CSCI568

Lecture 4: Data Fundamentals
Sep. 2, 2009

Data Mining Starts With Data! (duh)

- type
- quality
- preprocessing
- existing/obvious relationships

Types

Categorical
(qualitative)

Numeric
(quantitative)

Nominal	Ordinal	Interval	Ratio
zip codes	hardness	dates	temp in K
student ID	grades	temperature (in C/F)	money
eye color	street numbers		age
gender			lengths

(DM 26)

Data Transformations

(changes that don't affect the *meaning* of an attribute)

Categorical
(qualitative)

Numeric
(quantitative)

Nominal

preserve 1:1 mapping

Ordinal

preserve order

Interval

consistent scaling

Ratio

preserve ratio, but units can change

Bottom line: *statistical* operations should yield same results whether or not data is transformed.

(DM 27)

of possible values

- finite ... discrete
- infinite.. continuous (limited precision)

A "Special" Type: Asymmetric Attributes

- We only care about presence (or non-presence)
- Influences the meaning of "similarity"

Data Sets (collections of objects)

- Dimensionality (# of attributes)
- Sparsity (how often a value exists)
- Resolution
  - too fine? pattern buried in noise
  - too coarse? pattern not evident

Data Sets Have Types Too

DM 29 - 36

Record-Based

- General (most common)
- Transaction / Market Basket Data
    - One attribute includes a set of values
- Data Matrix
    - sparse data matrices
        - document-term matrices

Graph-Based

- Relationships among objects
  - data objects are mapped to nodes
  - relationships are represented as links
- Object attributes might be graphs

Ordered Data

- Temporal data
  - "candy sales peak before Halloween
  - purchase history, predictions
  - subject to temporal autocorrelation
- Sequence data (position, no timestamps)
- Spatial data (eg, weather data by location)
  - subject to spatial autocorrelation

Unstructured Data

- non-record data can become record data
- but you might not capture certain aspects of the data

Data Quality Issues

- collected for other purposes
- collected without specified application
- errors
- *often cannot be addressed at the source*

DM 36 - 44

Data Mining Includes...

- detection and correction of problems
- use of algorithms that tolerate poor quality

Data Is Rarely Perfect (damn!)

- human error
- limitation of measuring tools
- flaws in collection process
- missing values
- missing records
- duplicate records

Measurement Error

- recorded value differs from true value
- measure the measurement error (aka “error”) of continuous attributes

Collection Error

- eg, omitting data objects or attributes
- inappropriately including a data object
- collection process problems
- keyboard errors

Noise & Artifacts

- NOISE = random component of measurement error
    - distortion or added spurious objects
    - most often with temporal / spatial datasets
    - elimination is often difficult
        - so we use algorithms that are robust against noise
- ARTIFACTS = deterministic errors
    - think: “a repeated streak on a set of photos”

Precision, Bias, Accuracy (simple, but important!)

- Precision: closeness of repeated measurements
  - std. dev. of a set of values
- Bias: systematic imprecision
  - diff between mean of values and real value
- Accuracy: closeness to true value

DM 39

Outliers

- *data objects* that have characteristics that are very different from most of the other objects
- *value*s of an attribute that are very different from typical values of that attribute
- outliers are not noise per se
    - are often legitimate data objects/values

DM 40

Missing Values (and what to do about it)

- Eliminate data object or attribute
- Estimate missing values
- Ignore missing values during analysis

DM 40, 41

Inconsistent Values

- think: "incorrect city for a zip code"
- often easy to detect
  - eg, human height shouldn't be negative
- can use other valid data to resolve inconsistencies

DM 41, 42

Duplicate Data

- duplicates or "almost-duplicates"
- eg, John Smith vs John Q. Smith
    - Same object but slightly different attributes
- Values should be resolved/merged
    - aka "deduplication"

DM 42, 43

What Defines Quality Data?

Data is of "high quality" if it is suitable for its intended use.

Timeliness
Relevance
Documentation / Knowledge