HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

582364 Data mining, 4 cu
Lecture 2: Data Preprocessing

Spring 2010

Lecturer: Juho Rousu

Teaching assistant: Taru Itäpelto

Data mining process

- Data mining process consists of several interdependent steps
  - Preprocessing to make the data suitable for analysis
  - Data mining to find the patterns/build models
  - Postprocessing to make the results suitable for human analysis
- In reality: iterative process with feedback loops and human interaction

Figure 1.1. The process of knowledge discovery in databases (KDD).

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

Types of Attribute Values: Levels of Measurement

- Nominal
  - E.g., profession, ID numbers, eye color, zip codes
  - Operators: distinctness (=,≠), set membership
  - Central tendency: mode - the most frequent value
  - Measure of dispersion:
    - e.g. entropy $-\Sigma_i p_i \log p_i$ (p_i is the relative frequency of value i
  - Transformation that does not change the meaning: any permutation of values
    - e.g. reassigning student ID's would not change the meaning

Types of Attribute Values: Levels of Measurement

- Ordinal
  - E.g., rankings (e.g., army ranks), grades, height in {tall, medium, short}
  - Operators: distinctness (=,≠), order (<,>)
  - Central tendency: median – the middle element
  - Measure of dispersion: percentile
    - p-th percentile is the value that is at or above p% of the data
    - Median is the 50% percentile
  - Transformation that does not change the meaning: any order preserving transformation
    - $new\_value = f(old\_value)$ where f is a monotonic function.

Types of Attribute Values: Levels of Measurement

- Interval
  - E.g., calendar dates, body temperatures
  - Operations: distinctness, order, addition (+,-)
  - Central tendency: (arithmetic) mean, i.e. average value
  - Measure of dispersion: standard deviation (σ) and variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$
  - Transformation that does not change the meaning:
    - *new_value* =*a* * *old_value* + *b,* where a and b are constants

Types of Attribute Values: Levels of Measurement

- Ratio
  - E.g., temperature in Kelvin, length, time, counts
  - Operations: distinctness, order, addition (+,-), multiplication (*,/)
  - Central tendency: geometric mean
  - Measure of dispersion: coefficient of variation
    - CV = σ/μ
  - Transformation that does not change the meaning:
    - $new\_value = a * old\_value$

Types of Attribute Values: Discrete and Continuous Attributes

- Independently from the measurement scales, attributes can be characterized by the sets of possible values they take
- 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
  - Both ordinal and nominal attributes are discrete
  - In computer memory, discrete values are typically represented by integers
  - 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.

Asymmetric attributes

- In some data, only a small fraction of attributes have non-zero value
  - E.g. Items in customers shopping basket, as compared to all items in the supermarket
- Comparison of customers based on items they did not buy is not meaningful
  - we would get close to 100% precent similarity for most customers
- Analysis of the items they did buy may reveal much more
  - Frequent pattern discovery is based on this premise

Data quality and cleaning

- 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

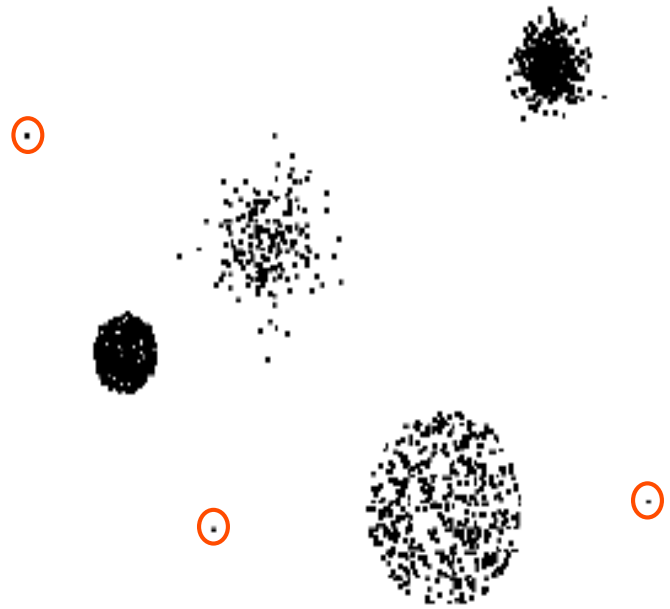
- The process of tackling the quality is often called *data cleaning*

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

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set
- Unlike noise, outliers can contain interesting information
- Deciding whether the outlier is caused by an error or is correct, generally requires a human expert
- In anomaly detection tasks (e.g industrial process monitoring), the *goal* is to detect an outlier



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**

Handling Missing values by 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**

Handling Missing values by 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 by 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**

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

Data Preprocessing

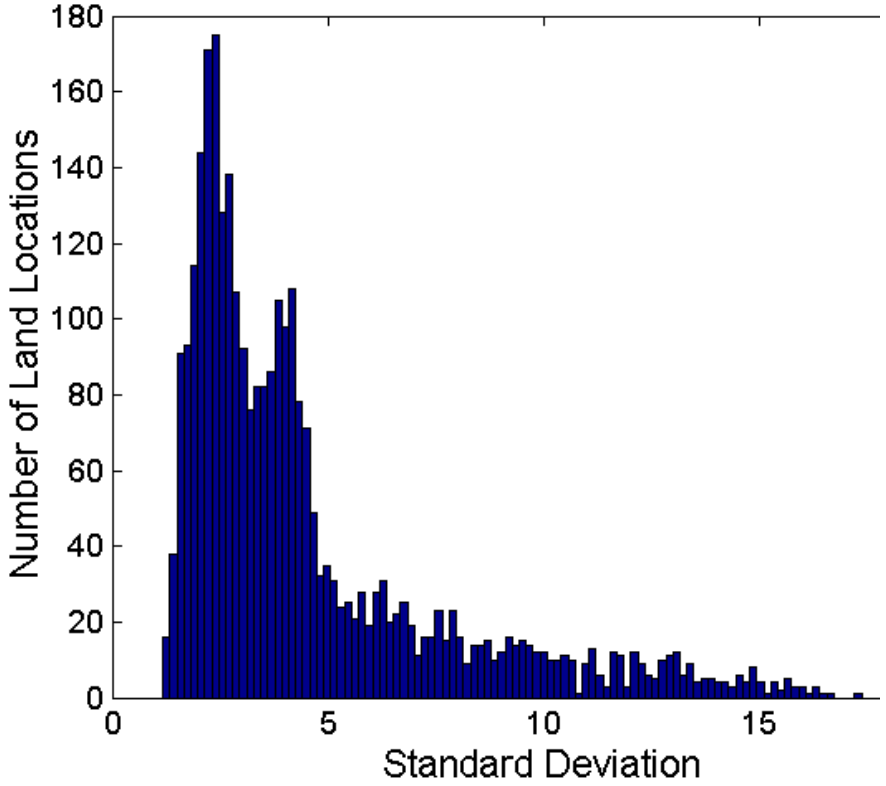
- After addressing the data quality by cleaning the data, it may still need further processing before it can be fed into a data mining algorithm
- Most important steps for frequent pattern discovery include
  - Aggregation
  - Sampling
  - Discretization and Binarization
  - Attribute Transformation
- Other preprocessing tasks, important in predictive data mining and clustering: dimensionality reduction, feature subset selection

Aggregation

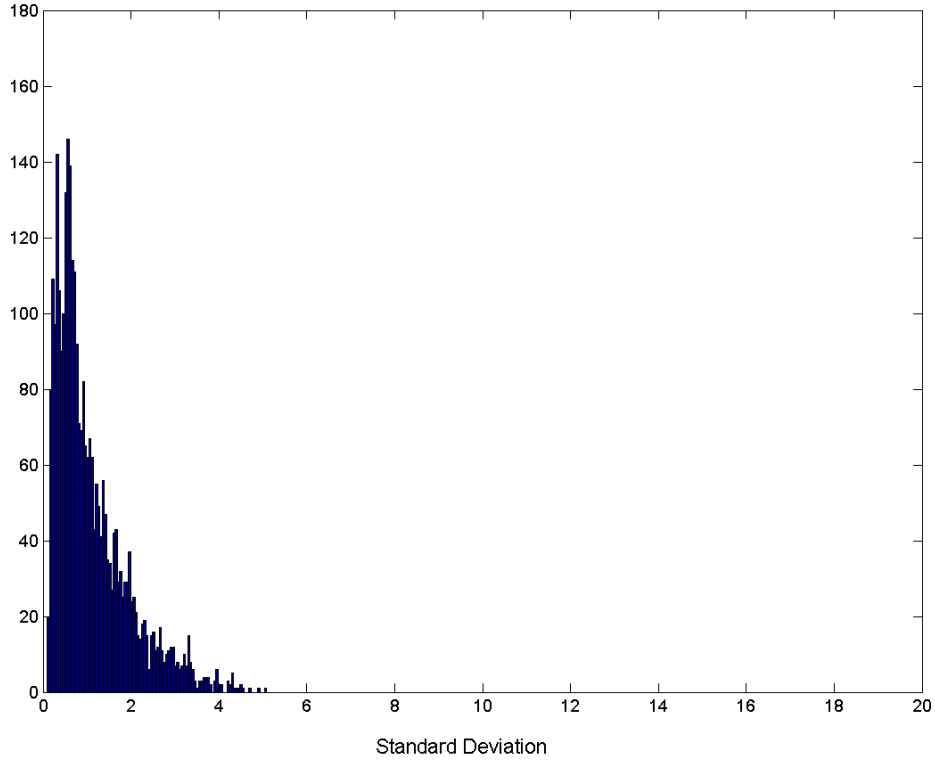
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
    - Faster to process, easier to fit to computer main memory
  - Change of scale
    - E.g. Cities aggregated into regions, states, countries,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

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

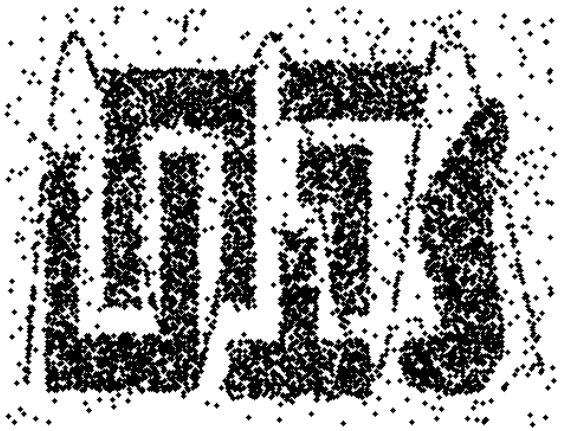
- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Reasons to use sampling
  - In statistics: *obtaining* the entire set of data of interest is too expensive or time consuming.
  - In data mining: *processing* the entire set of data of interest is too expensive or time consuming
- 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

Sampling

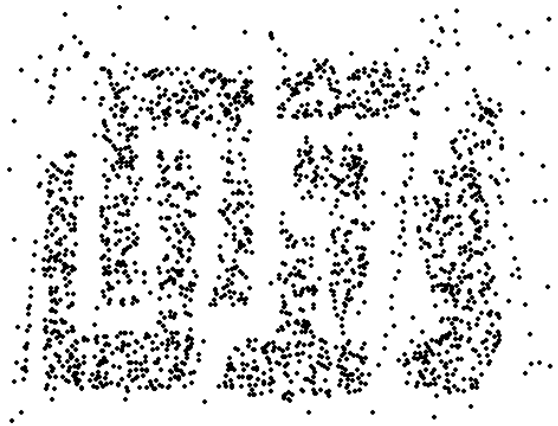
- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample.
- Simple random sampling does not work well with data that has many groups
  - Some groups many not get fair representation in the sample

Choosing the Sample Size

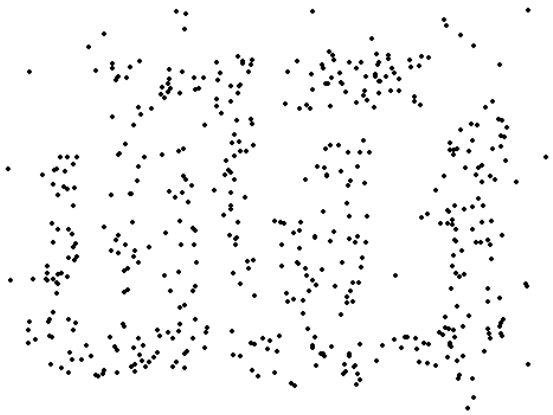
- It is important to choose a sample size that is
  - Large enough to enable to recover the structure in the original data (i.e. has approximately the same property than the orginal data)
  - Small enough to give as savings in processing time and space



8000 points



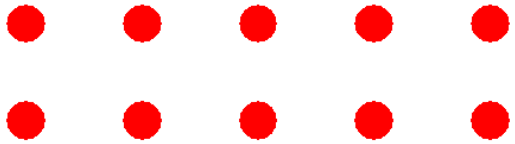
2000 Points

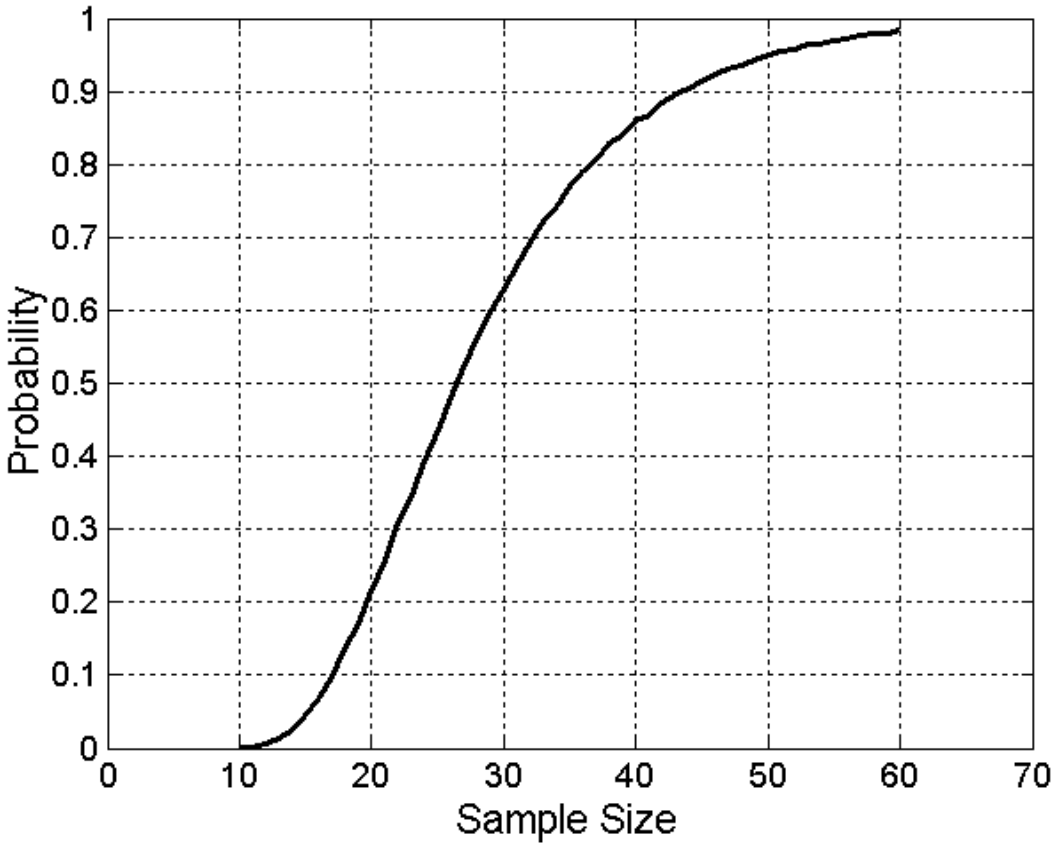


500 Points

Example: Representative Sample Size

- What sample size is necessary to get at least one object from each of 10 groups in random sampling





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

Discretization

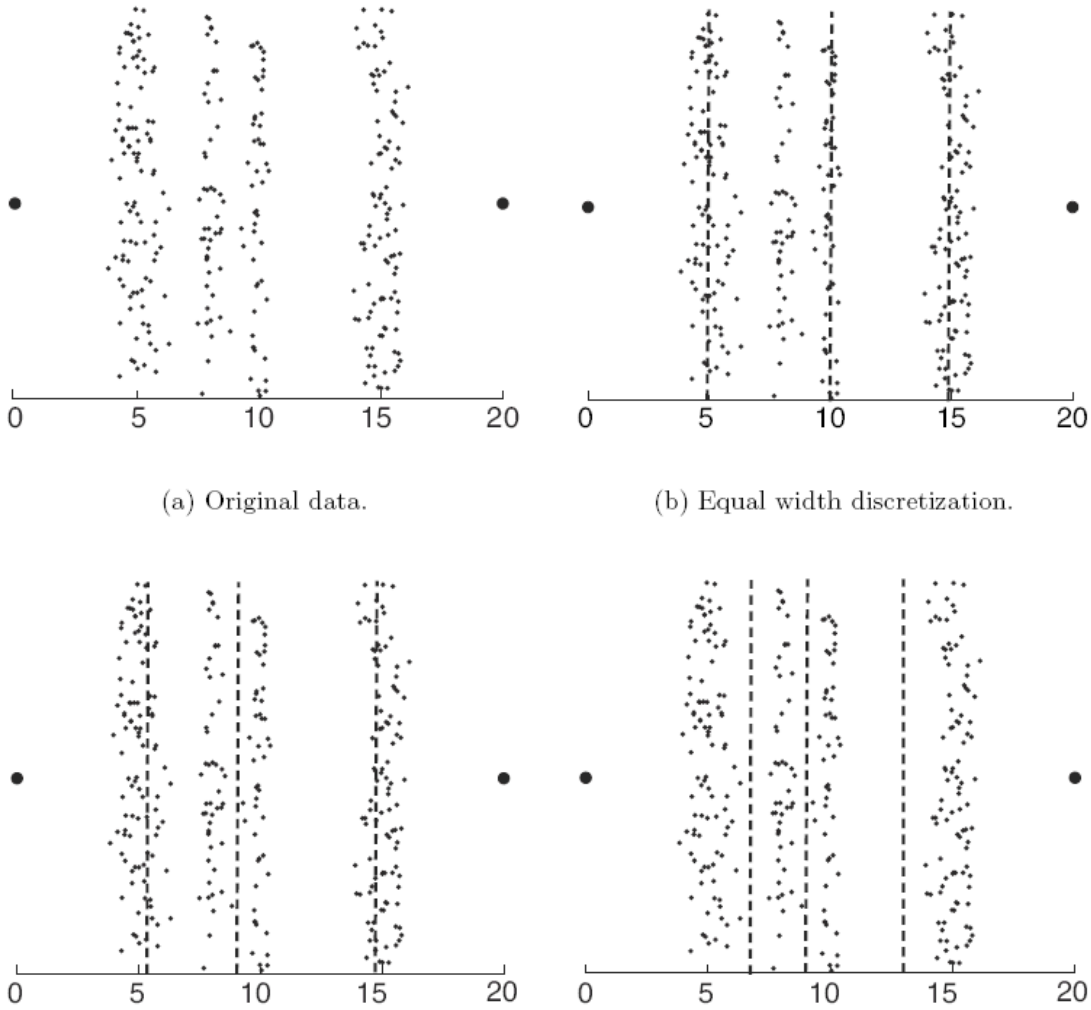
- Many data mining algorithms require the data to be discrete, often binary
- Discretization is the process of converting
  - Continuous-valued attributes, and
  - Ordinal attributes with high number of distinct values

  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



(a) Original data.

(b) Equal width discretization.

(c) Equal frequency discretization.

(d) K-means discretization.

Figure 2.13. Different discretization techniques.

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 *xnew_i* = *1* if and only if *xold* = *i*

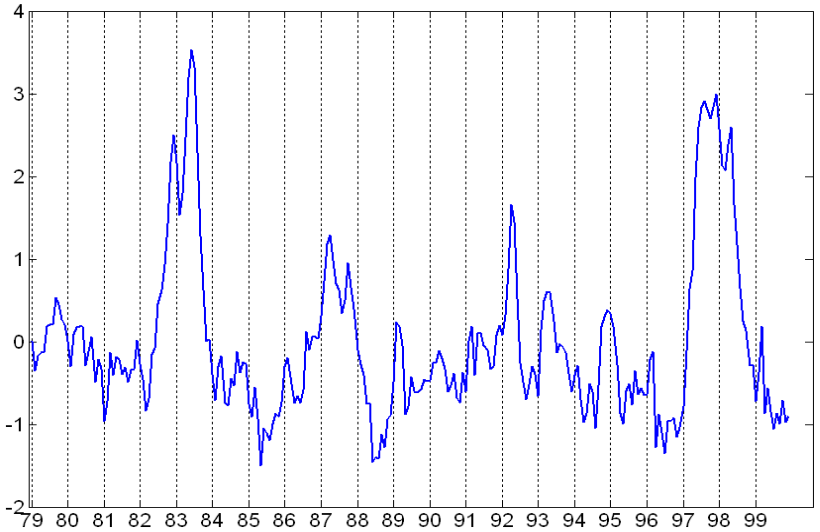
Binarization

- 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

Attribute Transformation

- 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|
  - Standardization and Normalization

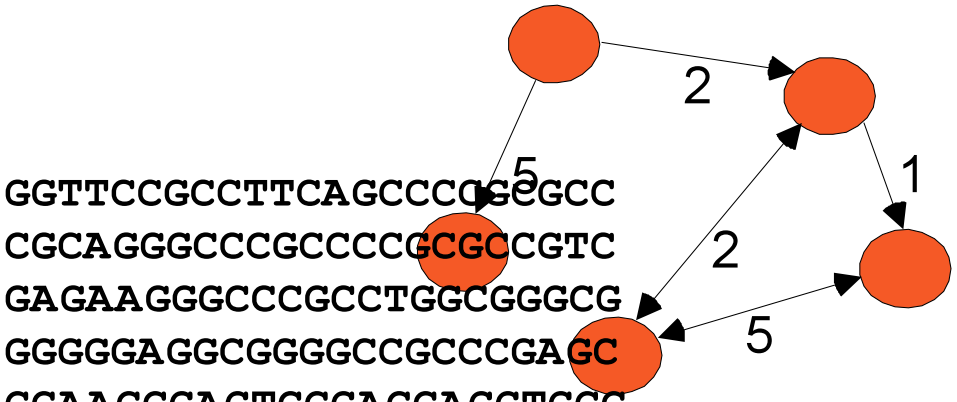


Normalization/standardization

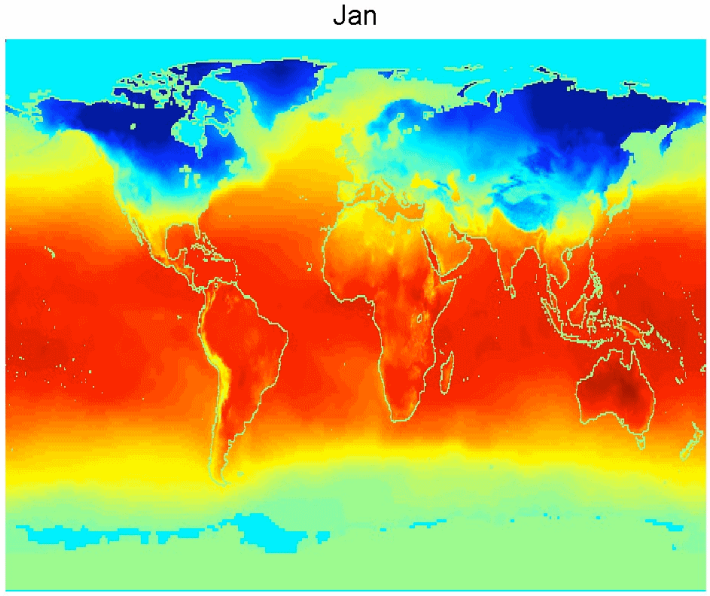
- In many data mining tasks, variables that have vastly different scales from each other may cause problems
  - e.g. one variable taking values in [0,1000] and another in [0,0.001]
- Normalization (in correct statistics terminology: standardization) is the process of converting the attributes to zero-mean, unit-variance attributes
  - $X_{new} = (x_{old} - x_{mean})/\sigma$

Types of data sets

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



GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCG
CCCTCTG
GCTCATT
GCCAAGT
TGGGCTG



Handling non-record data

- Most data mining algorithms have ben designed for record type data
- However, many times the data is in non-record format:
  - Text and Images are prime example
- A general approach to handle non-record data is to transform them to reacord format by computing features (attributes) of the data

Example: Text

- Text documents are frequently transformed into so called “bag of words” representation
- Document is represented by a record where there is a attribute for each possible word
- The attribute value is either the count of the words in the document or binary value (word occurs/does not occur)

	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: images

- Simple approach used in image processing is to use color histograms
- The number of pixels of certain color is one attribute
  - Colors can be discretized
- Color histogram is resistant to rotation and translation of the image