PREPROCESSING- DATA PREPARATION

DRAFT

1. DATA PREPROCESSING: AN OVERVIEW
    a. DATA QUALITY
    b. MAJOR TASKS IN DATA PREPROCESSING
2. DATA CLEANING
3. DATA INTEGRATION
4. DATA REDUCTION
5. DATA TRANSFORMATION AND DATA DISCRETIZATION

A. Data Quality: Why Preprocess the Data?

Measures for data quality: A multidimensional view

1. Accuracy: correct or wrong, accurate or not
2. Completeness: not recorded, unavailable…
3. Consistency: some modified but some not
4. Timeliness: timely update?
5. Believability: how trustable the data are correct?
6. Interpretability: how easily the data can be understood?

B. Major Tasks in Data Preprocessing

1. Data Cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

A. INCOMPLETE: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
   e.g., *Occupation*=" " (missing data)

B. NOISY: containing noise, errors, or outliers
   e.g., *Salary*="−10" (an error)

C. INCONSISTENT: containing discrepancies in codes or names, e.g.,
   Age="42", *Birthday*="03/07/2010"
   Was rating "1, 2, 3", now rating "A, B, C"
   Discrepancy between duplicate records

D. INTENTIONAL (e.g., *disguised missing* data)
   Jan. 1 as everyone's birthday?

A. Incomplete (Missing) Data

- Data is not always available
  E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to - Reasons
  1. equipment malfunction
  2. inconsistent with other recorded data and thus deleted
  3. data not entered due to misunderstanding
  4. certain data may not be considered important at the time of entry
  5. not register history or changes of the data
  6. not applicable – as number of children for a single
- Missing data may need to be inferred

How to Handle Missing Data?

1) Ignore the Missing Value During Analysis.
2) Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
3) Fill in the missing value manually: tedious + infeasible?
4) Fill in it automatically with
   - A global constant: e.g., "unknown", a new class?!
   - the attribute mean
   - the attribute median
   - the attribute mean for all samples belonging to the same class: smarter
   - the most probable value: inference-based such as Bayesian formula or decision tree
5) Replace with all possible values (weighted by their probabilities)

B. Noisy Data

Noise: random error or variance in a measured variable

- Incorrect attribute values may be due to
  a. faulty data collection instruments
  b. data entry problems
  c. data transmission problems
  d. technology limitation
  e. inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

1. Binning

First sort data and partition into (equal-frequency) bins
then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc. There exist two methods

A. Equal-depth (frequency) partitioning

* Sort data and partition into bins, each containing approximately same number of samples
* Smooth by bin means, bin median, bin boundaries, etc.
* Good data scaling
* Managing categorical attributes can be tricky

* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

B. Equal-width (distance) partitioning

* divide the range into N intervals of equal size
* if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B-A)/N.
* Most straightforward
* Outliers may dominate presentation
* Skewed data is not handled well

2. REGRESSION
   Smooth by fitting the data into regression functions
   As f(x) = 2x +5
3. CLUSTERING
   Detect and remove outliers
4. COMBINED COMPUTER AND HUMAN INSPECTION
   Detect suspicious values and check by human (e.g., deal with possible outliers)

C. INCONSISTENCY

Very often, in large data sets, there exist samples that do not comply with the general behavior of the data model. Such samples, which are significantly different or inconsistent with the remaining set of data, are called outliers. Outliers can be caused by measurement error or they have be the result of inherent data variability. If, e.g., the display of a person's age in the database is −1 the value is obviously not correct, and error could have been caused by a default setting of the field "unrecorded age" in the computer program. On the other hand, if, in the database, the number of children for one person is 25 this datum is unusual and has to be checked. The value could be a typographical error, or it could be correct and represent real variability for the given attribute.
Many data-mining algorithms try to minimize the influence of outliers on the final model, some of the outliers detection methods are:

STATISTICAL APPROACH: Select a threshold point, and then decide that all points out of the range are considered as potential outliers
Threshold = Mean ± Constant * Standard Deviation
DISTANCE-BASED OUTLIER DETECTION: In this method, we assign 2 threshold values p for proportion (Frequency) and d for Euclidean distance, for example, if we set p>=4, d>=3 then the set of values that have distance >=3 and repeated 5 or more in each set are potential outliers.

EXAMPLE: if you have a set s={(2,4),(3,2),(1,1),(4,3),(1,6),(5,3),(4,2)} then after drawing the Euclidian Distance Table, the potential outliers are

	Euclidean Distance						
	1	2	3	4	5	6	7
1	**.000**	2.236	3.162	2.236	2.236	3.162	2.828
2	2.236	**.000**	2.236	1.414	4.472	2.236	1.000
3	3.162	2.236	**.000**	3.606	5.000	4.472	3.162
4	2.236	1.414	3.606	**.000**	4.243	1.000	1.000
5	2.236	4.472	5.000	4.243	**.000**	5.000	5.000
6	3.162	2.236	4.472	1.000	5.000	**.000**	1.414
7	2.828	1.000	3.162	1.000	5.000	1.414	**.000**

for the values of thresholds p ≥ 4 and d ≥ 3	
Sample	P
S1	2
S2	1
S3	5
S4	2
S5	5
S6	3
S7	2

We can conclude from the right table that S3, S5, and S6 are potential outliers.

DEVIATION-BASED TECHNIQUES: These techniques simulate the way in which humans can distinguish unusual samples from a set of other similar samples

DATA CLEANING

Process of dealing with duplicate data issues
Data set may include data objects that are duplicates, or almost duplicates of one another
Major issue when merging data from heterogeous sources

DATA CLEANING AS A PROCESS

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

DATA INTEGRATION

Integration of multiple databases, data cubes, or files
Combines data from multiple sources into a coherent store

SCHEMA INTEGRATION: e.g., A.cust-id ≡ B.cust-#

Integrate metadata from different sources

ENTITY IDENTIFICATION PROBLEM:

Identify real world entities from multiple data sources, e.g., U.S.A = U.S

DETECTING AND RESOLVING DATA VALUE CONFLICTS

For the same real world entity, attribute values from different sources are different
Possible reasons: different representations, different scales, e.g., metric vs. British units

HANDLING REDUNDANCY IN DATA INTEGRATION

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue

- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*

Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

CORRELATION ANALYSIS (NOMINAL DATA)

- **X^2 (chi-square) test**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Employed Observed /Expected	Not Employed Observed /Expected	Sum (row)
Male	300(110)	250(300)	550
Female	150(200)	600(520)	750
Sum(col.)	450	850	1300

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
- It shows that Male and Employed are correlated in the group

CORRELATION ANALYSIS (NUMERIC DATA)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

Where n is the number of tuples, and are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

2. Data reduction

- Dimensionality reduction
- Numerosity(Large Number) reduction
- Data compression

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies

Dimensionality reduction, e.g., remove unimportant attributes

- Wavelet transforms
- Principal Components Analysis (PCA)
- Feature subset selection, feature creation

Numerosity reduction (some simply call it: Data Reduction)

- Regression and Log-Linear Models
- Histograms, clustering, sampling
- Data cube aggregation: Combining two or more attributes (or objects) into a single attribute (or object)

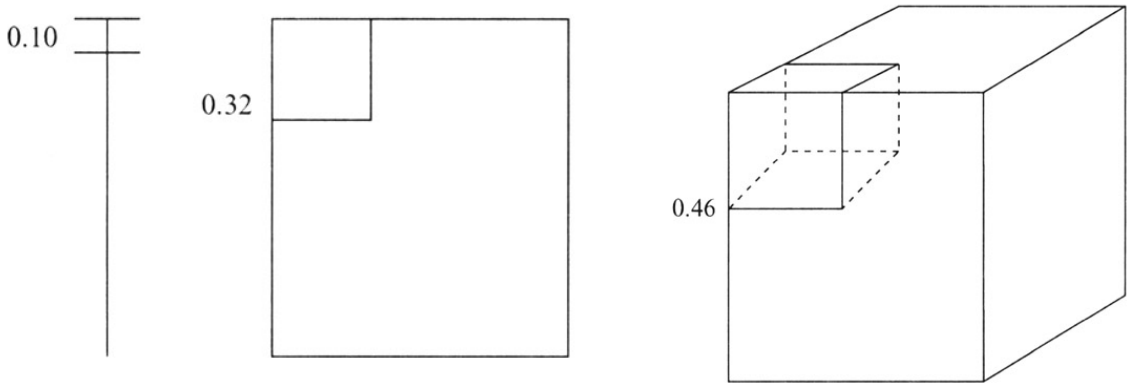
Data compression

Data Reduction: Dimensionality Reduction

Curse of dimensionality

When dimensionality increases, data becomes increasingly sparse
Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
The possible combinations of subspaces will grow exponentially

- The size of a data set yielding the same density of data points in an n-dimensional space increases exponentially with dimensions
  - A larger radius is needed to enclose a fraction of the data points in a high-dimensional space. For a given fraction of samples, it is possible to determine the edge length e of the hypercube using the formula $E(p)=p^{1/d}$
  - where p is the prespecified fraction of samples and d is the number of dimensions. For example, if one wishes to enclose 10% of the samples (p = 0.1), the corresponding edge for a two-dimensional space will be $e_2(0.1)$ = 0.32, for a three-dimensional space $e_3(0.1)$ = 0.46, and for a 10-dimensional space e10(0.1) = 0.80. Graphical interpretation of these edges is given in



- Almost every point is closer to an edge than to another sample point in a high-dimensional space. For a sample size n, the expected distance D between data points in a d-dimensional spaces is $D(d,n) = \frac{1}{2}(1/n)^{1/d}$

  For example, for a two-dimensional space with 10000 points the expected distance is D(2,10000) = 0.0005 and for a 10-dimensional space with the same number of sample points D(10,10000) = 0.4. Keep in mind that the maximum distance from any point to the edge occurs at the center of the distribution, and it is 0.5 for normalized values of all dimensions.

- Almost every point is an outlier. As the dimension of the input space increases. the distance between the prediction point and the center of the classified points increases. For example, when d = 10, the expected value of the prediction point is 3.1 standard deviations away from the center of the data belonging to one class. When d = 20, the distance is 4.4 standard deviations. From this standpoint, the prediction of every new point looks like an outlier of the initially classified data. This is illustrated conceptually in Figure 2.1, where predicted points are mostly in the edges of the porcupine, far from the central part.

DIMENSIONALITY REDUCTION BENEFITS

1. Avoid the curse of dimensionality
2. Help eliminate irrelevant features and reduce noise
3. Reduce time and space required in data mining
4. Allow easier visualization

- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

ATTRIBUTE SUBSET SELECTION

Sampling is the main technique employed for data selection
Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

- The key principle for effective sampling is the following:

- using a sample will work almost as well as using the entire data sets, if the sample is representative
- A sample is representative if it has approximately the same property (of interest) as the original set of data

Types of Sampling

SIMPLE RANDOM SAMPLING
- There is an equal probability of selecting any particular item

SAMPLING WITHOUT REPLACEMENT
- As each item is selected, it is removed from the population

SAMPLING WITH REPLACEMENT
- Objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked up more than once

STRATIFIED SAMPLING
- Split the data into several partitions; then draw random samples from each partition

Another way to reduce dimensionality of data

REDUNDANT ATTRIBUTES
- Duplicate much or all of the information contained in one or more other attributes
  E.g., purchase price of a product and the amount of sales tax paid

IRRELEVANT ATTRIBUTES

Contain no information that is useful for the data mining task at hand
e.g., students' ID is often irrelevant to the task of predicting students' GPA
e.g., Attribute that has atomic value.

ATTRIBUTE CREATION (FEATURE GENERATION)

Create new attributes (features) that can capture the important information in a data set more effectively than the original ones

- Three general methodologies

1. FEATUE EXTRACTION
   - Domain-specific
2. MAPPING DATA TO NEW SPACE (SEE: DATA REDUCTION)

   E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
3. FEATUE CONSTRUCTION

Example: Calculate a new field BMI according to the formula
BMI (Metric)= Weigt / Height^2

18.5 or less	Underweight
18.5-24.9	Normal Weight
25.0-25.9	Overweight
30.0-34.9	Obese
35.0-39.9	Obese
40 or greater	Extremely Obese

3. DATA TRANSFORMATION AND DATA DISCRETIZATION

Data smoothing techniques are used to eliminate "noise" and extract real trends and patterns.

Smoothing: Remove noise from data

Some methods used in data smoothing:

SIMPLE ROUNDING TECHNIQUE

Eliminate the Minor differences between values that are not significant and may degrade the performance of the method and the final results.

Example:

If the set of values for the given feature F is {0.93, 1.01, 1.001, 3.02, 2.99, 5.03, 5.01, 4.98}, then it is obvious that smoothed values will be $F_{smoothed}$ = {1.0, 1.0, 1.0, 3.0, 3.0, 5.0, 5.0, 5.0}.

MOVING AVERAGES

- Attribute/feature construction
  - New attributes constructed from the given ones

  Why: the given attributes might be unuseful in our analysis

  How: using any suitable function/formula

- Aggregation: Summarization, data cube construction
- Normalization: Scaled to fall within a smaller, specified range

NORMALIZATION

A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values

DECIMAL SCALING: moves the decimal point but still preserves most of the original digit value. The typical scale maintains the values in a range of −1 to 1

$V'(i)=V(i)/10^k$ For the smallest k such that max (|v'(i)|) < 1.

Example:

For example, if the largest value in the set is 455 and the smallest value is −834, then the maximum absolute value of the feature becomes .834, and the divisor for all v(i) is 1000 (k = 3).

Suppose that the data for a feature v are in a range between 150 and 250. Then, the previous method of normalization will give all normalized data between .15 and .25, To obtain better distribution of values on a whole, normalized interval, e.g., [0, 1], we can use the min-max formula

MIN-MAX NORMALIZATION:

$V'(i)=\left(V(i) - \min(V(i))\right)/\left(\max(V(i)) - \min(V(i))\right)$

Z-SCORE NORMALIZATION / STANDARD DEVIATION NORMALIZATION:

$V'(i)=\left(V(i)-\text{mean}(V(i))\right) /\text{sd}(v)$

For example, if the initial set of values of the attribute is v = {1, 2, 3}, then *mean* (v) = 2, *sd*(v) = 1, and the new set of normalized values is v* = {−1, 0, 1}.

References

P. Adriaans and D. Zantinge. *Data Mining*. Addison-Wesley: Harlow, England, 1996.
R.J. Brachman, T. Anand. The process of knowledge discovery in databases. Advances in Knowledge
Discovery and Data Mining. AAAI/MIT Press, 1996.
D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of
ACM, 42:73-78, 1999.
M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans.
Knowledge and Data Engineering, 8:866-883, 1996.
U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and
Data Mining*. AAAI/MIT Press, 1996.
T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Communications of ACM*,
39:58-64, 1996.
Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data
Engineering, 20(4), December 1997.
D. Keim, Visual techniques for exploring databases. Tutorial notes in KDD'97, Newport Beach, CA, USA,
1997.
D. Keim, Visual data mining. Tutorial notes in VLDB'97, Athens, Greece, 1997.
D. Keim, and H.P. Krieger, Visual techniques for mining large databases: a comparison. IEEE Transactions on
Knowledge and Data Engineering, 8(6), 1996.
W. Kloesgen, Explora: A multipattern and multistrategy discovery assistant. In U.M. Fayyad, et al. (eds.),
Advances in Knowledge Discovery and Data Mining, 249-271. AAAI/MIT Press, 1996.