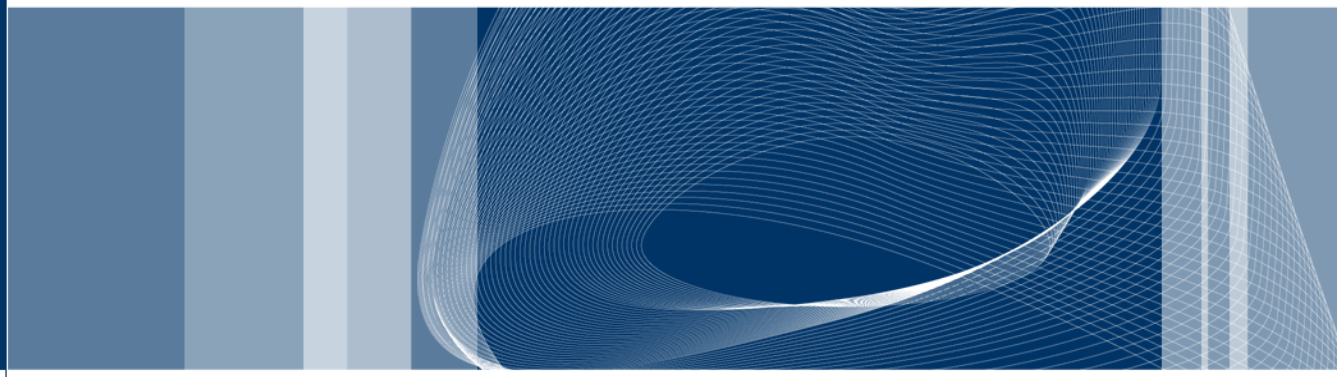
POLITECNICO DI MILANO



Data Representation

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

Lecture outline

- Components of the input
  - Concepts, instances, and attributes
- Attributes types
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Missing values and inaccurate values

Components of the input

- Concepts
  - Kinds of things that can be learned
- Instances
  - the individual, independent examples of a concept
- Attributes
  - measuring aspects of an instance

What's in an example?

- Instance: specific type of example
- Thing to be classified, associated, or clustered
- Individual, independent example of target concept
- Characterized by a predetermined set of attributes
- Input to learning scheme: set of instances/dataset
- Represented as a single relation/flat file
- Rather restricted form of input
- No relationships between objects
- Most common form in practical data mining

What's in an attribute?

- Each instance is described by a fixed predefined set of features, its "attributes"
- But the number of attributes may vary in practice
- Possible solution: "irrelevant value" flag
- Related problem: existence of an attribute may depend of value of another one
- Possible attribute types ("levels of measurement"):
- Nominal, ordinal, interval and ratio

The contact lenses data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

The weather problem

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

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
  - No distance measure
- Only equality tests can be performed

Ordinal quantities

- Impose order on values
- No distance between values defined
- Example:
  - The attribute “temperature” in weather data
  - Values: “hot” > “mild” > “cool”
- Addition and subtraction don’t make sense
- Example rule:
  temperature < hot Þ 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
- Examples
  - Attribute “temperature” expressed in degrees
  - Attribute “year”
- Difference of two values makes sense
- Sum or product doesn’t make sense
- Zero point is not defined

Ratio quantities

- Ratio quantities are ones for which the measurement scheme defines a zero point
- Example
  - Attribute “distance”
  - Distance between an object and itself is zero
- Ratio quantities are treated as real numbers
- All mathematical operations are allowed
- Is there an “inherently” defined zero point?
  - It depends on scientific knowledge
  - E.g. Fahrenheit knew no lower limit to temperature

Attribute types used in practice

- Most schemes accommodate just two levels of measurement: nominal and ordinal
- Nominal attributes are also called “categorical”, ”enumerated”, or “discrete”
- But: “enumerated” and “discrete” imply order
- Special case: dichotomy (“boolean” attribute)
- Ordinal attributes are called “numeric”, or “continuous”
- But: “continuous” implies mathematical continuity

Why specify attribute types?

- Why Machine Learning algorithms need to know about attribute type?
- To be able to make right comparisons and learn correct concepts
- Example
  - Outlook > "sunny" does not make sense, while
  - Temperature > "cool" or
  - Humidity > 70 does
- Additional uses of attribute type
  - Check for valid values
  - Deal with missing, etc.

Nominal vs. ordinal

- Attribute "age" nominal
  - If age = young and astigmatic = no
    and tear production rate = normal
    then recommendation = soft
  - If age = pre-presbyopic and astigmatic = no
    and tear production rate = normal
    then recommendation = soft
- Attribute "age" ordinal
  (e.g. "young" < "pre-presbyopic" < "presbyopic")
  - If age≤pre-presbyopic and astigmatic = no
    and tear production rate = normal
    then recommendation = soft

Missing values

- Frequently indicated by out-of-range entries
- Types: unknown, unrecorded, irrelevant
- Reasons:
  - Malfunctioning equipment
  - Changes in experimental design
  - Collation of different datasets
  - Measurement not possible
- Missing value may have significance in itself
  - E.g. missing test in a medical examination
- Most schemes assume that is not the case
  - "missing" may need to be coded as additional value
- Does absence of value have some significance?
  - If it does, "missing" is a separate value
  - If it does not, "missing" must be treated in a special way

Inaccurate values

- What is the reason?
  - data has not been collected for mining it
- What is the result?
  - errors and omissions that don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes, thus values need to be checked for consistency
- Typographical and measurement errors in numeric attributes, thus outliers need to be identified
- Errors may be deliberate (e.g. wrong zip codes)
- Other problems: duplicates, stale data

Getting to know the data

- Simple visualization tools are very useful
- Nominal attributes: histograms
- Numeric attributes: graphs
- 2-D and 3-D plots show dependencies
- Need to consult domain experts
- When there is too much data to inspect, take a sample!

Example: the ARFF format


```
%
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
```

ARFF: Additional attribute types

- ARFF supports string attributes:

```
@attribute description string
```

  - Similar to nominal attributes but list of values is not pre-specified

- ARFF also supports date attributes:

```
@attribute today date
```

  - Uses the ISO-8601 combined date and time format yyyy-MM-dd-THH:mm:ss

ARFF: Attribute Types & Interpretation

- Interpretation of attribute types in ARFF depends on the mining scheme
- Numeric attributes are interpreted as
  - Ordinal scales if less-than and greater-than are used
  - Ratio scales if distance calculations are performed (normalization/standardization may be required)
- Instance-based schemes define distance between nominal values (0 if values are equal, 1 otherwise)
- Integers in some given data file: nominal, ordinal, or ratio scale?

Summary

- Data are defined in terms of concepts, instances, & attributes
- Different types of attributes: nominal, ordinal, interval, etc.
- Attribute values can be missing or inaccurate
- Missing values can be unknown, unrecorded, irrelevant, etc.