Data Mining :: Unit-2

(Data Types, Attributes, Data Preprocessing)

Er. Dinesh Baniya Kshatri
(Lecturer)

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attribute Values

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Scales of Measurement

Description	Scale	Type
Differences between measurements, true zero exists	Ratio Data	Quantitative Data
Differences between measurements but no true zero	Interval Data	Quantitative Data
Ordered Categories (rankings, order, or scaling)	Ordinal Data	Qualitative Data
Categories (no ordering or direction)	Nominal Data	Qualitative Data

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $= \ \neq$
  - Order: $< \ >$
  - Addition: $+ \ -$
  - Multiplication: $* \ /$

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Attribute Type	Description	Examples
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ($=, \neq$)	zip codes, employee ID numbers, eye color, sex: {*male, female*}
Ordinal	The values of an ordinal attribute provide enough information to order objects. ($<, >$)	hardness of minerals, {*good, better, best*}, grades, street numbers
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ($+, -$)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	For ratio variables, both differences and ratios are meaningful. ($*, /$)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Categorical: Nominal variables

Person	Occupation
Eirini	archaeologist
Erich	engineer
Kostas	doctor
Jane	engineer
Emily	teacher
Markus	driver

- *No ordering* in the categories/ states.
- Only *distinctness relationships* apply, i.e.,
  - *equal (=)* and
  - *different* (≠)

Categorical: Ordinal variables

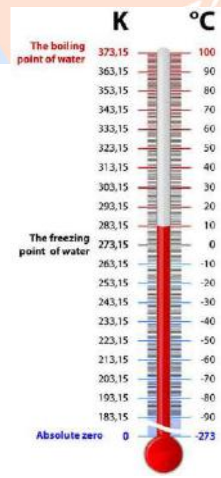
Person	A beautiful mind	Titanic
Eirini	5*	3*
Erich	5*	1*
Kostas	3*	3*
Jane	1*	2*
Emily	2*	5*
Markus	4*	3*

- Allows to apply *order relationships*, i.e., >, ≥, <, ≤
- However, the *difference* and *ratio* between these values has no meaning.
  - E.g., 5*-3* is the same as 3*-1* or, 4* is 2 times better than 2*?

Numeric: Interval Variables

- Differences between values are meaningful
  - The difference between 90° and 100° *temperature* is the same as the difference between 40° and 50° *temperature*.
- Examples:
  - *Calendar dates* , *Temperature* in Farenheit or Celsius, ...
- Ratio still has no meaning
  - A *temperature* of 2° Celsius is not much different than a *temperature* of 1° Celsius.
  - The issue is that the 0° point of the Celsius scale is in a physical sense arbitrary and therefore the ratio of two *Celsius temperatures* is not physically meaningful.

Numeric: Ratio-scale Variables



- Both differences and ratios have a meaning
  - E.g., a 100 kgs person is twice heavy as a 50 kgs person.
  - E.g., a 50 years old person is twice old as a 25 years old person.
- Meaningful (unique and non-arbitrary) zero value
- Examples:
  - *age, weight, length, number of sales*
  - *temperature in Kelvin*
    - When measured on the Kelvin scale, a temperature of 2° is, in a physical meaningful way, twice that of a 1°.
      - The zero value is absolute 0, represents the complete absence of molecular motion

Important Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Types of Data Sets

- Record
  - Tables
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

Tabular Data

- A collection of records
  - Each record is characterized by a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Document Data

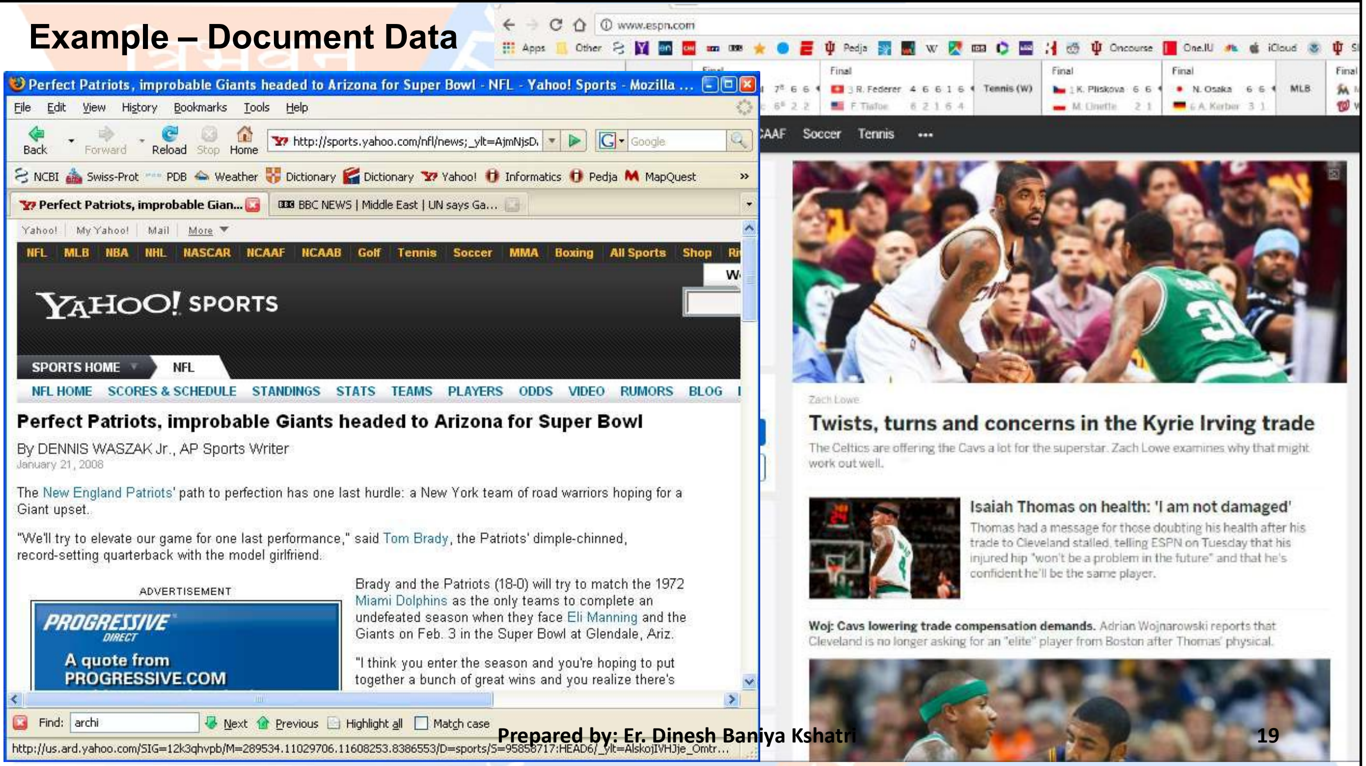
- It includes textual data that can be **semi-structured** or **unstructured**
  - Plain text can be organized in sentences, paragraphs, sections, documents
- Text acquired in different contexts may have a structure and/or a semantics
  - **Web pages** are enriched with tags
  - **Documents in digital libraries** are enriched with metadata
  - **E-learning documents** can be annotated or partly highglihted

Document Data

- Each document becomes a 'term' vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

	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

Example – Document Data



Prepared by: Er. Dinesh Baniya Kshatri 19

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Prepared by: Er. Dinesh Baniya Kshatri 20

Recommendations Data

- Sparse matrix
  - each row is a person
  - each column is a movie (book, disease, …)
  - each number is a rating

Movies

Persons

	Spiderman	Ocean's 11	Matrix	Titanic	JFK	Star wars	Creed	Rocky
Person 1		3		4				
Person 2								5
Person 3							4	5
Person 4	1		3				2	

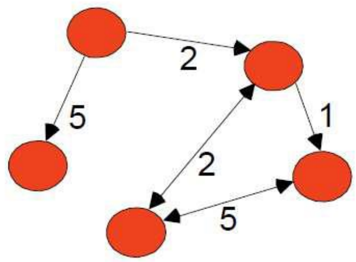
Transaction Data

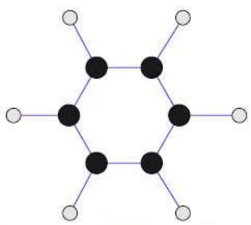
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

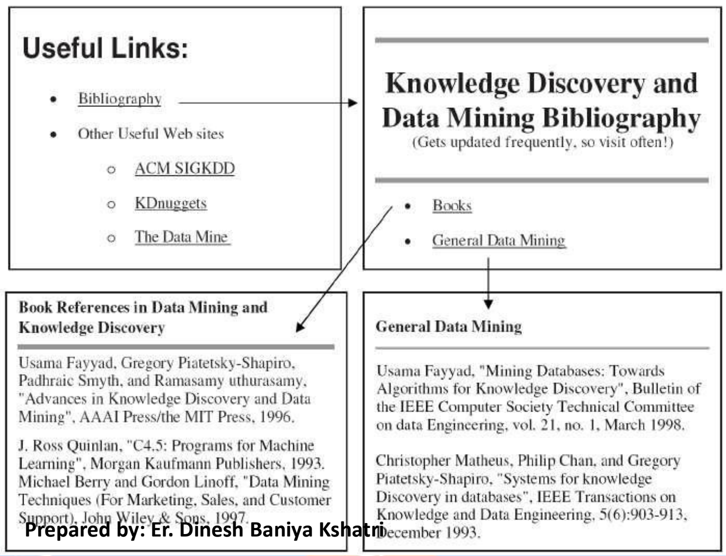
Graph Data

- Examples: Generic graph, a molecule, and webpages



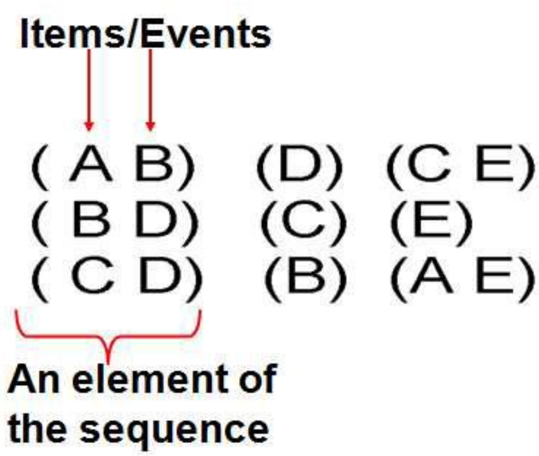


Benzene Molecule: C6H6



Ordered Data

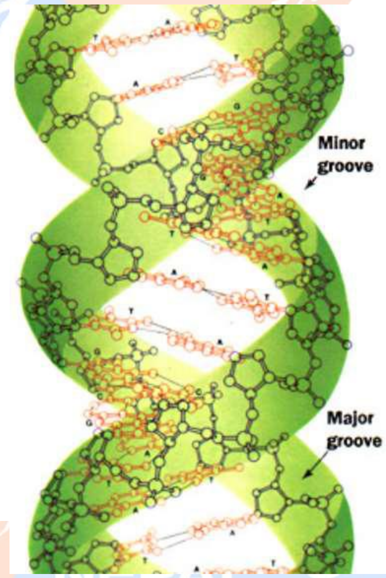
- Sequences of transactions



Ordered Data

- Genomic sequence data

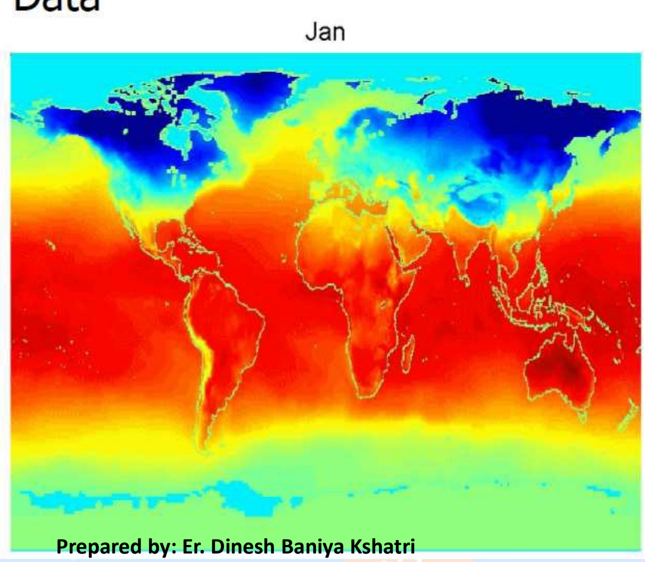
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



Ordered Data

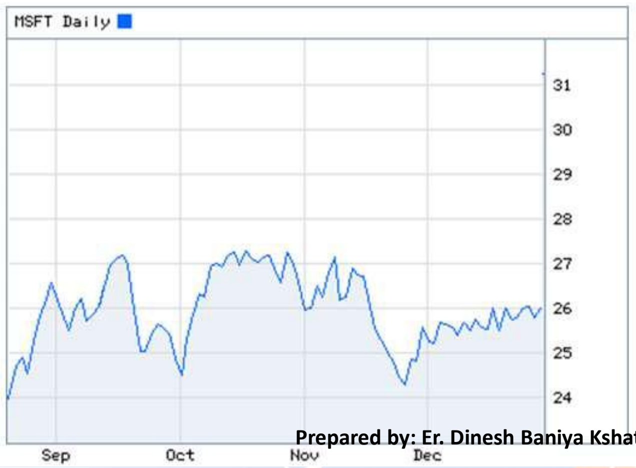
- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Ordered Data

- Time series
  - Sequence of ordered (over "time") numeric values.



Data Quality

- Poor data quality negatively affects many data processing efforts

"The most important point is that poor data quality is an unfolding disaster.

– Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

Thomas C. Redman, DM Review, August 2004

Data Quality Problems

- Examples of data quality problems
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

Examples of Dirty Data

- Data in the real world is dirty
  - **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=" "
  - **noisy:** containing errors or outliers
    - e.g., Salary="-10"
  - **inconsistent:** containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"

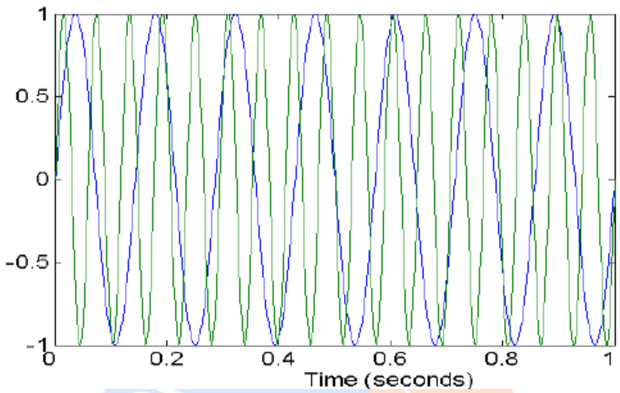
Why is Data Dirty?

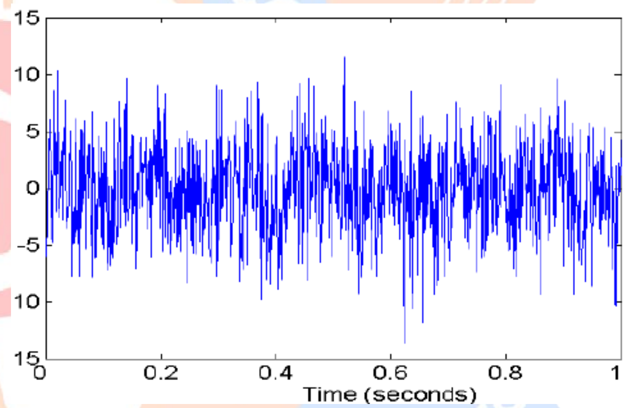
- **Incomplete data** may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- **Noisy data** (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- **Inconsistent** data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

Noise

- Noise is the random component of measurement error
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
- In general hard to remove the noise without losing some of the useful information (signal)
  - For data with temporal (e.g. speech) or spatial component (images), there are noise reduction techniques that can *partially* solve this problem
- As an alternative, development of algorithms that are robust with respect to noisy data (i.e. do not completely break down) is an important theme in data mining

Examples of Noisy Data

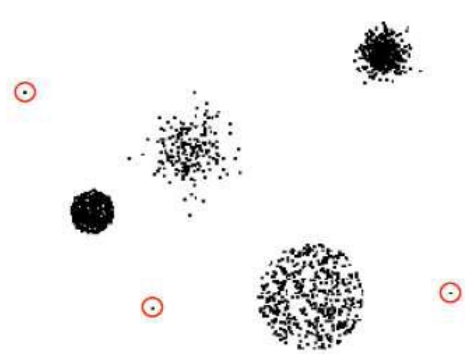




Two Sine Waves **Two Sine Waves + Noise**

Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



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)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	?	**No**
8	No	Single	?	**Yes**
9	No	Married	?	**No**
10	No	Single	90K	**Yes**

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
- Examples:
  - Same person with multiple email addresses
  - Laboratory experiments that has been performed as duplicate
    - very common practise in, e.g. biological sciences
- Need to
  - Detect whether two records represent the same object
  - Merge only if they do
  - For merging need to resolve inconsistencies in values
    - averaging or selecting one representative value

Measures of Data Quality

- A well-accepted multidimensional view of data quality:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability
  - Accessibility

Data Preprocessing

- Integration
- Data cleaning
- Aggregation
- Sampling
- Dimensionality reduction
- Feature subset selection
- Feature creation
- Discretization and binarization
- Data transformation

Data Integration

- Data integration:
  - 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., Bill Clinton = William Clinton
- 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

Data Integration (Handling Redundancy)

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

Data Cleaning

- Data cleaning tasks:
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Prepared by: Er. Dinesh Baniya Kshatri 41

Data Cleaning
How to Handle Missing Data?

- (1) Ignore the tuple (record) : usually done when class label is missing (assuming the tasks in classification)
- It is not effective when the percentage of missing values per attribute varies considerably.
- (2) Fill in the missing value manually: tedious + infeasible?

Prepared by: Er. Dinesh Baniya Kshatri 42

Data Cleaning
How to Handle Missing Data?

- **(3) Use a global constant** to fill in the missing value (be careful- it introduces a new class)
- **(4) Use the attribute values mean** to fill in the missing value
- **(5) Use the attribute values mean for all** samples belonging to the same class to fill in the missing value: smarter then **(4)** in case of classification
- **(6) Use the most probable value** to fill in the missing value
- **(7)** Use regresion methods

Prepared by: Er. Dinesh Baniya Kshatri 43

Handling Missing Values
(Eliminating Data Objects)

- Eliminating data objects with missing values is simple and effective
- If too large fraction of data contains missing values, we may not be able to make reliable analysis with the remaining data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	?	**No**
8	No	Single	?	**Yes**
9	No	Married	?	**No**
10	No	Single	90K	**Yes**

Prepared by: Er. Dinesh Baniya Kshatri 44

Handling Missing Values
(Eliminating Attributes)

- Eliminating attributes with missing values is an alternative
- Should be performed with caution, since the attribute we are removing may be crucial for the analysis

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	?	No
8	No	Single	?	Yes
9	No	Married	?	No
10	No	Single	90K	Yes

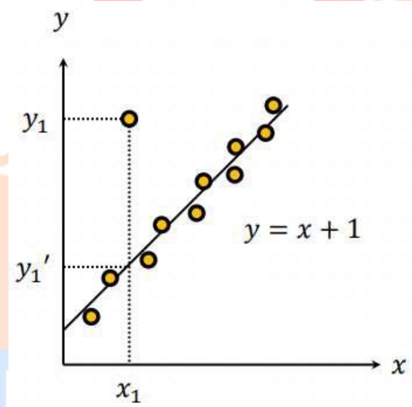
Handling Missing Values
(Estimating Missing Values)

- In some cases it is possible to estimate the missing value from the values of other data points
- If the data has temporal or spatial structure, interpolation between points close in time or space can give a good result
- In record based data, we can look for similar records and use the central value (mean, median, or mode)
- Methods estimating the missing values are often called *imputation methods*

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	?	No
8	No	Single	?	Yes
9	No	Married	80K	No
10	No	Single	90K	Yes

Handling Missing Values
Linear Regression

- Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line



Data Cleaning
Noisy Data

- **Noise:** random error or variance in a measured variable (numeric attribute value)
- **Incorrect attribute values** may due to
  - faulty data collection instruments,
  - data entry problems,
  - data transmission problems,
  - technology limitation,
  - inconsistency in naming convention

Data Cleaning
Handle Noisy Data

- Binning
  - sort data and partition into (equi-depth) bins
  - smooth by bin means, bin median, bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values automatically and check by human

Data Cleaning
How to Handle Noisy Data?

- Equal-width (distance) partitioning
  - 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
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Data Cleaning
How to Handle Noisy Data

- Binning method:
  - first **sort data** (values of the attribute we consider) and **partition them** into (equal-depth) bins
  - Apply one of the methods:
  - **smooth by bin means** - replace noisy values in the bin by the **bin mean**
  - **smooth by bin median** - replace noisy values in the bin by the **bin median**)
  - smo**oth by bin boundaries** - replace noisy values in the bin by the **bin boundaries**)

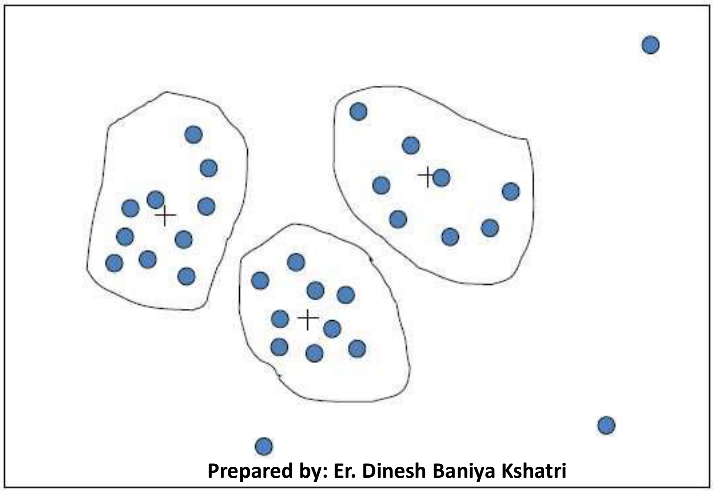
Data Cleaning
Binning Methods for Data Smoothing

- Sorted data (attribute values) **for price (attribute: price in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
- **Partition into (equal-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**
- **Replace all values in a BIN by ONE value (smoothing values)**

Data Cleaning

Cluster Analysis

Clustering: detect and remove outliers

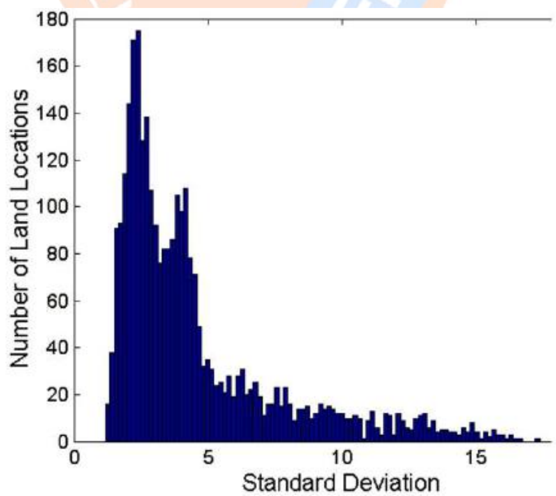


Aggregation

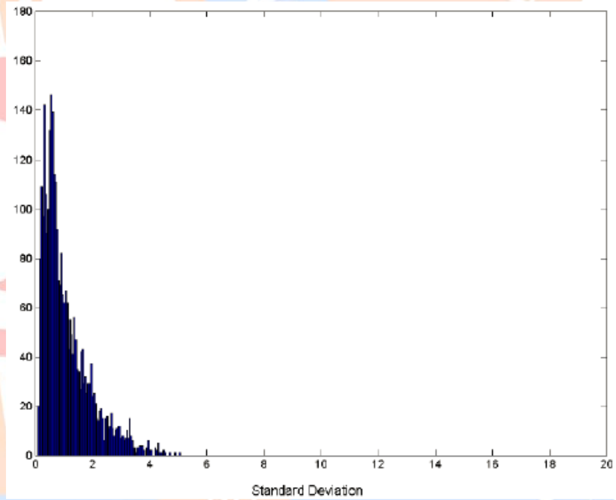
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
    - Faster to process, easier to fit to computer main memory
  - Change of scale
    - E.g. Cities aggregated into regions, states, countries, etc
  - More "stable" data
    - Aggregated data tends to have less variability due to random effects (less noise, less outliers)

Variation of Precipitation in Australia

(From the period 1982 to 1993)



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

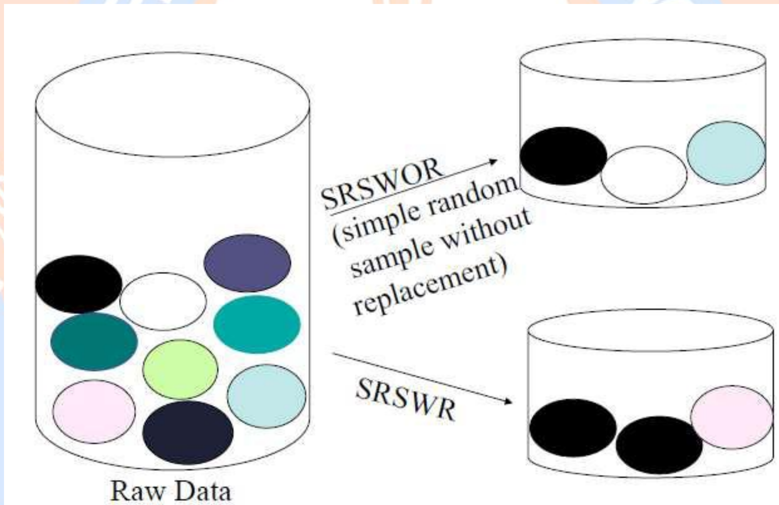
Sampling …

- 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
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition
- 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. The same object can be picked up more than once.

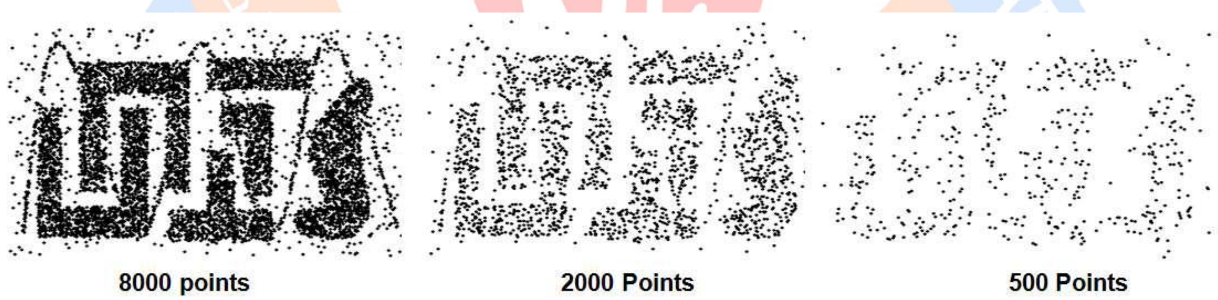
Sampling: with or without Replacement



Prepared by: Er. Dinesh Baniya Kshatri 59

Choosing the Sample Size

- It is important to choose a sample size that is
  - Large enough to enable to recover the structure in the original data (i.e. has approximately the same property than the orginal data)
  - Small enough to give as savings in processing time and space



Prepared by: Er. Dinesh Baniya Kshatri 60

Stratified Sampling

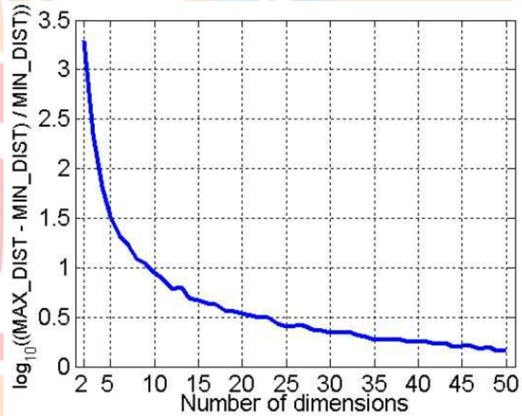
- Stratified sampling works better for data with many different groups
  - Divide the data into the groups
  - Sample from each group
    - Equal number of samples, or
    - With probability proportional to the group size
- For example, think about a questionaire to 1000 european people
  - Simple random sampling might results in no or very few samples from small population countries such as Finland
  - Stratified sampling would guarantee samples form each of ca. 50 countries
  - Stratified sampling weighted with population, large countries (e.g. Germany) would get more samples than small countries

Dimensionality Reduction

- Purpose:
  - avoid curse of dimensionality
  - reduce amount of time and memory required by data mining algorithms
  - allow data to be more easily visualized
  - may help to eliminate irrelevant features or reduce noise
  - may help to avoid stability problems
- Techniques
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Others: supervised and non-linear techniques

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

Prepared by: Er. Dinesh Baniya Kshatri 63

Feature Subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

Prepared by: Er. Dinesh Baniya Kshatri 64

Feature Subset Selection Techniques

- **Brute-force approach:**
  - Try all possible feature subsets as input to data mining algorithm
- **Embedded approaches:**
  - Feature selection occurs naturally as part of the data mining algorithm
- **Filter approaches (usually one pass through data):**
  - Features are selected before data mining algorithm is run
- **Wrapper approaches (usually many passes through data):**
  - Use the data mining algorithm as a black box to find best subset of attributes

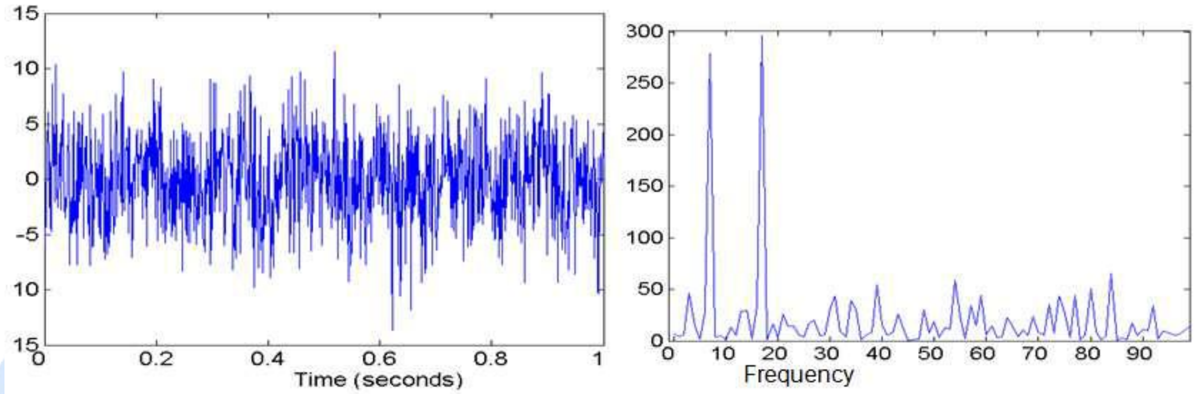
	The	Game	Play	Football	Baseball	Brady	Deflate	Gate
Document 1	12	2	3	14		4	4	6
Document 2	18	5	5		3			5
Document 3	24						4	5
Document 4	56	15					24	

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to a New Space

- Fourier transform
- Wavelet transform



Two Sine Waves + Noise   Frequency

Discretization

- Many data mining algorithms require the data to be discrete, often binary
- Discretization is the process of converting
  - Continuous-valued attributes, and
  - Ordinal attributes with high number of distinct values

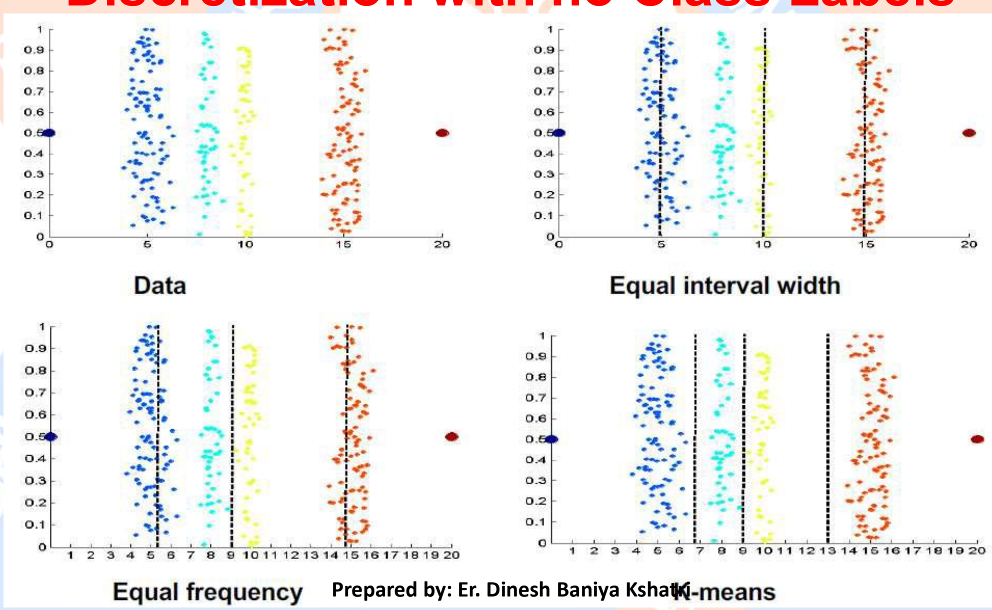
  into discrete variables with a small number of values
- Discretization is performed by
  - choosing one or more threshold values from the range of the attribute to create intervals of the original value range, and
  - then putting values inside each interval into a common bin
- Choosing the best number of bins is an open problem, typically trial and error process

Unsupervised Discretization

- Used in descriptive data mining tasks
- Discretization aims to produce equal-sized groups
  - Equal-width discretization: aims for close to same length intervals
  - Equal-frequency discretization: aims for close to same frequencies of values in each bin
  - K-means discretization: finds clusters of values and puts each cluster into a common bin

Unsupervised Discretization

Discretization with no Class Labels



Supervised Discretization
Information/Entropy

- Given probabilitites p_1, p_2, .., p_s whose sum is 1, ***Entropy*** is defined as:

$$\boxed{H(p_1, p_2, ..., p_s) = \sum_{i=1}^{s}(p_i log(1/p_i))}$$

- Entropy measures the amount of randomness or surprise or uncertainty.

- Only takes into account non-zero probabilities

Supervised Discretization
Entropy-Based Discretization

- Given a set of samples S, if S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2)$$

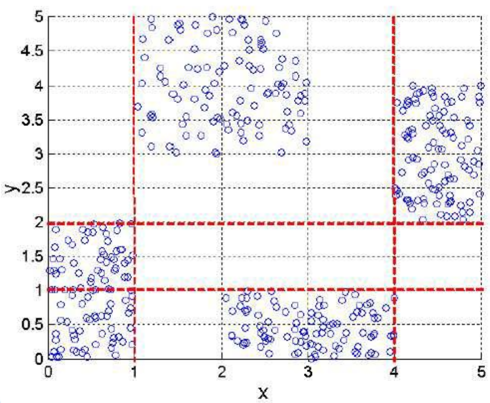
- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T,S) > \delta$$

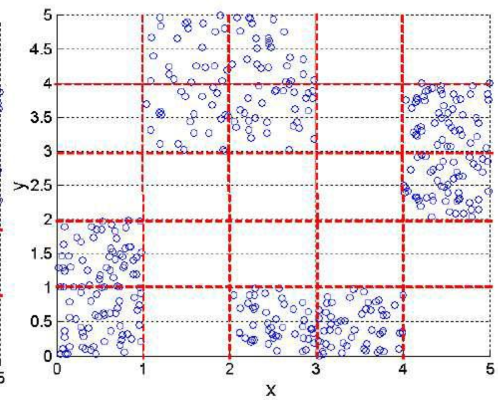
- Experiments show that it may reduce data size and improve classification accuracy

Supervised Discretization
Discretization Using Class Labels

- Entropy based approach



3 categories for both x and y



5 categories for both x and y

Binarization

- Many of the methods for finding frequent patterns rely on binary data
- For them we need to *binarize*
  - Attributes measured at ordinal, interval and ratio scales
    - this can be done via discretization methods by choosing the number of bins = 2
  - Multi-valued nominal (categorical) attributes
    - We create a separate binary attribute for each distinct value of the original attribute $x_{new}(i) = 1$ if and only if $x_{old} = i$

	x_{old}	$x_{new}(1)$	$x_{new}(2)$	$x_{new}(3)$
Helsinki	1	1	0	0
Tampere	2	0	1	0
Oulu	3	0	0	1

- Continuous attribute: first map the attribute to a categorical one
  - Example: height measured as {low, medium, high}
- Categorical attribute
  - Mapping to a set of binary attributes
  - Example: Low, medium, high as 1 0 0, 0 1 0, 0 0 1

Normalization

- It is a type of data transformation
  - The values of an attribute are scaled so as to fall within a small specified range, typically [-1,+1] or [0,+1]

- Min-max normalization: to [new_min_A, new_max_A]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  – Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600-12,000}{98,000-12,000}(1.0-0)+0=0.716$

Normalization …

- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  – Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600-54,000}{16,000}=1.225$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|v'|) < 1

Normalization …

Decimal Scaling Normalization

Suppose that the recorded values of F range from − 986 to 917. The maximum absolute value of F is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$) so that −986 normalizes to −0.986 and 917 normalizes to 0.917.

Prepared by: Er. Dinesh Baniya Kshatri 77

Variable Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$

- For large numbers, it may be advantageous to express them using log transformation
- For representing negative numbers as positive, the absolute value can be used

Prepared by: Er. Dinesh Baniya Kshatri 78