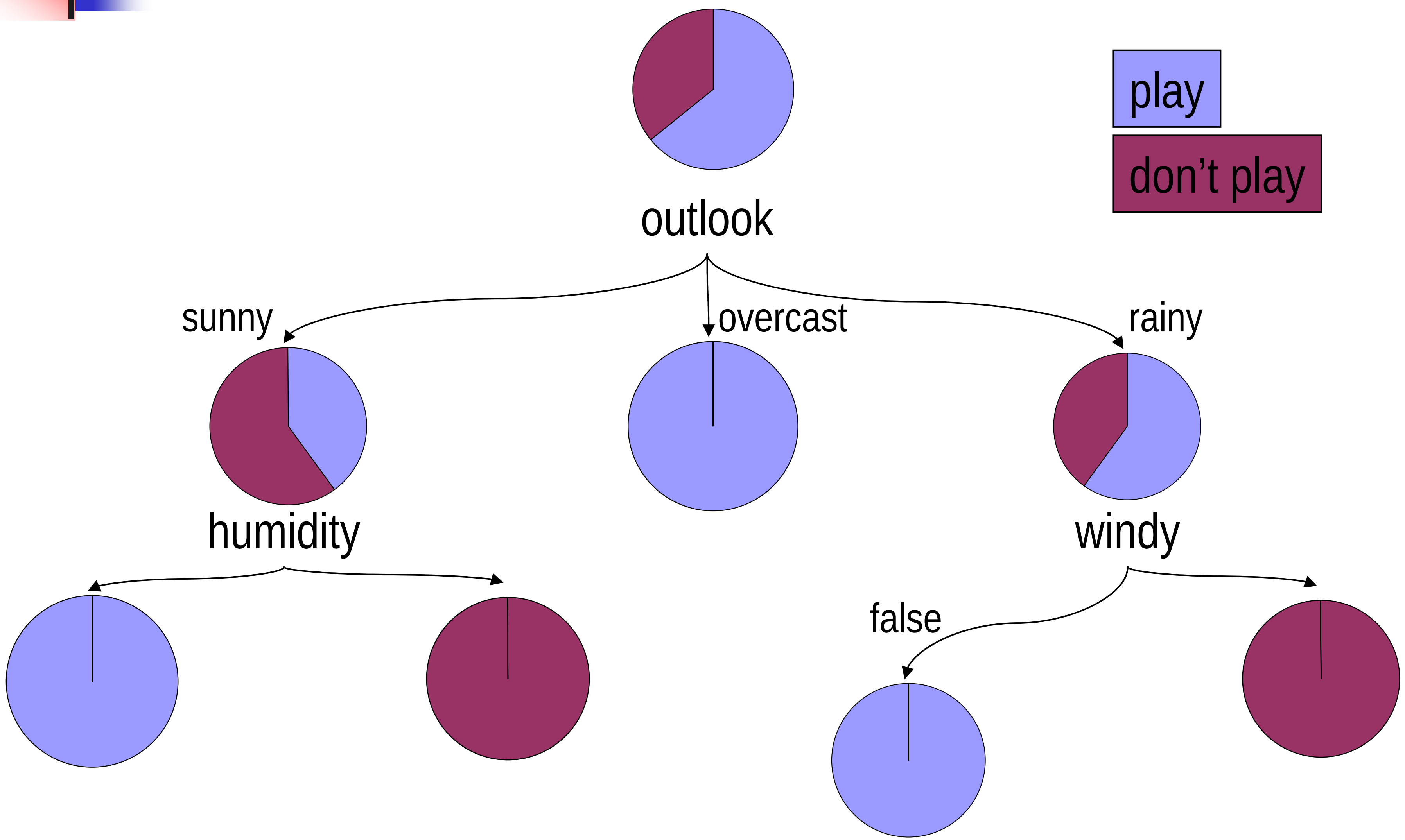
gain: 0.02 bits

Decision trees Building



outlook	temperature	humidity	windy	play
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no

-- Decision trees: Final tree

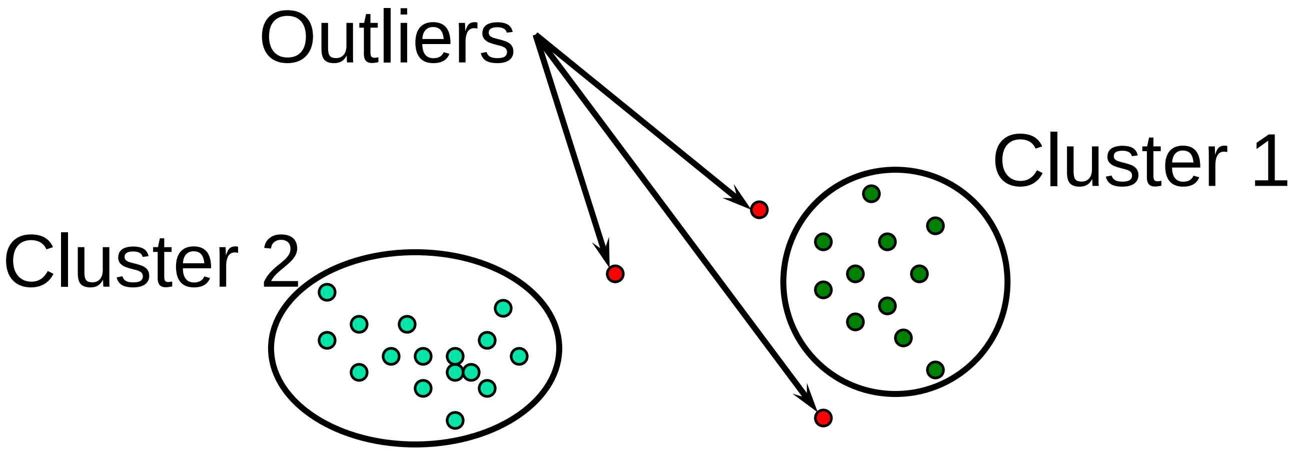


Decision tree: Basic Algorithm

- Initialize top node to all examples
- While impure leaves available
  - select next impure leave **L**
  - find splitting attribute **A** with maximal information gain
  - for each value of **A** add child to **L**

-- Clustering

- Group data into clusters
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
  - Unsupervised learning: no predefined classes



Approaches to Other Data Mining Problems

- Discovery of sequential patterns
- Discovery of Patterns in Time Series
- Discovery of Classification Rules
- Regression
- Neural Networks
- Genetic Algorithms
- Clustering and Segmentation

Applications of Data Mining

- Data mining can be applied to a large variety of decision-making contexts in business like
  - Marketing
  - Finance
  - Manufacturing
  - Health care

…