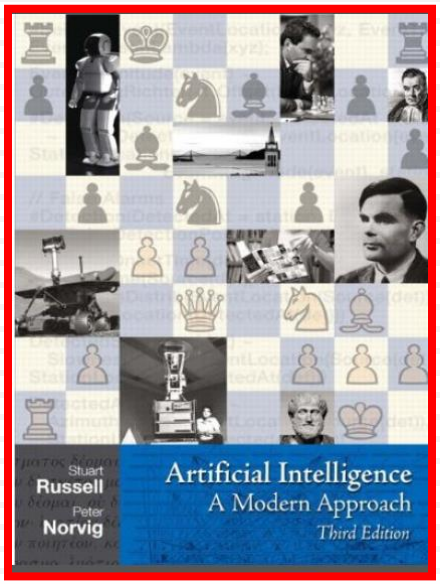
Princess Nora University
Faculty of Computer & Information Systems

جامعة الأميرة نورة بنت عبد الرحمن
Princess Nora Bint Abdul Rahman University

ARTIFICIAL INTELLIGENCE

(CS 370D)

جامعة الأميرة نورة بنت عبد الرحمن
Princess Nora Bint Abdul Rahman University



College of Computer and Information Sciences

(CHAPTER-18)
LEARNING FROM EXAMPLES
DECISION TREES

Outline

1- Introduction

2- know your data

3- Classification tree construction

schema

Dr. Abeer Mahmoud
(course coordinator)

Know your data?

Dr. Abeer Mahmoud
(course coordinator)

Types of Data Sets

Record

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vector
- Transaction data

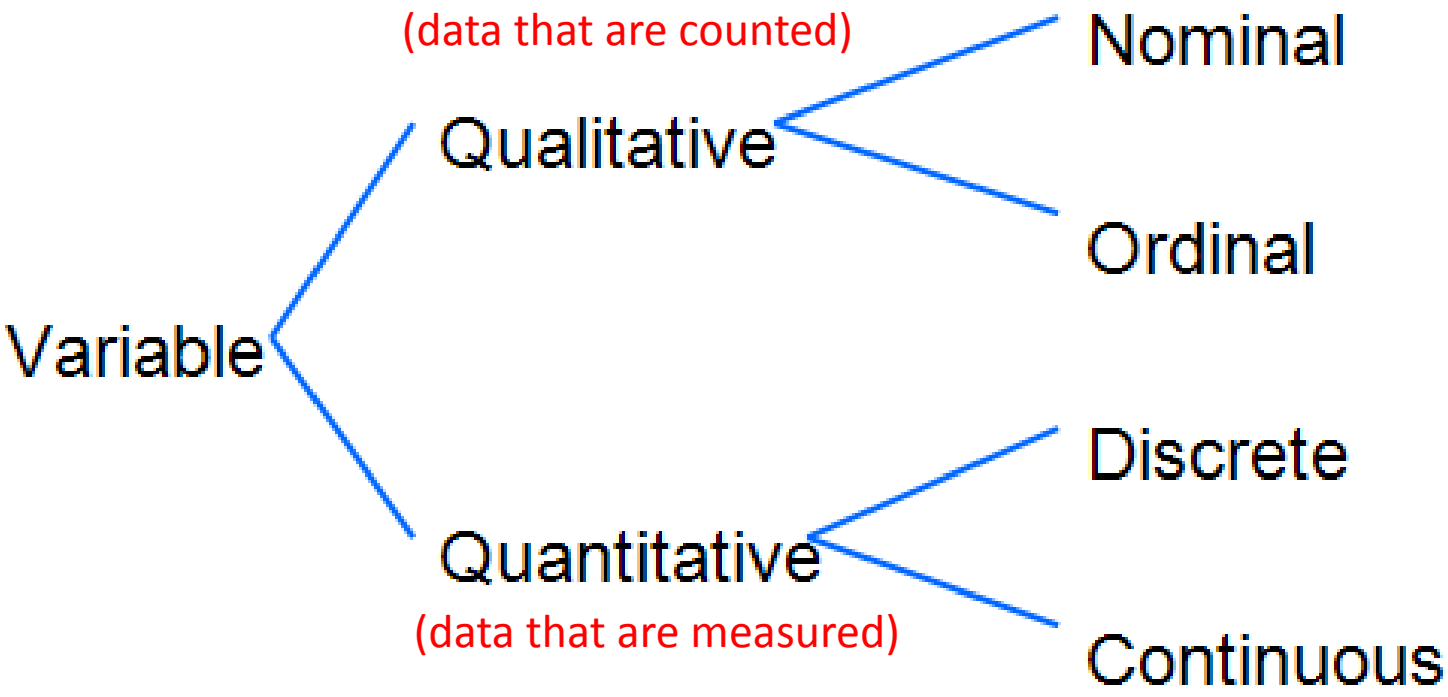
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Dr. Abeer Mahmoud
(course coordinator)

Types of Variables



Dr. Abeer Mahmoud
(course coordinator)

Types of Variables

"named", i.e. classified into one or more qualitative categories or description

(data that are counted)

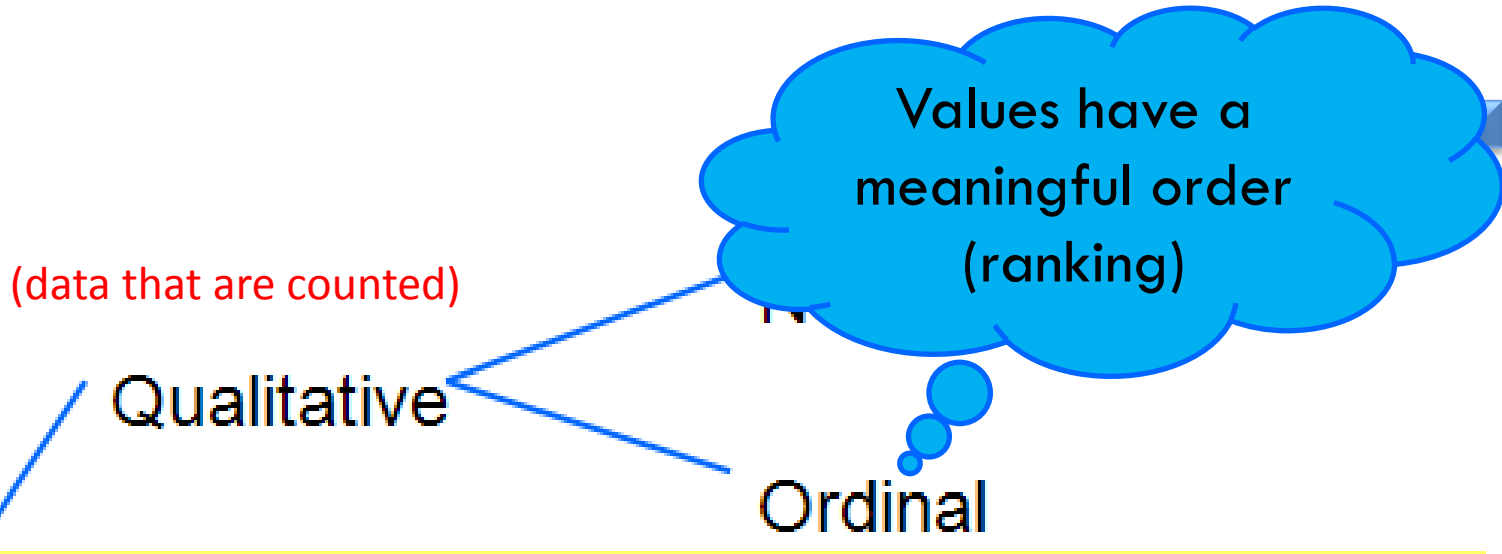
Nominal

In medicine, nominal variables are often used to describe the patient. Examples of nominal variables might include:

- Gender (male, female)
- Eye color (blue, brown, green, hazel)
- Surgical outcome (dead, alive)
- Blood type (A, B, AB, O)

Dr. Abeer Mahmoud
(course coordinator)

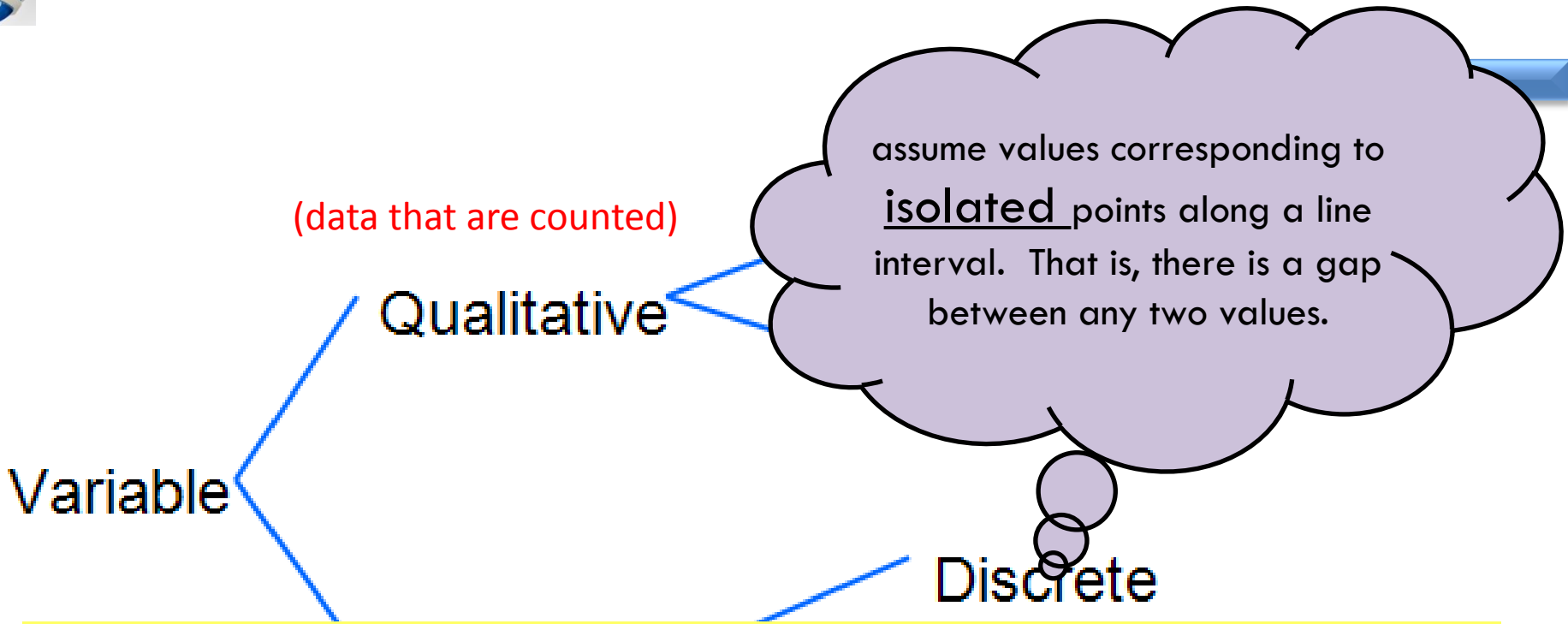
Types of Variables



In medicine, ordinal variables often describe the patient's characteristics, attitude, behavior, or status. Examples of ordinal variables might include:

- Stage of cancer (stage I, II, III, IV)
- Education level (elementary, secondary, college)
- Pain level (mild, moderate, severe)
- Satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)
- Agreement level (strongly disagree, disagree, neutral, agree, strongly agree)

Types of Variables

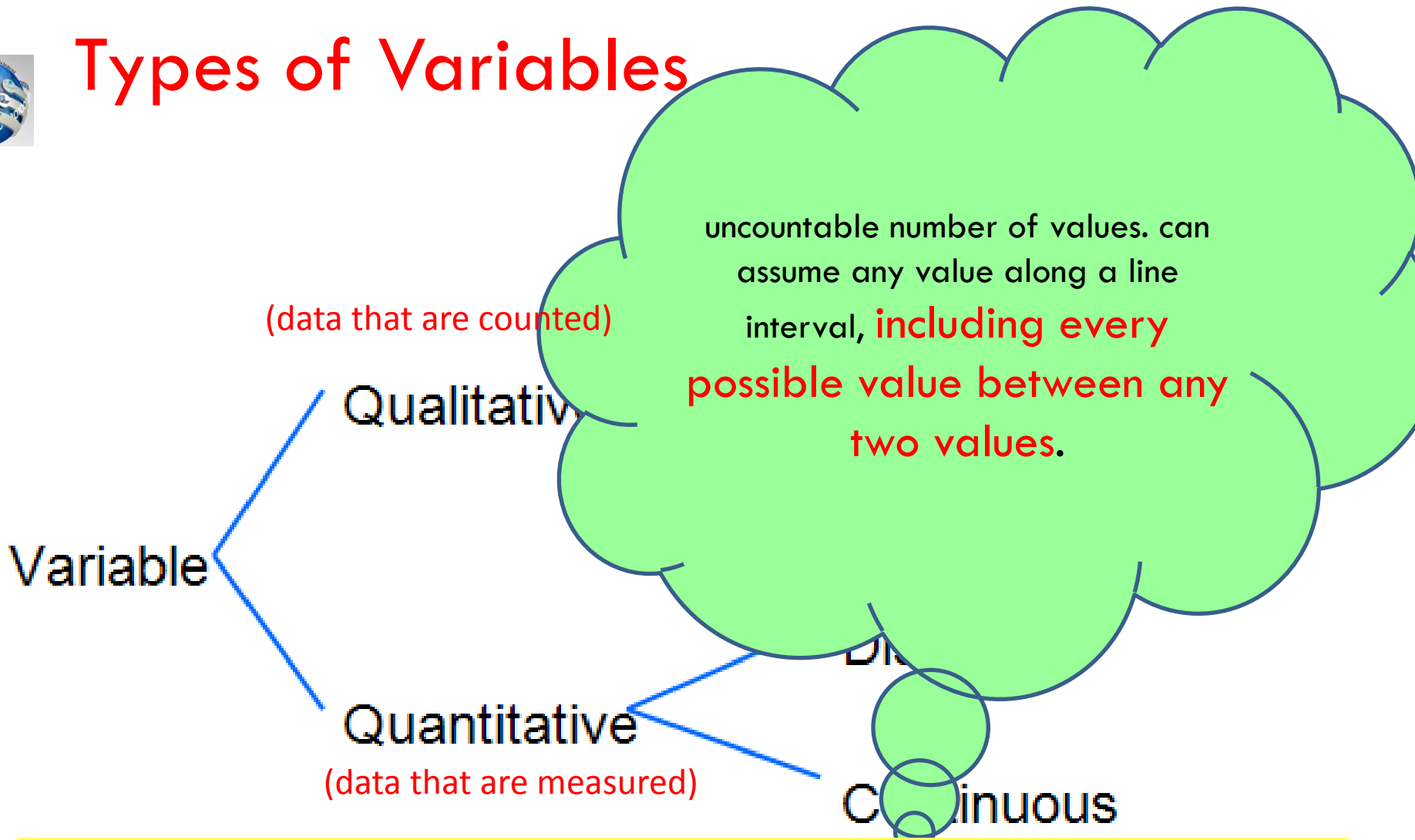


Discrete Variables that have constant, equal distances between values, but the zero point is arbitrary.

Examples of interval variables:

- Intelligence (IQ test score of 100, 110, 120, etc.)
- Pain level (1-10 scale)
- Body length in infant

Types of Variables



- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

Classification

Dr. Abeer Mahmoud
(course coordinator)

Classification: Definition

- Given a collection of records (*training set*)
- Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- **Goal**: previously unseen records should be assigned a class as accurately as possible.

- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Illustrating Classification Task

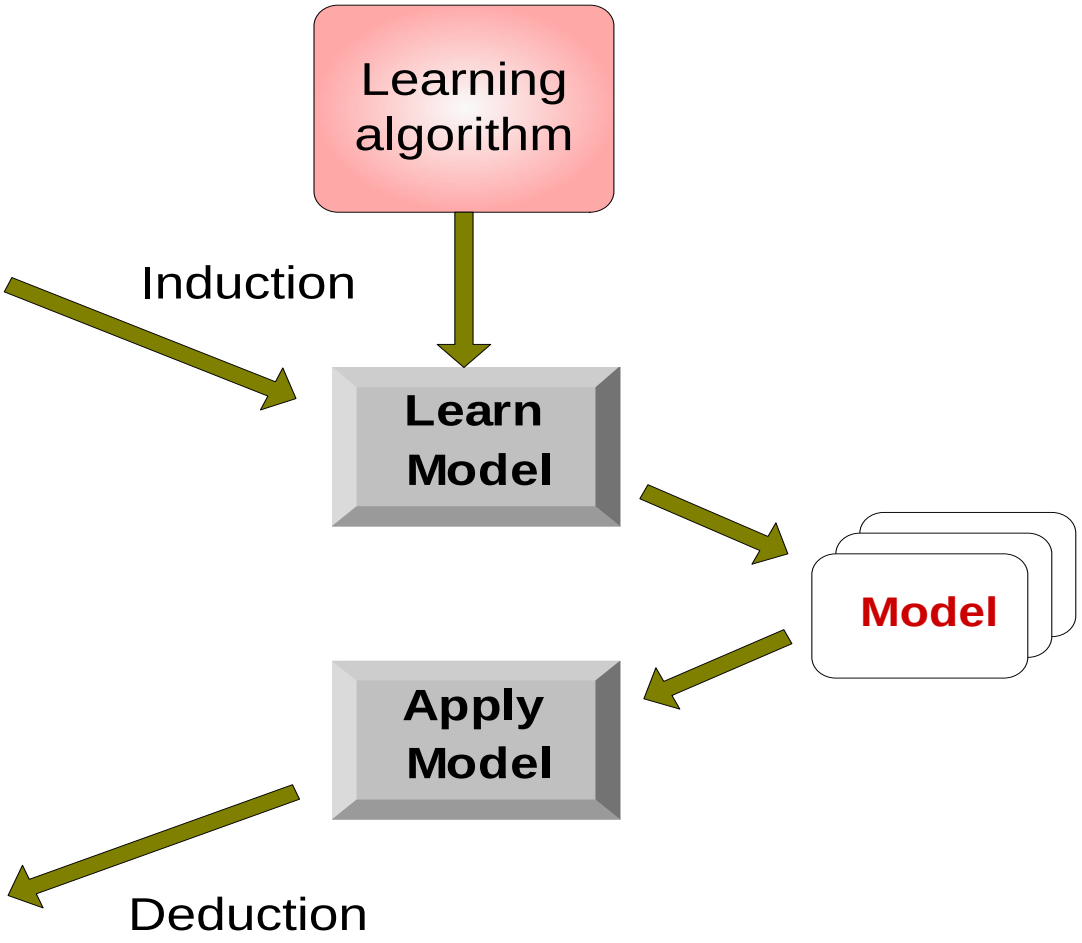
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

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

13



Dr. Abeer Mahmoud
(course coordinator)

Decision Tree Learning

Dr. Abeer Mahmoud
(course coordinator)

Definition

- Decision tree is a classifier in the form of a tree structure
  - Decision node: specifies a test on a single attribute
  - Leaf node: indicates the value of the target attribute
  - Arc/edge: split of one attribute
  - Path: a disjunction of test to make the final decision

- Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node.

Dr. Abeer Mahmoud
(course coordinator)

key requirements

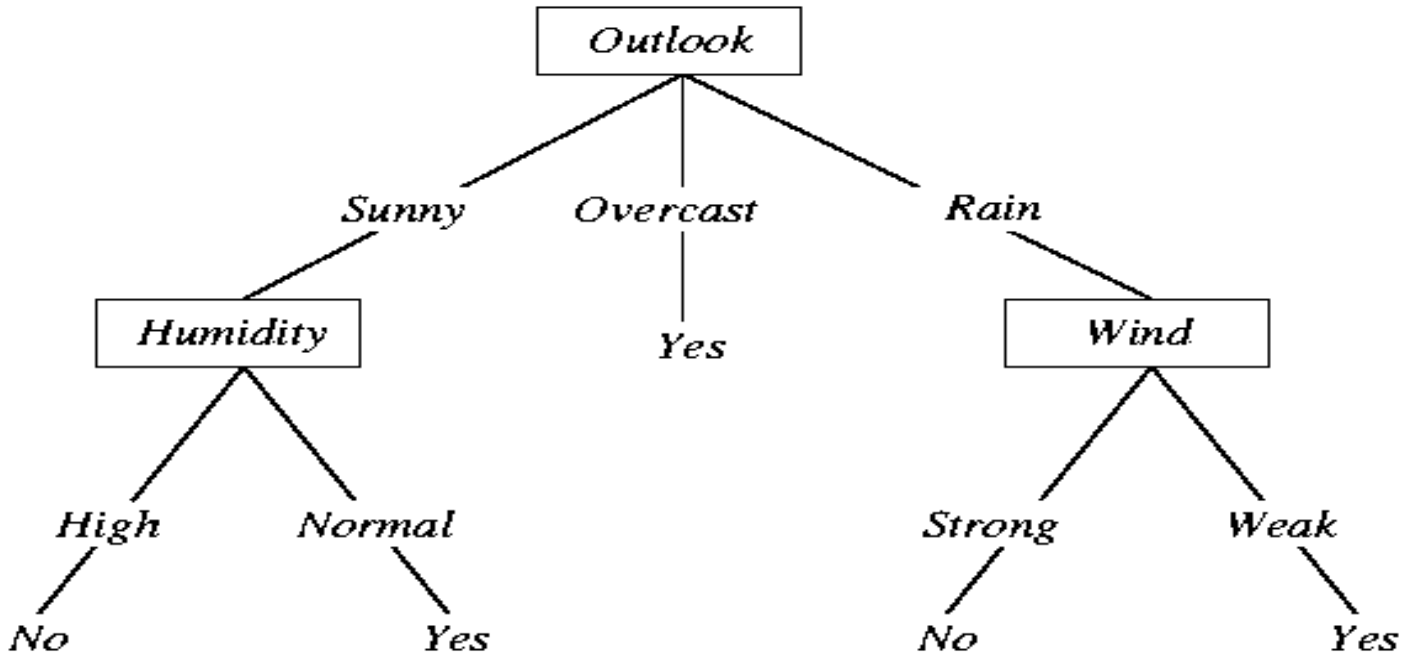
- **Attribute-value description**: object or case must be expressible in terms of a fixed collection of properties or attributes (e.g., hot, mild, cold).
- **Predefined classes (target values)**: the target function has **discrete output values** (single or multiclass)
- **Sufficient data**: enough training cases should be provided to learn the model.

Dr. Abeer Mahmoud
(course coordinator)

Training Examples

Day	Outlook	Temperature	Humidity	Wind	PlayTenr
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

Decision Tree for *PlayTennis*



Shall we play tennis today?

Decision Tree Construction

- Top-down tree construction schema:
- Examine training database and find best splitting predicate for the root node
- Partition training database
- Recourse on each child node

BuildTree(Node *t*, Training database *D*, Split Selection Method ***S***)
(1) Apply ***S*** to *D* to find splitting criterion
(2) **if** (*t* is not a leaf node)
(3) Create children nodes of *t*
(4) Partition *D* into children partitions
(5) Recurse on each partition
(6) **endif**

oud
(course coordinator)

Advantages of using DT

- Fast to implement
- Simple to implement because it perform classification without much computation
- Can convert result to a set of easily interpretable rules that can be used in knowledge system such as database, where rules are built from the label of the nodes and the labels of the arcs.
- Can handle continuous and categorical variables
- Can handle noisy data
- provide a clear indication of which fields are most important for prediction or classification

Dr. Abeer Mahmoud
(course coordinator)

Disadvantages of using DT

- "Univariate" splits/partitioning using only one attribute at a time so limits types of possible trees
- large decision trees may be hard to understand
- Perform poorly with many class and small data.

Dr. Abeer Mahmoud
(course coordinator)

Decision Tree Construction (cont..)

Two important algorithmic components:

1. Splitting (ID3, CART, C4.5, QUEST, CHAID, ……)
2. Missing Values

Dr. Abeer Mahmoud
(course coordinator)

1-Splitting

Dr. Abeer Mahmoud
(course coordinator)

2.1- Splitting

Depends on

Data base
Attributes types

Statistical calculation
technique or method

(ID3, CART, C4.5, QUEST, CHAID, ……)

Dr. Abeer Mahmoud
(course coordinator)

2.1- Split selection

- Selection of an attribute to test at each node - choosing the most useful attribute for classifying examples.

attributes or features

Class label or decision

Day	Outlook	Temperature	Humidity	Wind	PlayTenn
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

- How to specify the test condition at each node??

Dr. Abeer Mahmoud
(course coordinator)

How to specify the test condition at each node??

- Some possibilities are:
  - **Random**: Select any attribute at random
  - **Least-Values**: Choose the attribute with the smallest number of possible values (fewer branches)
  - **Most-Values**: Choose the attribute with the largest number of possible values (smaller subsets)
  - **Max-Gain**: Choose the attribute that has the largest expected information gain, i.e. select attribute that will result in the smallest expected size of the subtrees rooted at its children.

Dr. Abeer Mahmoud
(course coordinator)

How to specify the test condition at each node??

Depends on attribute types

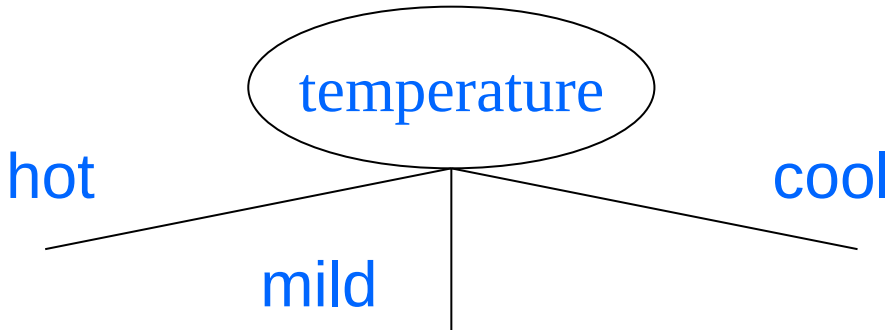
- 2.1.1 Nominal
- 2.1.2 Continuous

Dr. Abeer Mahmoud
(course coordinator)

2.1.1 Nominal or Categorical (Discrete)

- Nominal or categorical (Discrete): Domain is a finite set without any natural ordering (e.g., occupation, marital status (single, married, divorced…))

•Each non-leaf node is a test, its edge partitioning the attribute into subsets (easy for discrete attribute).



Dr. Abeer Mahmoud
(course coordinator)

2.1.2. Continuous or Numerical

- Continuous or Numerical: Domain is ordered and can be represented on the real line (e.g., age, income, temperatures degree)

- Convert to Binary
- Create a new boolean attribute A_c , looking for a threshold c,

$$A_c = \begin{cases} true & \text{if } A_c < c \\ false & \text{otherwise} \end{cases}$$

How to find best threshold??

- Discretization to form an ordinal categorical attribute where ranges can be found by equal intervals

Dr. Abeer Mahmoud
(course coordinator)

2.1- Splitting

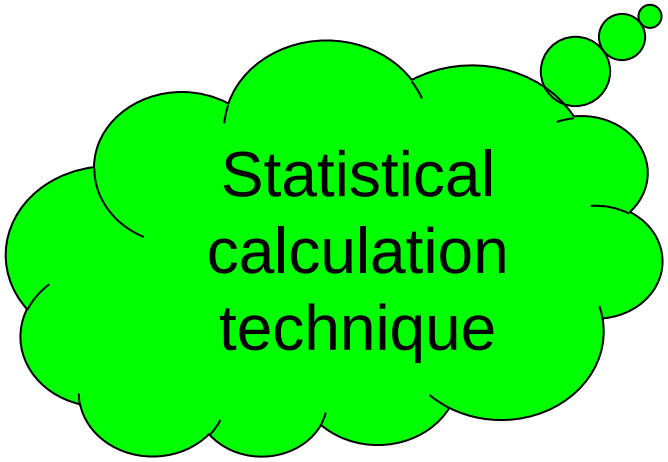
Depends on

Data base
Attributes types

Statistical calculation
technique or method

(ID3, CART, C4.5, QUEST, CHAID, ……)

Dr. Abeer Mahmoud
(course coordinator)



Entropy

Information gain

Gain ratio

Dr. Abeer Mahmoud
(course coordinator)

Entropy

- Entropy at a given node t:

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

(NOTE: $p(j \mid t)$ is the relative frequency of class **j** at node **t**).

- Measures homogeneity of a node.
  - Maximum **(log n_c)** when records are **equally distributed** among all classes implying least information
  - Minimum **(0.0)** when all records belong to **one class**, implying most information

Dr. Abeer Mahmoud
(course coordinator)

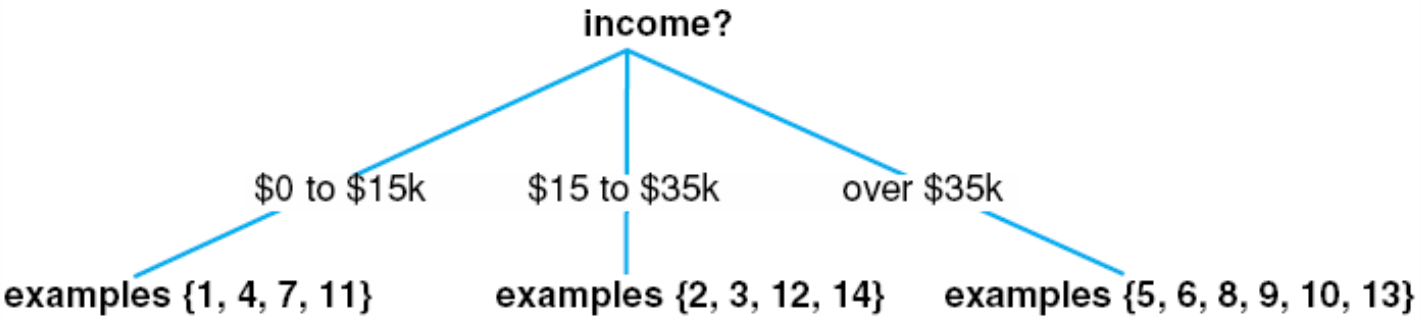
Example

No	Risk	Credit history	Debt	Collateral	Income
1	high	bad	high	none	0-15 $
2	high	unknown	high	none	15-35$
3	moderate	unknown	low	none	15-35$
4	high	unknown	low	none	0-15 $
5	low	unknown	low	none	Over 35$
6	low	unknown	low	adequate	Over 35$
7	high	bad	low	none	0-15 $
8	moderate	bad	low	adequate	Over 35$
9	low	good	low	none	Over 35$
10	low	good	high	adequate	Over 35$
11	high	good	high	none	0-15 $
12	moderate	good	high	none	15-35$
13	low	good	high	none	Over 35$
14	High	bad	high	none	15-35$

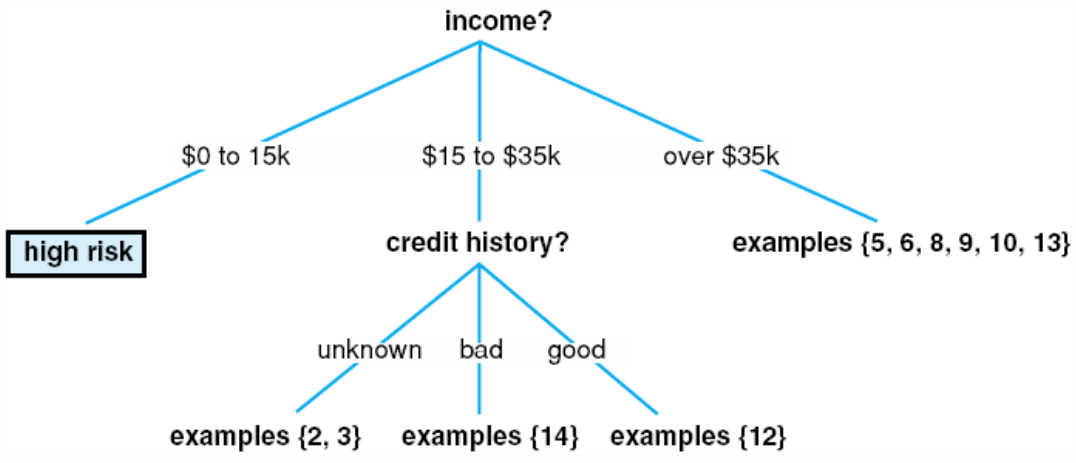
Example

starting with the population of loans

- suppose we first select the income property
- this separates the examples into three partitions



- all examples in leftmost partition have same conclusion – HIGH RISK
- other partitions can be further subdivided by selecting another property



ID3 & information theory

- the selection of which property to split on next is based on **information theory** the *information content* of a tree is defined by

$$\mathbf{I[tree] = \Sigma\ -prob(classification_i) * \log_2(\ prob(classification_i)\ )}$$

  - e.g., In credit risk data, there are 14 samples
    prob(high risk) = 6/14 prob(moderate risk) = 3/14
    prob(low risk) = 5/14

the information content of a tree that correctly classifies these examples:

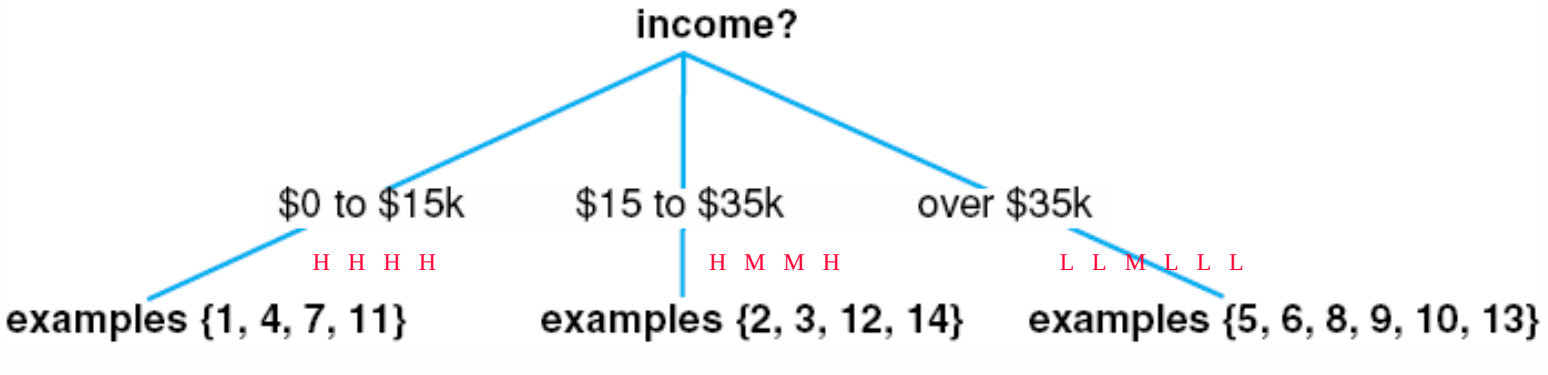
- $I[tree] = -6/14 * \log_2(6/14) + -3/14 * \log_2(3/14) + -5/14 * \log_2(5/14)$
  $= -6/14 * -1.222 + -3/14 * -2.222 + -5/14 * -1.485$
  $= 1.531$

Dr. Abeer Mahmoud
(course coordinator)

ID3 & more information theory- example

- after splitting on a property, consider the expected (or remaining) content of the subtrees

E[property] = Σ (# in subtree$_i$ / # of samples) * I [subtree$_i$]



E[income] = 4/14 * I[subtree$_1$] + 4/14 * I[subtree$_2$] + 6/14 * I[subtree$_3$]

= 4/14 * **(-4/4 log$_2$(4/4) + -0/4 log$_2$(0/4) + -0/4 log$_2$(0/4))** +

4/14 * (-2/4 log$_2$(2/4) + -2/4 log$_2$(2/4) + -0/4 log$_2$(0/4)) +

6/14 * (-0/6 log$_2$(0/6) + -1/6 log$_2$(1/6) + -5/6 log$_2$(5/6))

= 4/14 * (0.0+0.0+0.0) + 4/14 * (0.5+0.5+0.0) + 6/14 * (0.0+0.43+0.22)

= 0.0 + 0.29 + 0.28

= 0.57

Dr. Abeer Mahmoud
(course coordinator)

Credit risk example (cont.)

- what about the other property options?
  - E[debt]? E[history]? E[collateral]?
  - after further analysis
    E[income] = 0.57
    E[debt] = 1.47
    E[history] = 1.26
    E[collateral] = 1.33

the ID3 selection rules splits on the property that produces the **minimal E[property]**

- in this example, income will be the first property split
- then repeat the process on each subtree

Dr. Abeer Mahmoud
(course coordinator)

2.3- Missing Values

- What is the problem?
- During computation of the splitting predicate, we can selectively **ignore** records with missing values (note that this has some problems)
- But if a record r misses the value of the variable in the splitting attribute, r can not participate further in tree construction Algorithms for missing values address this problem.
- Simplest algorithm to solve this problem :

-If X is numerical (categorical), impute the overall mean
- if X is discrete attribute set the most common value

Dr. Abeer Mahmoud
(course coordinator)

Thank you

End of Chapter 18-part 2

Dr. Abeer Mahmoud
(course coordinator)