Data Mining

1.3 Input

Fall 2008

Instructor: Dr. Masoud Yaghini

Outline

- Instances
- Attributes
- References

Instances

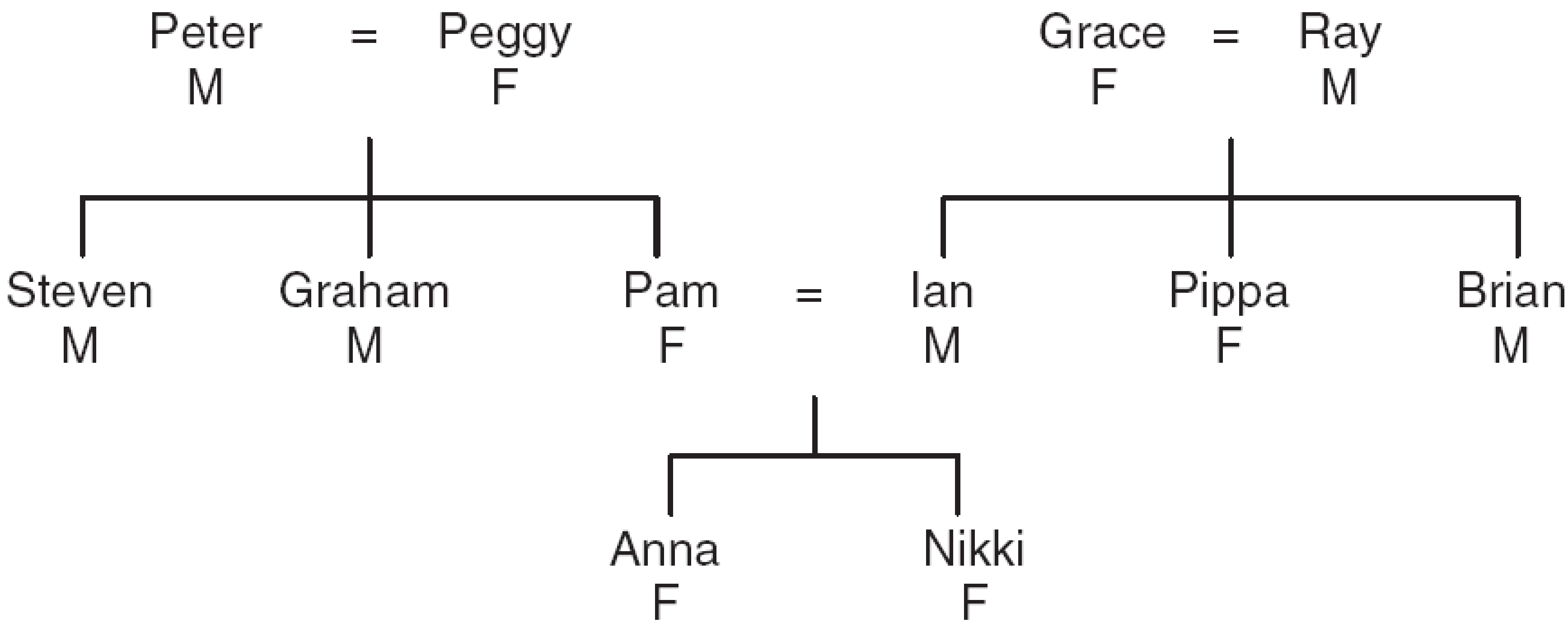
Instances

- **Instance**:
  - Individual, independent example of the concept to be learned.
  - Characterized by a predetermined set of attributes
  - Input to learning process: set of instances/dataset
- Each dataset is represented as a matrix of instances versus attributes
  - Represented as a single table or **flat file**
- Rather restricted form of input
  - No relationships between objects

Instances

- Problems often involve relationships between objects rather than separate, independent instances.
- Example:
  - a family tree is given, and we want to learn the concept *sister.*
  - This tree is the input to the learning process, along with a list of pairs of people and an indication of whether they are sisters or not.

An example: A family tree



Two ways of expressing the sister-of relation

first person	second person	sister of?
Peter	Peggy	no
Peter	Steven	no
...	......	
Steven	Peter	no
Steven	Graham	no
Steven	Pam	yes
Steven	Grace	no
...	......	
Ian	Pippa	yes
...	......	
Anna	Nikki	yes
...	.....	
Nikki	Anna	yes

first person	second person	sister of?
Steven	Pam	yes
Graham	Pam	yes
Ian	Pippa	yes
Brian	Pippa	yes
Anna	Nikki	yes
Nikki	Anna	yes
All the rest		no

- Neither table is of any use without the family tree itself.

Family tree represented as a table

Name	Gender	Parent1	Parent2
Peter	male	?	?
Peggy	female	?	?
Steven	male	Peter	Peggy
Graham	male	Peter	Peggy
Pam	female	Peter	Peggy
Ian	male	Grace	Ray
. . .			

- These tables do not contain independent sets of instances because values in the Name, Parent1, and Parent2 columns refer to rows of the family tree relation.

The sister-of relation represented in a table

First person				Second person				
Name	Gender	Parent1	Parent2	Name	Gender	Parent1	Parent2	Sister of?
Steven	male	Peter	Peggy	Pam	female	Peter	Peggy	yes
Graham	male	Peter	Peggy	Pam	female	Peter	Peggy	yes
Ian	male	Grace	Ray	Pippa	female	Grace	Ray	yes
Brian	male	Grace	Ray	Pippa	female	Grace	Ray	yes
Anna	female	Pam	Ian	Nikki	female	Pam	Ian	yes
Nikki	female	Pam	Ian	Anna	female	Pam	Ian	yes
			all the rest					no

- Each of instance is an individual, independent example of the concept that is to be learned.

A simple rule for the sister-of relation

- A simple rule for the sister-of relation is as follows:

```
If second person's gender = female
   and first person's parent1 = second person's parent1
   then sister-of = yes
```

Denormalization

- **Denormalization or flattening:**
  - Several relations are joined together to make one
  - to recast data into a set of independent instances
- Possible with any finite set of finite relations
- Problems:
  - Relationships without prespecified number of objects
    - Example: concept of nuclearfamily
  - Denormalization may produce false regularities that reflect structure of database
    - Example: “supplier” predicts “supplier address”

Summary

- The input to a data mining scheme is generally expressed as a table of independent instances of the concept to be learned.
- The instances are the rows of the tables the attributes are the columns.

Attributes

Attributes

- Each instance is described by a fixed predefined set of **features** or **attributes**

- Problem: Number of attributes may vary in different instances
  - Example: the instances were transportation vehicles
  - Possible solution: to make each possible feature an attribute and to use a special flag value to indicate that a particular attribute is not available for a particular case.

Attributes

- Another problem: existence of an attribute may depend of value of another one
  - Spouse’s name depends on the value of married or single attribute

Attributes Types

- Possible attribute types (“levels of measurement”):
  - **Nominal**
  - **ordinal**
  - **interval**
  - **ratio**

Nominal quantities

- Nominal attributes take on values in a prespecified, finite set of possibilities and are sometimes called **categorical**.
- Nominal quantities values are distinct symbols
  - Values themselves serve only as labels or names
- Example: attribute “outlook” from weather data
  - Values: “sunny”, “overcast”, and “rainy”
- No relation is implied among nominal values (no ordering or distance measure)

Nominal quantities

- Note: addition, subtraction, and comparing don't make sense
- Only equality tests can be performed
  - Example:

```
outlook: sunny    → no
         overcast → yes
         rainy    → yes
```

Ordinal quantities

- Ordinal quantities are ones that make it possible to rank order the categories.
- But: no distance between values defined
- Example: attribute “temperature” in weather data
  - Values: “hot” > “mild” > “cool”
- Note: it makes sense to compare two values, but addition and subtraction don’t make sense
- Example rule:
  - temperature < hot => play = yes
- Distinction between nominal and ordinal not always clear (e.g. attribute “outlook”)

Interval quantities

- Interval quantities are not only ordered but measured in fixed and equal units
- Example 1: attribute “temperature” expressed in degrees Fahrenheit
- Example 2: attribute “date” (year)
- Difference of two values makes sense
- Sum or multiplication doesn’t make sense
  - E.g. sum of the years 1939 and 1945 (3884)
  - Or, three times the year 1939 (5817)
- Zero point is not defined!
  - Example: the year 300

Ratio quantities

- Ratio quantities are ones for which the measurement method defines a zero point
- Example: attribute "distance"
  - Distance between an object and itself is zero
  - It does make sense to talk about three times the distance and even to multiply one distance by another to get an area.
- Ratio quantities are treated as real numbers
  - All mathematical operations are allowed

Attribute types used in practice

- Most data mining processes accommodate just two levels of measurement:
  - Nominal
  - Ordinal
- Nominal attributes are also called “categorical”, “enumerated”, or “discrete”
  - Enumerated is the standard term used in computer science to denote a categorical data type
  - But: “enumerated” and “discrete” imply order
  - Special case: “boolean” attribute
    - Example: true/false or yes / no

Attribute types used in practice

- Ordinal attributes are generally called **numeric**.

References

References

- Ian H. Witten and Eibe Frank, **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd Edition, Elsevier Inc., 2005. (Chapter 2)

The end