DATA MINING : DATA

LECTURE 8

- DATA QUALITY PROBLEMS
  - NOISE
  - OUTLIER
  - MISSING VALUES
  - DUPLICATE DATA
- WHAT IS DATA PREPROCESSING?
- WHY DATA PREPROCESSING?
- MEASURE OF DATA QUALITY
- MAJOR TASKS IN DATA PREPROCESSING
- DATA CLEANING

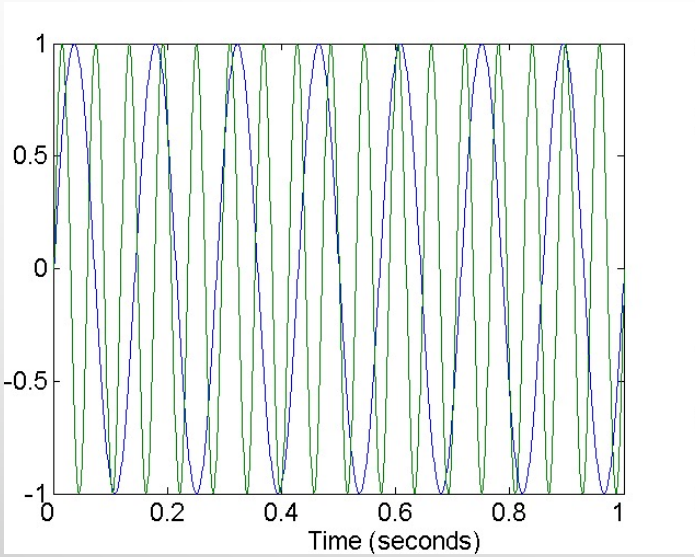
DATA QUALITY

- WHAT KINDS OF DATA QUALITY PROBLEMS?
- HOW CAN WE DETECT PROBLEMS WITH THE DATA?
- WHAT CAN WE DO ABOUT THESE PROBLEMS?

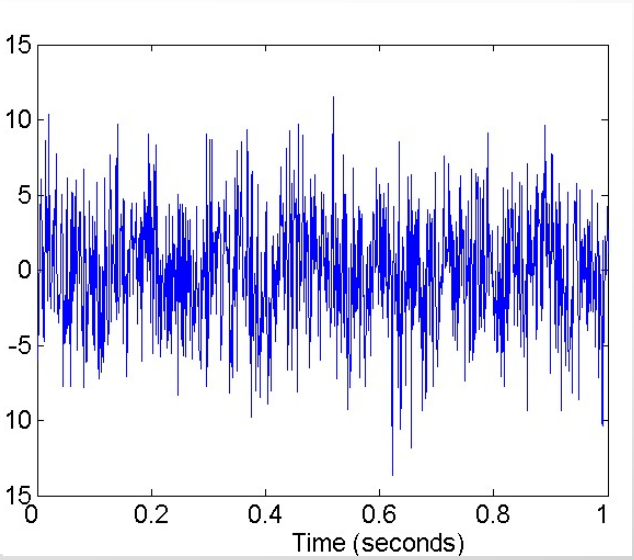
- EXAMPLES OF DATA QUALITY PROBLEMS:
  - NOISE AND OUTLIERS
  - MISSING VALUES
  - DUPLICATE DATA

NOISE

- NOISE REFERS TO MODIFICATION OF ORIGINAL VALUES
  - EXAMPLES: DISTORTION OF A PERSON'S VOICE WHEN TALKING ON A POOR PHONE AND "SNOW" ON TELEVISION SCREEN



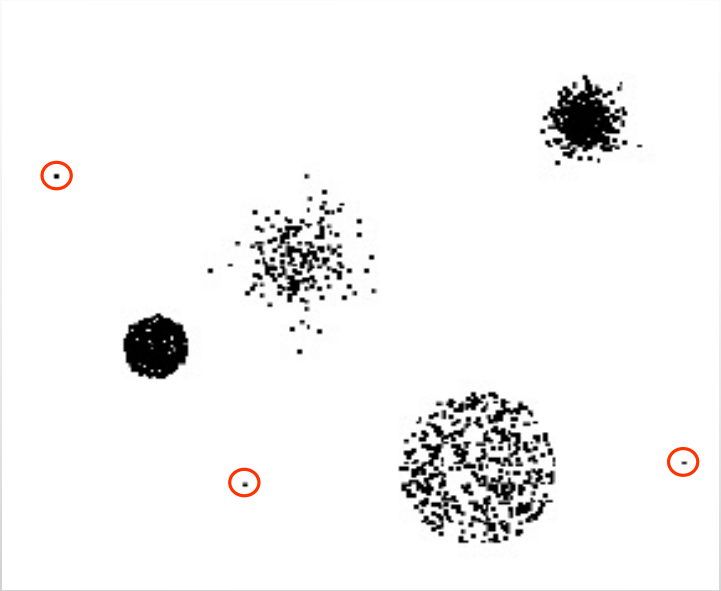
Two Sine Waves



Two Sine Waves + Noise

OUTLIERS

- OUTLIERS ARE DATA OBJECTS WITH CHARACTERISTICS THAT ARE CONSIDERABLY DIFFERENT THAN MOST OF THE OTHER DATA OBJECTS IN THE DATA SET



MISSING VALUES

- REASONS FOR MISSING VALUES
  - INFORMATION IS NOT COLLECTED
    (E.G., PEOPLE DECLINE TO GIVE THEIR AGE AND WEIGHT)
  - ATTRIBUTES MAY NOT BE APPLICABLE TO ALL CASES
    (E.G., ANNUAL INCOME IS NOT APPLICABLE TO CHILDREN)

- HANDLING MISSING VALUES
  - ELIMINATE DATA OBJECTS
  - ESTIMATE MISSING VALUES
  - IGNORE THE MISSING VALUE DURING ANALYSIS
  - REPLACE WITH ALL POSSIBLE VALUES (WEIGHTED BY THEIR PROBABILITIES)

DUPLICATE DATA

- DATA SET MAY INCLUDE DATA OBJECTS THAT ARE DUPLICATES, OR ALMOST DUPLICATES OF ONE ANOTHER
  - MAJOR ISSUE WHEN MERGING DATA FROM HETEROGEOUS SOURCES
- EXAMPLES:
  - SAME PERSON WITH MULTIPLE EMAIL ADDRESSES
- DATA CLEANING
  - PROCESS OF DEALING WITH DUPLICATE DATA ISSUES

WHAT IS PREPROCESSING?

- DATA PREPROCESSING DESCRIBES ANY TYPE OF PROCESSING PERFORMED ON RAW DATA TO PREPARE IT FOR *ANOTHER* PROCESSING PROCEDURE.

- COMMONLY USED AS A PRELIMINARY DATA MINING PRACTICE, DATA PREPROCESSING TRANSFORMS THE DATA INTO A FORMAT THAT WILL BE MORE EASILY AND EFFECTIVELY PROCESSED FOR THE PURPOSE OF THE USER

WHY DATA PREPROCESSING?

- DATA IN THE REAL WORLD IS DIRTY
  - INCOMPLETE: LACKING ATTRIBUTE VALUES, LACKING CERTAIN ATTRIBUTES OF INTEREST, OR CONTAINING ONLY AGGREGATE DATA
  - NOISY: CONTAINING ERRORS OR OUTLIERS
  - INCONSISTENT: CONTAINING DISCREPANCIES IN CODES OR NAMES
- NO QUALITY DATA, NO QUALITY MINING RESULTS!
  - QUALITY DECISIONS MUST BE BASED ON QUALITY DATA
  - DATA WAREHOUSE NEEDS CONSISTENT INTEGRATION OF QUALITY DATA

DATA QUALITY

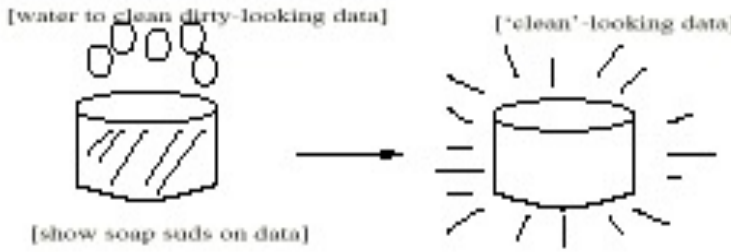
- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

MAJOR TASKS IN DATA PREPROCESSING

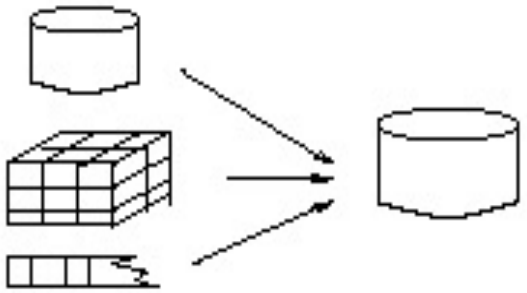
- DATA CLEANING
  - FILL IN MISSING VALUES, SMOOTH NOISY DATA, IDENTIFY OR REMOVE OUTLIERS, AND RESOLVE INCONSISTENCIES
- DATA INTEGRATION
  - INTEGRATION OF MULTIPLE DATABASES, DATA CUBES, OR FILES
- DATA TRANSFORMATION
  - NORMALIZATION AND AGGREGATION
- DATA REDUCTION
  - OBTAINS REDUCED REPRESENTATION IN VOLUME BUT PRODUCES THE SAME OR SIMILAR ANALYTICAL RESULTS
- DATA DISCRETIZATION
  - PART OF DATA REDUCTION BUT WITH PARTICULAR IMPORTANCE, ESPECIALLY FOR NUMERICAL DATA

FORMS OF DATA PREPROCESSING

Data Cleaning

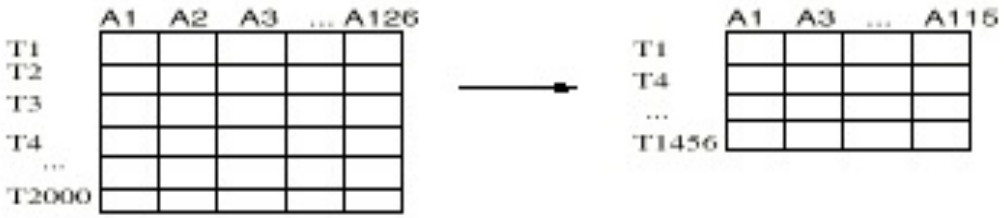


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



DATA CLEANING

- DATA CLEANING TASKS
  - FILL IN MISSING VALUES
  - IDENTIFY OUTLIERS AND SMOOTH OUT NOISY DATA
  - CORRECT INCONSISTENT DATA

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
  - EQUIPMENT MALFUNCTION
  - INCONSISTENT WITH OTHER RECORDED DATA AND THUS DELETED
  - DATA NOT ENTERED DUE TO MISUNDERSTANDING
  - CERTAIN DATA MAY NOT BE CONSIDERED IMPORTANT AT THE TIME OF ENTRY
  - NOT REGISTER HISTORY OR CHANGES OF THE DATA
- MISSING DATA MAY NEED TO BE INFERRED.

MISSING DATA EXAMPLE
BANK ACCT TOTALS - HISTORICAL

Name	SSN	Address	Phone #	Date	Acct Total
John Doe	111-22-3333	1 Main St Bedford, Ma	111-222-3333	2/12/1999	2200.12
John W. Doe		Bedford, Ma		7/15/2000	12000.54
John Doe	111-22-3333			8/22/2001	2000.33
James Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444	12/22/2002	15333.22
Jim Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444		12333.66
Jim Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444		

HOW SHOULD WE HANDLE THIS?

HOW TO HANDLE MISSING DATA?

- IGNORE THE TUPLE: USUALLY DONE WHEN CLASS LABEL IS MISSING (ASSUMING THE TASKS IN CLASSIFICATION—NOT EFFECTIVE WHEN THE PERCENTAGE OF MISSING VALUES PER ATTRIBUTE VARIES CONSIDERABLY.
- FILL IN THE MISSING VALUE MANUALLY: TEDIOUS + INFEASIBLE?
- USE A GLOBAL CONSTANT TO FILL IN THE MISSING VALUE: E.G., "UNKNOWN", A NEW CLASS?!
- USE THE ATTRIBUTE MEAN TO FILL IN THE MISSING VALUE
- USE THE ATTRIBUTE MEAN FOR ALL SAMPLES BELONGING TO THE SAME CLASS TO FILL IN THE MISSING VALUE: SMARTER
- USE THE MOST PROBABLE VALUE TO FILL IN THE MISSING VALUE: INFERENCE-BASED SUCH AS BAYESIAN FORMULA OR DECISION TREE

NOISY DATA

- NOISE: RANDOM ERROR OR VARIANCE IN A MEASURED VARIABLE
- INCORRECT ATTRIBUTE VALUES MAY BE DUE TO
  - FAULTY DATA COLLECTION INSTRUMENTS
  - DATA ENTRY PROBLEMS
  - DATA TRANSMISSION PROBLEMS
  - TECHNOLOGY LIMITATION
  - INCONSISTENCY IN NAMING CONVENTION
- OTHER DATA PROBLEMS WHICH REQUIRES DATA CLEANING
  - DUPLICATE RECORDS
  - INCOMPLETE DATA
  - INCONSISTENT DATA

NOISY DATA EXAMPLE
BANK ACCT TOTALS - HISTORICAL

Name	SSN	Address	Phone #	Date	Acct Total
John Doe	111-22-3333	1 Main St Bedford, Ma	111-222-3333	2/12/1999	2200.12
John Doe	111-22-3333	1 Main St Bedford, Ma	111-222-3333	2/12/1999	2233.67
James Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444	12/22/2002	15333.22
James Smith	222-33-4444	2 Oak St Boston, Ma	222-333-4444	12/23/2003	15333000.00

HOW SHOULD WE HANDLE THIS?

HOW TO HANDLE NOISY DATA?

- BINNING METHOD:
  - FIRST SORT DATA AND PARTITION INTO (EQUI-DEPTH) BINS
  - THEN ONE CAN SMOOTH BY BIN MEANS, SMOOTH BY BIN MEDIAN, SMOOTH BY BIN BOUNDARIES, ETC.
- CLUSTERING
  - DETECT AND REMOVE OUTLIERS
- COMBINED COMPUTER AND HUMAN INSPECTION
  - DETECT SUSPICIOUS VALUES AND CHECK BY HUMAN
- REGRESSION
  - SMOOTH BY FITTING THE DATA INTO REGRESSION FUNCTIONS

DISCRETIZATION

- THREE TYPES OF ATTRIBUTES
  - NOMINAL—VALUES FROM AN UNORDERED SET, E.G., COLOR, PROFESSION
  - ORDINAL—VALUES FROM AN ORDERED SET, E.G., MILITARY OR ACADEMIC RANK
  - NUMERIC—REAL NUMBERS, E.G., INTEGER OR REAL NUMBERS
- DISCRETIZATION: DIVIDE THE RANGE OF A CONTINUOUS ATTRIBUTE INTO INTERVALS
  - INTERVAL LABELS CAN THEN BE USED TO REPLACE ACTUAL DATA VALUES
  - REDUCE DATA SIZE BY DISCRETIZATION
  - SUPERVISED VS. UNSUPERVISED
  - SPLIT (TOP-DOWN) VS. MERGE (BOTTOM-UP)
  - DISCRETIZATION CAN BE PERFORMED RECURSIVELY ON AN ATTRIBUTE
  - PREPARE FOR FURTHER ANALYSIS, E.G., CLASSIFICATION

DATA DISCRETIZATION METHODS

- TYPICAL METHODS: ALL THE METHODS CAN BE APPLIED RECURSIVELY
  - BINNING
    - TOP-DOWN SPLIT, UNSUPERVISED
  - HISTOGRAM ANALYSIS
    - TOP-DOWN SPLIT, UNSUPERVISED
  - CLUSTERING ANALYSIS (UNSUPERVISED, TOP-DOWN SPLIT OR BOTTOM-UP MERGE)
  - DECISION-TREE ANALYSIS (SUPERVISED, TOP-DOWN SPLIT)
  - CORRELATION (E.G., χ^2) ANALYSIS (UNSUPERVISED, BOTTOM-UP MERGE)

SIMPLE DISCRETIZATION METHODS: BINNING

- EQUAL-WIDTH (DISTANCE) PARTITIONING:
  - IT DIVIDES THE RANGE INTO *N* INTERVALS OF EQUAL SIZE: UNIFORM GRID
  - IF *A* AND *B* ARE THE LOWEST AND HIGHEST VALUES OF THE ATTRIBUTE, THE WIDTH OF INTERVALS WILL BE: $W = (B-A)/N$.
  - THE MOST STRAIGHTFORWARD
  - BUT OUTLIERS MAY DOMINATE PRESENTATION
  - SKEWED DATA IS NOT HANDLED WELL.
- EQUAL-DEPTH (FREQUENCY) PARTITIONING:
  - IT DIVIDES THE RANGE INTO *N* INTERVALS, EACH CONTAINING APPROXIMATELY SAME NUMBER OF SAMPLES
  - GOOD DATA SCALING
  - MANAGING CATEGORICAL ATTRIBUTES CAN BE TRICKY.

BINNING METHODS FOR DATA SMOOTHING

* SORTED DATA FOR PRICE (IN DOLLARS): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* PARTITION INTO (EQUI-WIDTH) BINS:
  - BIN 1 (4-14): 4, 8, 9
  - BIN 2(15-25): 15, 21, 21, 24
  - BIN 3(25-34): 25, 26, 28, 29, 34
* SMOOTHING BY BIN MEANS:
  - BIN 1: 7, 7, 7
  - BIN 2: 20, 20, 20, 20
  - BIN 3: 28, 28, 28, 28, 28
* SMOOTHING BY BIN BOUNDARIES:
  - BIN 1: 4, 4, 4
  - BIN 2: 15, 24, 24, 24
  - BIN 3: 25, 25, 25, 25, 34

BINNING METHODS FOR DATA SMOOTHING (CONTINUED)

* 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

REFERENCES

1. Pang-Ning Tan, Vipin Kumar, Michael Steinbach: **Introduction to Data Mining, chapter 2,** Pearson, 2012.
2. Jiawei Han and Micheline Kamber: **Data Mining - Concepts and Techniques, chapter 3,** 3rd Edition, MorganKaufmann Publisher, 2014