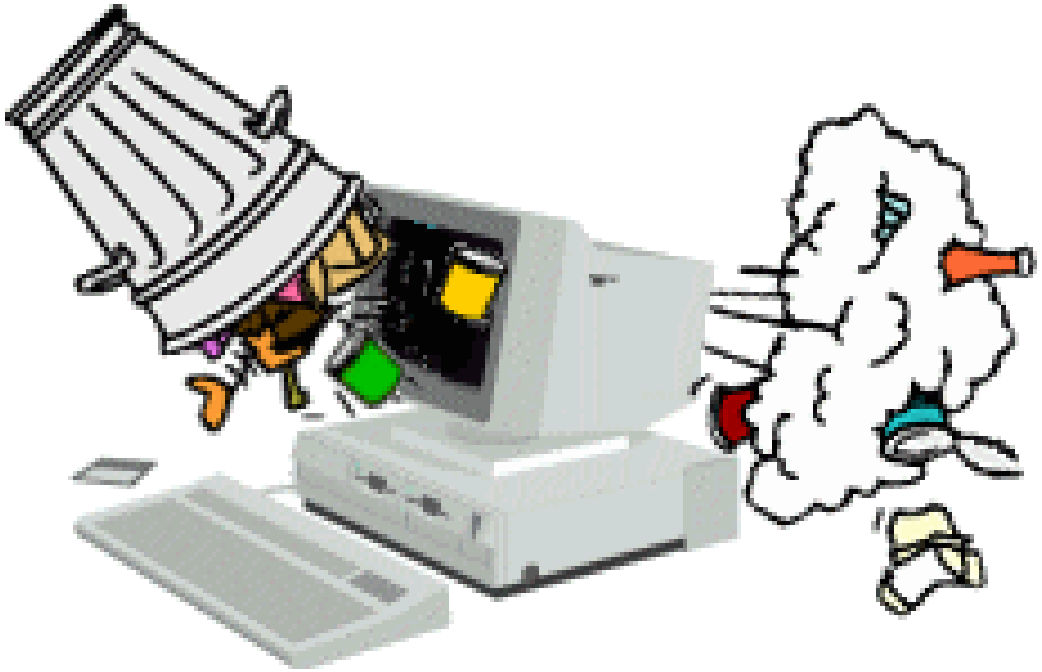
DATA PREPROCESSING

Garbage In, Garbage Out



Quality of data mining output depends on quality of input data

Data Quality Issues

- Data mining is often applied to opportunistic samples (data that have already been collected)
  - Preventing data quality issues (noise, missing values, duplicate data, etc) through careful design of experiment is not an option

- Data preprocessing is needed to help alleviate many of the data quality issues

Data – Things to consider

- Type of data: determines which tools to analyze the data
- Quality of data:
  - Tolerate some levels of imperfection
  - Improve quality of data improves the quality of the results
- Preprocessing: modify the data to better fit data mining tools:
  - Change length into short, medium, long
  - Reduce number of attributes

Data

- Collection of objects or records
- Each object described by a number of attributes
- Attribute: property of the object, whose value may vary from object to object or from time to time

Patient ID	Gender	DOB	Systolic	Diastolic	Heart Rate	Smoker	…
…	…				…		
1029345	Male	1/24/1957	151	92	62	Yes	
1029346	Male	5/3/1983	124	80	66	No	
1029347	Female	9/20/1991	110	74	54	Yes	
…	…				…		

Types of Attributes

- Nominal: Differentiates between values based on names
  - *Gender, eye color, patient ID*
- Ordinal: Allows a rank order for sorting data but does not describe the degree of difference
  - *{low, medium, high}, grades {A, B, C, D, F},*
- Interval: Describes the degree of difference between values
  - *Dates, temperatures in C and F*
- Ratio: Both degree or difference and ratio are meaningful
  - *Temperatures in K, lengths, age, mass, …*

Binary Attributes

- A nominal attribute with only two categories: 0 and 1 (or True/False)
- **Symmetric**: if both states of the variable are equally valuable

  attendance: 0 denotes No, 1 denotes Yes

- **Asymmetric**: if the states are not equally important

  Medical Test: 0 denotes negative, 1 denotes positive

Attribute Properties

- Distinctness: = and ≠
- Order: <, ≤, ≥, and >
- Addition: + and -
- Multiplication: * and /

	Type	Description	Examples	Operations
Categorical Or Qualitative	Nominal	Provide enough information to distinguish by name. =, ≠	Zip code, employee ID numbers, eye color, gender	Mode, entropy, contingency
	Ordinal	Provide enough information to sort. <, >	Hardness of minerals {good, better, best}, street numbers	Median, percentiles, rank correlation
Numeric Or Quantitative	Interval	Differences between values are meaningful. +, -	Calendar dates, temps in Celsius and Fahrenheit	Mean, standard deviation, Pearson's correlations
	Ratio	Differences and ratios are meaningful *, /	Temps in Kelvin, monetary quantities, counts, age, mass	Geometric mean, harmonic mean, percent variation

Transformations

	Type	Transformation	Comments
Categorical Or Qualitative	Nominal	Any one to one mapping	If all employee numbers are reassigned, it will not make a difference
	Ordinal	Any order preserving function	{0.5, 1, 10} => {1, 2, 3}
Numeric Or Quantitative	Interval	new = a*old+b	Celsius to/from Fahrenheit
	Ratio	New = a*old	Length can be measured in meters or feet

Discrete and Continuous Attributes

- Discrete Attributes:
  - Finite or countably infinite set of values
  - Categorical (zipcode, empIDs) or numeric (counts)
  - Often represented as integers
  - Special Case: binary attributes (yes/no, true/false, 0/1)
- Continuous Attributes:
  - Real numbers
  - Examples: temperatures, height, weight, …
  - Practically, can be measured with limited precision

Asymmetric Attributes

- Only presence of attribute is considered important
- Can be binary, discrete or continuous
- Examples:
  - Words in document
  - Courses taken by students
  - Items purchased by customers

	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

Exploring the data

- Number of attributes
- Type of attributes
- Data range, tendency, standard deviation
- Correlations between attributes
- Visualization:
  - Box plots
  - Scatter plots
  - Histograms

Similarity and Dissimilarity

- Similarity: a numerical measure of the degree to which the objects are alike
- Dissimilarity: a numerical measure of the degree to which the objects are different
- Can convert a dissimilarity measure to similarity and vice versa
- The proximity of objects with several attributes is the combination of proximities of individual attributes
- Distance as measure for similarity/dissimilarity:
  - Euclidean, SMC, Cosine, Jaccard

Proximity for simple attributes

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 \text{ if } x = y \\ 1 \text{ if } x \neq y \end{cases}$	$s = \begin{cases} 1 \text{ if } x = y \\ 0 \text{ if } x \neq y \end{cases}$
Ordinal (map values to 0..n-1)	$d = \lvert x - y \rvert / (n-1)$	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$ $\quad s = e^{-d}$ $\quad s = \frac{1}{1+d}$ $\quad s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Similarity for Binary Data

- Objects have only binary attributes

$$x = (1,0,1,1,1,0,\dots,1) \qquad y = (1,0,0,0,1,0,\dots,0)$$

- Compute the similarity using:
  - f_{00} = number of attributes where x is 0 and y is 0
  - f_{01} = number of attributes where x is 0 and y is 1
  - f_{10} = number of attributes where x is 1 and y is 0
  - f_{11} = number of attributes where x is 1 and y is 1

Counts the presences and absences equally

- Simple Matching Coefficient:

$$SMC = \frac{\text{Number of matching attributes}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

Similarity for Binary Data – Jaccard Coeff

- Objects have only binary attributes

$$x = (1,0,1,1,1,0,\ldots,1) \qquad y = (1,0,0,0,1,0,\ldots,0)$$

- Jaccard Coefficient:

$$J = \frac{\text{Number of matching presences}}{\text{number of attributes not in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Does not include matching absences (i.e., f_{00})

Example: Jaccard Coefficient

Cases	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	1	0	0	0	1	1	1	1
2	0	1	0	0	1	0	1	0	0	1

Assume all the activities are *asymmetric* binary variables. What is the Jaccard coefficient between case 1 and case 2?

A contingency table

		case1	
		1	0
case2	1	3 (f11)	1 (f01)
	0	3 (f10)	3 (f00)

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{3}{1+3+3} = \frac{3}{7}$$

Cosine Similarity

- Ignores 0-0 matches but handles non-binary data (document matrices)

$$x = (3,0,2,5,1,0, \dots, 7) \qquad y = (1,5,1,0,0,0, \dots, 2)$$

- Cosine Similarity: $\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$

where $\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2}$ $\qquad x \cdot y = \sum_{k=1}^{n} x_k y_k$

- Measures the angle between vectors x and y
- If cos(x, y) = 1, the angle is 0 and x and y are the same except the magnitude
- If cos(x, y) = 0, the angle is 90, x and y do not share any terms

Issues

- Standardize when attributes have different scales
- Combine similarities when attribute types are different
  - For the k-th attribute, compute similarity s_k in the range [0, 1]
  - Define δ_k as:
    - 0 if the attribute is asymmetric and a 0-0 match for both objects, or one of the object has a missing value.
    - 1 otherwise
  - Define the similarity as $similarity(p,q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$
- Some attributes are more important than others:
  - use weights (between 0 and 1, summing to 1)

$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$ $\quad$ $distance(p,q) = (\sum_{k=1}^{n} w_k |p_k - q_k|^r)^{1/r}$

Why to Preprocess the data?

- Transform data into specific form
- Handle data quality issues

Data cleaning
Data transformation
Data reduction
Data integration

Handling non record data

- Most data mining algorithm are designed for record data
- Ignore the data row:
  - many attributes are missing from the row
  - poor performance if number is too high
- Non-record data can be transformed into record data

  Example:

  Input: Chemical structures data

  Expert knowledge: common substructures

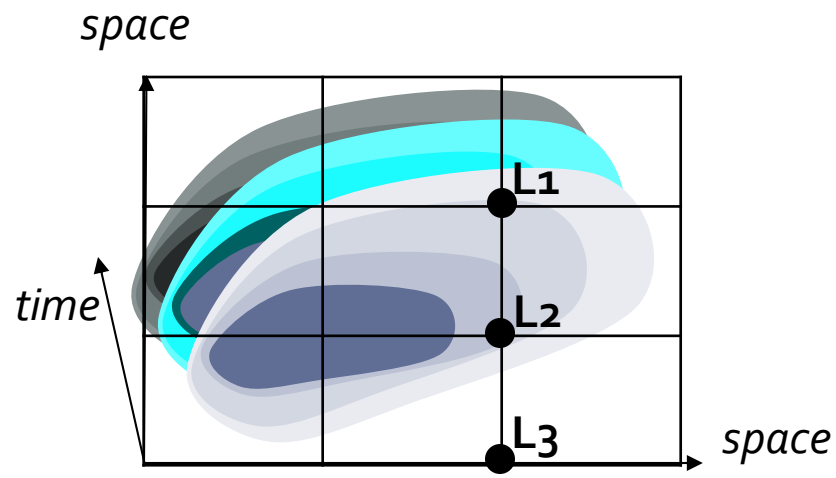
  Transformation: For each compound, add a record with common substructures as binary attributes

	Substruc 1	Substruc 2	…	Substruc 11
CH0001	1	1		0
CH0002	1	1		1

Handling non record data

May lose some information during transformation

Example: spacio-temporal data: time series for each data on a grid



	T1	T2	T3	…
L1	76	75	75	
L2	84	86	90	
L3	84	86	90	
…				

Are T2 and T3 independent? L2 and L3?

The transformation does not capture time/location relationships

Data quality Issues

- Data collected for purposes other than data mining
- May not be able to fix quality issues at the source
- To improve quality of data mining results:
  - Detect and correct the data quality issues
  - Use algorithms that tolerate poor data quality

Data quality Issues

- Caused by human errors, limitations of measuring devices, flaws in the data collection process, duplicates, missing data, inconsistent data
- Measurement error: the value recorded is different from true value
- Data collection error: an attribute/object is omitted or an object is inappropriately included.

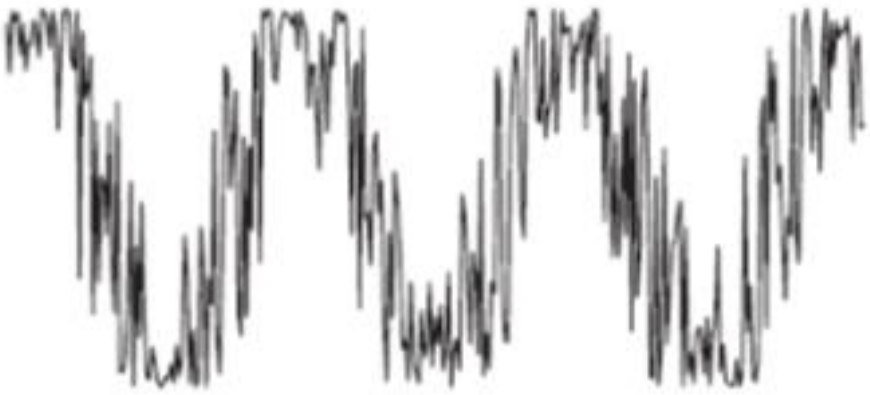
Data Quality Issues

- Noise
- Outliers
- Missing values
- Duplicate data

Noise

- Random component of a measurement error
- Distortion of the value or addition of spurious objects

(a) Time series.



(b) Time series with noise.

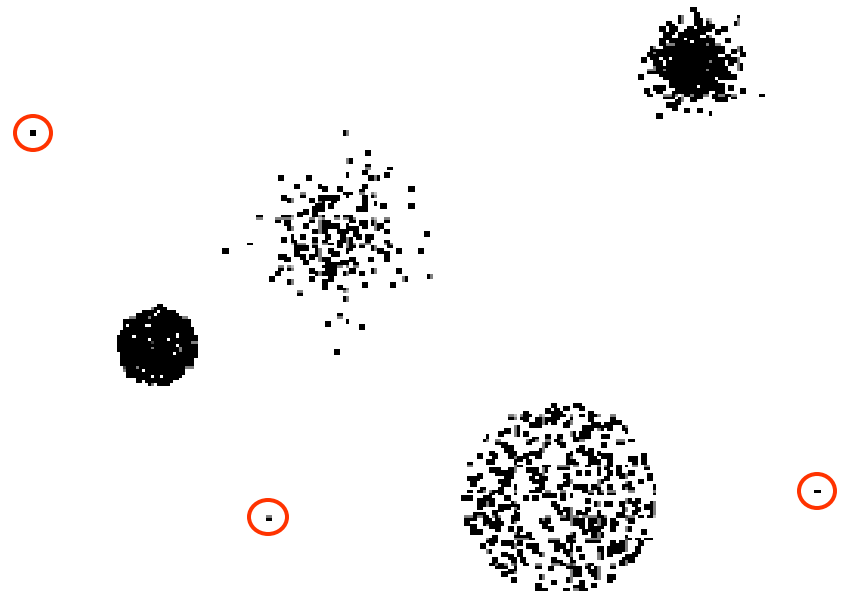
Noise

- How do data get corrupted?
  - Error in measurement due to faulty or low-resolution sensors
  - Error in data recording
  - External (environmental) factors that affect the measurement process
- Are noisy data useful or should they be discarded?
- Are there any reasons to intentionally add noise to the data?

Outliers

- Data objects with characteristics different than most other data objects
- Attributes values unusual with respect typical values

- Legitimate objects or values
- Applications: Intrusion detection, fraud detection



Missing Values

- Not collected
- Not applicable

ID	Age	Income	Prior Participation	Year	
12345	23		Yes	2012	…
34354	34		No		…
23545	58	75,000	Yes	2011	…

Missing Values

- Not collected
- Not applicable

ID	Age	Income	Prior Participation	Year	
12345	23	*unknown*	Yes	2012	…
34354	34	*unknown*	No	*N/A*	…
23545	58	75,000	Yes	2011	…

Missing Values

- Eliminate data objects or attributes: (Complete-case analysis)
  - Too many eliminated objects make the analysis unreliable
  - Eliminated attributes may be critical to the analysis

- Estimate missing values
  - Regression
  - Average of nearest neighbors (if continuous)
  - Mean for samples in same class
  - Most common attribute (if categorical)

Inapplicable Values

- Categorical Attributes: introduce special value ‘N/A’ as a new category
- Continuous Attributes: introduce special value outside the domain for example -1 for sales commission

Inconsistent Data

- Inconsistent Values:
  - Entry error, reading error, multiple sources
  - Negative values for weight
  - Zip code and city not matching

- Correction: requires information from external source or redundant information

Duplicate Data

- Duplicate Data

ID	Product Name	Price
r_1	iPad Two 16GB WiFi White	$490
r_2	iPad 2nd generation 16GB WiFi White	$469
r_3	iPhone 4th generation White 16GB	$545
r_4	Apple iPhone 4 16GB White	$520
r_5	Apple iPhone 3rd generation Black 16GB	$375
r_6	iPhone 4 32GB White	$599
r_7	Apple iPad2 16GB WiFi White	$499
r_8	Apple iPod shuffle 2GB Blue	$49
r_9	Apple iPod shuffle USB Cable	$19

- Deduplication: the process of dealing with duplicate data issues
  - Inconsistencies in duplicate records
  - Identify duplicate records

Duplicate Data

- Are these two records the same?

First Name	Last Name	Address	Phone
bob	roberts	1600 Pennsylvania ave.	555-0123
Robert	Roberts	1600 Pensylvania Avenue	

Deduplication Methods

- **Data preparation step**: Data transformation and standardization
  - Transform attributes from one data type or form to another, rename fields, standardize units, …
- **Distance based**:
  - Edit distance:
    - How many edits (insert/delete/replace) needed to transform one string to another
    - Same/Different cost operations

For example, the edit distance between "cake" and "asked" is 3:

1. *cake --→ aake (substitution of "a" for "c")*
2. *aake --→ aske (substitution of "s" for "a")*
3. *aske --→ asked (insertion of "d" at the end)*

Deduplication Methods

- Distance based methods: used to match individual fields
- Records consists of multiple fields

First Name	Last Name	Address	Phone
Robert	Jones	1600 Pensylvania Avenue	
Robert	Roberts	1600 Pensylvania Avenue	555-0123

- **Clustering based**: extends distance/similarity-based beyond pairwise comparison

Deduplication Methods

- **Classification based**:
  - Create a training set of duplicate and non-duplicate pairs of records
  - Train a classifier to distinguish duplicate from non-duplicate pairs

- *L. Breiman, L. Friedman, and P. Stone, (1984). Classification and Regression. Wadsworth, Belmont, CA.*
- *Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Classification and Regression Trees. Wadsworth and Brooks/Cole, 1984*

- *R. Agrawal, R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB-94, 1994.*
- *Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules In Proc. Of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994*

** Sarawagi and Bhamidipaty, Interactive Deduplication using Active Learning, KDD 2002*

Summary

- Types of data
- Attribute types
- Preliminary data exploration
  - Data point similarity
  - Data quality