Data Warehouse and Data Mining

Dhruv Gupta

dhruv.gupta@ntnu.no

31-January-2023

NTNU | Norwegian University of Science and Technology

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Data
   - Data Analysis Perspectives
   - Data Quality
   - Data Preprocessing
     - Aggregation
     - Sampling
     - Curse of Dimensionality
     - Feature Selection
     - Feature Creation
     - Discretization and Binarization
     - Attribute Transformations
     - Similarity and Dissimilarity
     - Correlation

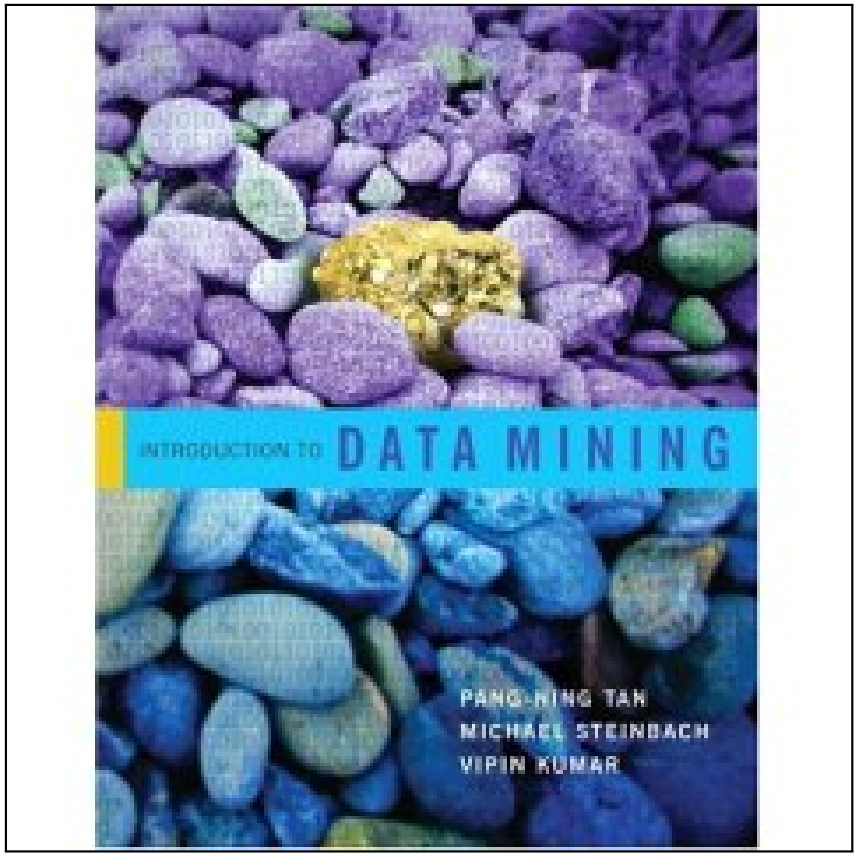
3. Summary

Administrative

1. First Assignment
   - Available and due by 09.February.2023.
2. Volunteers for feedback regarding course
   - Interested? Please contact me by email!

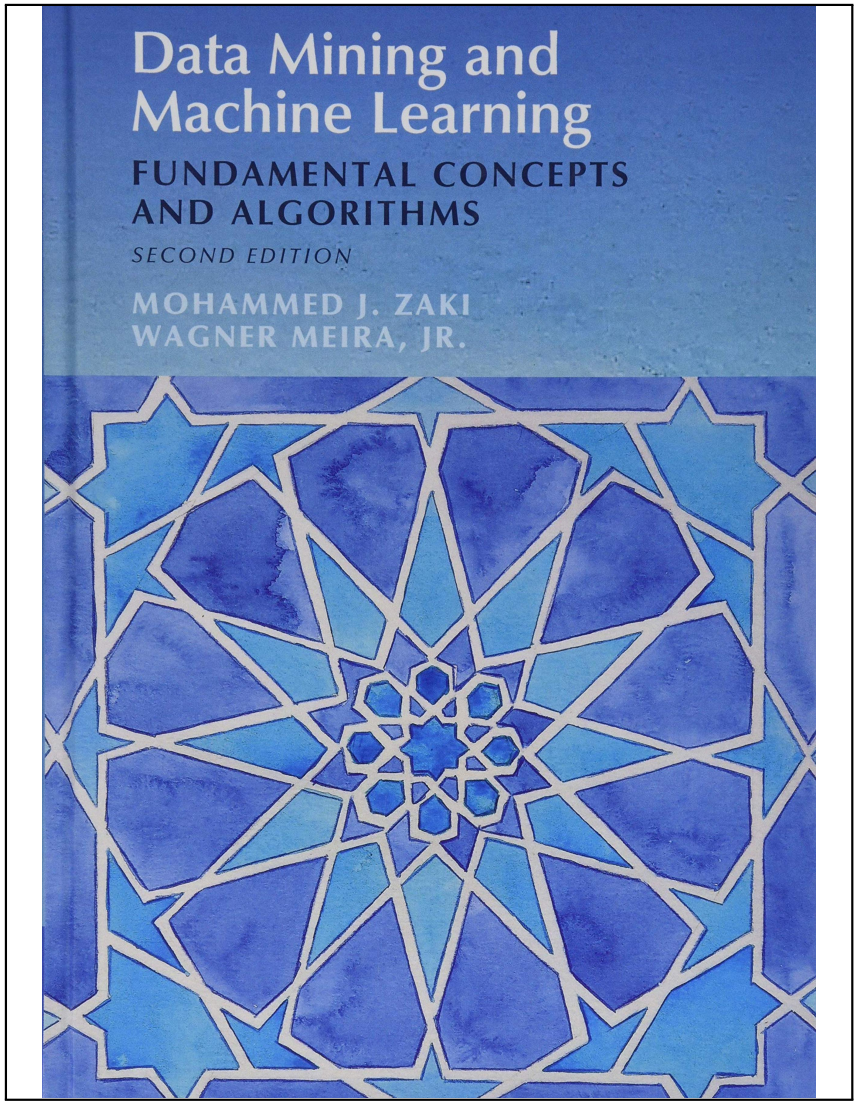
References for "Data"

1. Book: Tan et al. *"Introduction to Data Mining"*, 1st Edition, 2006, Pearson Education Inc.
2. Text and images for majority of slides in "Data" subsection are based on the book by Tan et al.



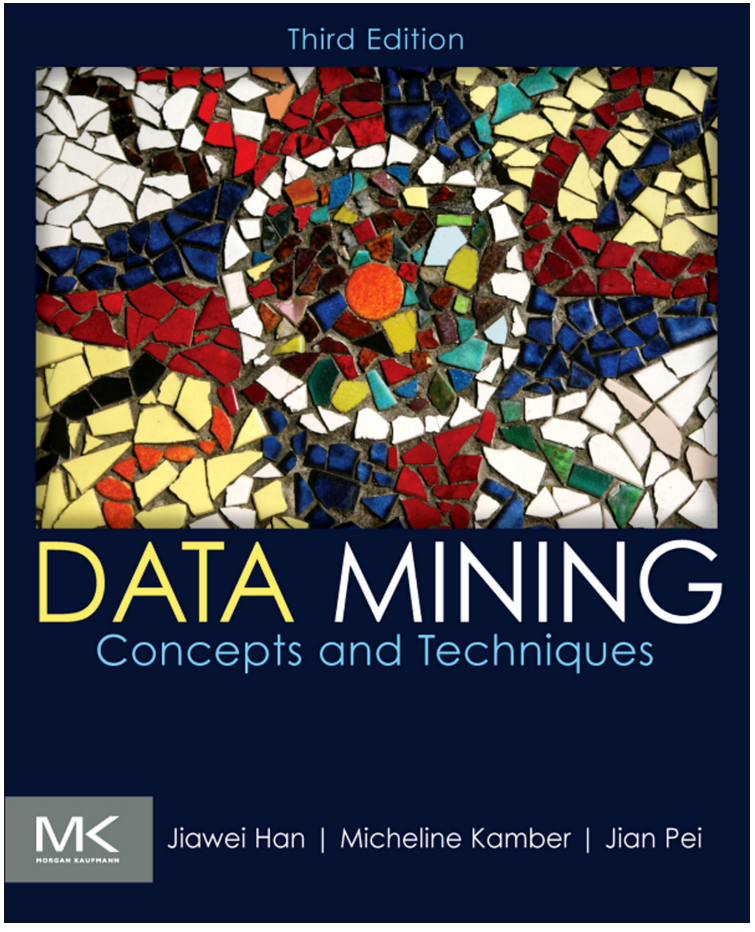
References for "Data"

1. Book: Zaki and Meira. *"Data Mining and Machine Learning: Fundamental Concepts and Algorithms"*, 2nd Edition, 2020, Cambridge University Press.
2. Text and images for some slides in "Data" subsection are based on the book by Zaki and Meira.



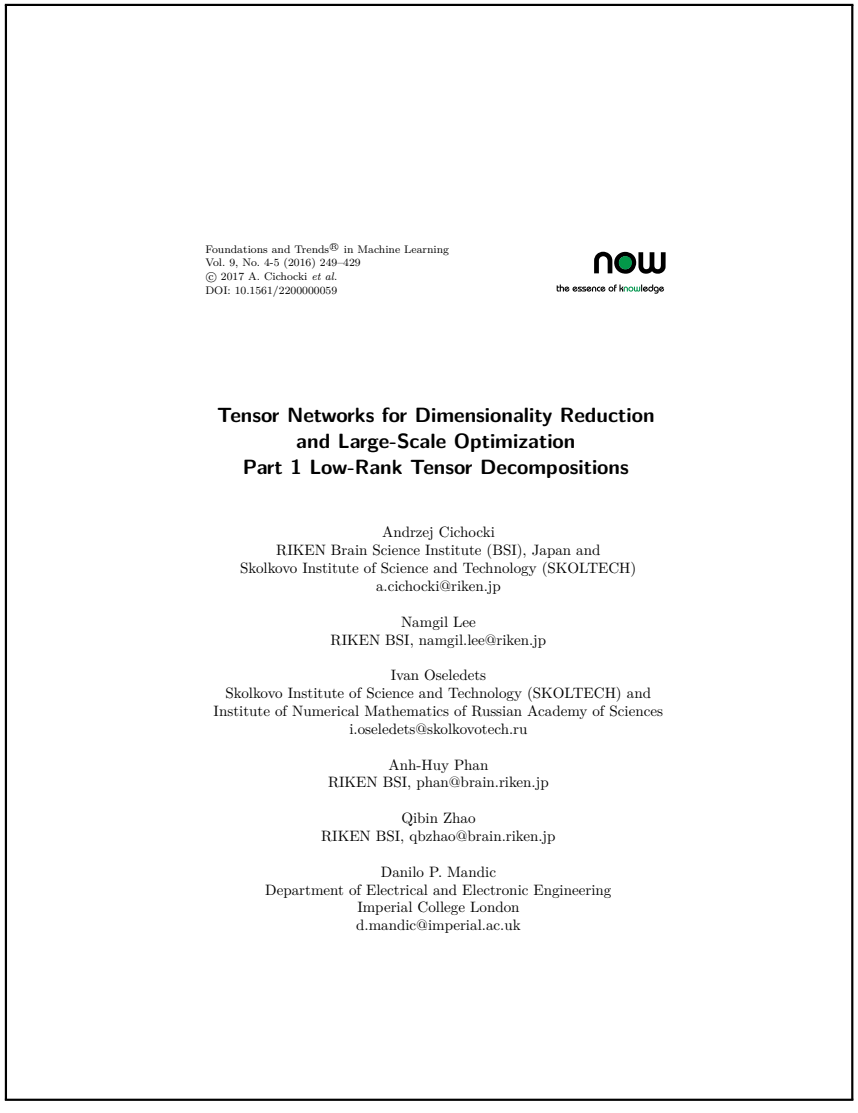
References for "Data"

1. Book: Han et al. *"Data Mining Concepts and Techniques"*, 3rd Edition, 2012, Morgan Kaufmann Publishers.
2. All text and images for some slides in "Data" are based on the book by Han et al.



References for "Data"

1. Paper: Cichocki et al. "Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Part 1 Low-Rank Tensor Decompositions", Vol. 9 No. 4–5 (2016), Foundations and Trends in Machine Learning.
2. Text for some slides in "Data" subsection are based on the book by Zaki and Meira.



Foundations and Trends® in Machine Learning
Vol. 9, No. 4-5 (2016) 249–429

DOI: 10.1561/2200000059

now
the essence of knowledge

Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Part 1 Low-Rank Tensor Decompositions

Andrzej Cichocki
RIKEN Brain Science Institute (BSI), Japan and
Skolkovo Institute of Science and Technology (SKOLTECH)
a.cichocki@riken.jp

Namgil Lee
RIKEN BSI, namgil.lee@riken.jp

Ivan Oseledets
Skolkovo Institute of Science and Technology (SKOLTECH) and
Institute of Numerical Mathematics of Russian Academy of Sciences
i.oseledets@skolkovotech.ru

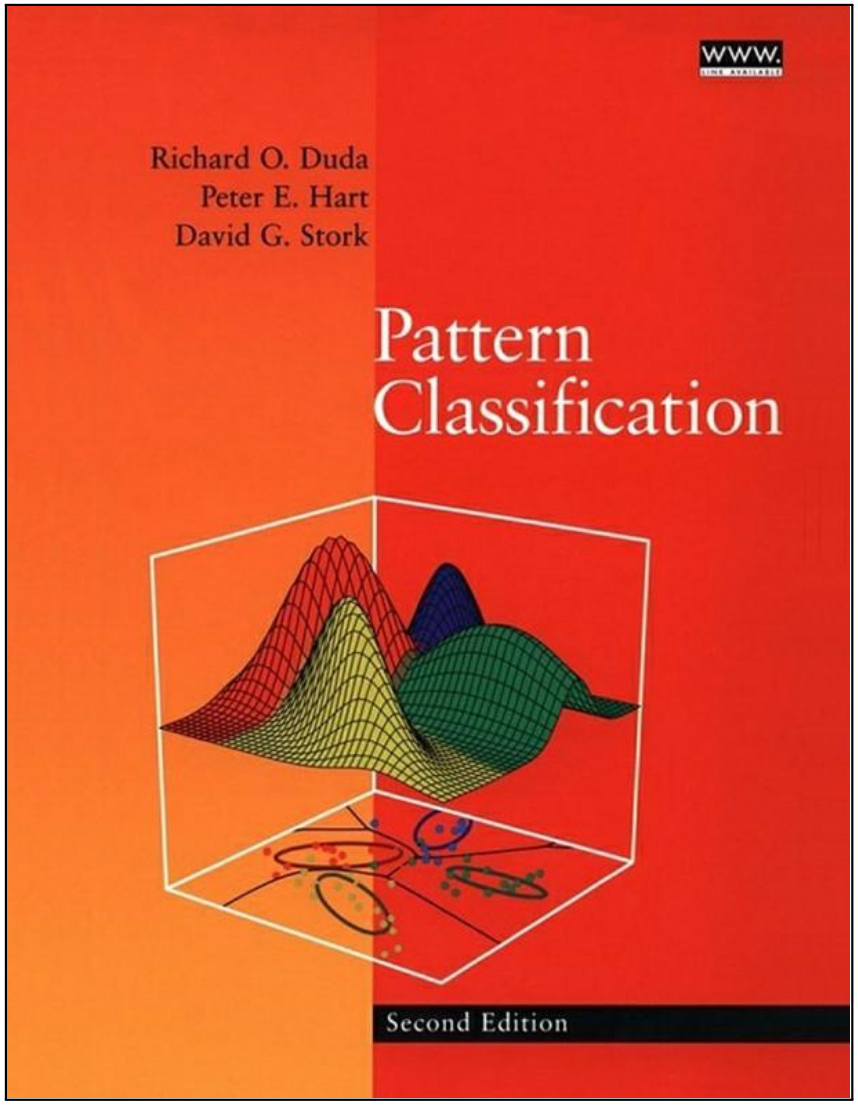
Anh-Huy Phan
RIKEN BSI, phan@brain.riken.jp

Qibin Zhao
RIKEN BSI, qbzhao@brain.riken.jp

Danilo P. Mandic
Department of Electrical and Electronic Engineering
Imperial College London
d.mandic@imperial.ac.uk

References for "Data"

1. Book: Duda et al. *"Pattern Classification"*, 2nd Edition, 2001, John Wiley & Sons, Inc.
2. Text and images for some slides in "Data" subsection are based on the book by Zaki and Meira.



1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Data
   - Data Analysis Perspectives
   - Data Quality
   - Data Preprocessing
     - Aggregation
     - Sampling
     - Curse of Dimensionality
     - Feature Selection
     - Feature Creation
     - Discretization and Binarization
     - Attribute Transformations
     - Similarity and Dissimilarity
     - Correlation

3. Summary

Data Set Types

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

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Based on Measurements

	Type	Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$.	ZIP codes, employee ID numbers, eye color, gender.	Mode, entropy, contingency correlation, χ^2 test.
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(>, <)$.	Hardness of minerals, {good,better,best}, grades, street numbers.	Median, percentiles, rank correlation, run tests, sign tests.
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$.	Calendar dates, temperature in Celsius or Fahrenheit.	Mean, standard deviation, Pearson's correlation, t and F tests.
	Ratio	For ratio variables, both differences and ratios are meaningful. $(\times, \div)$.	Temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current.	Geometric mean, harmonic mean, percent variation.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Attribute Types — Time Example

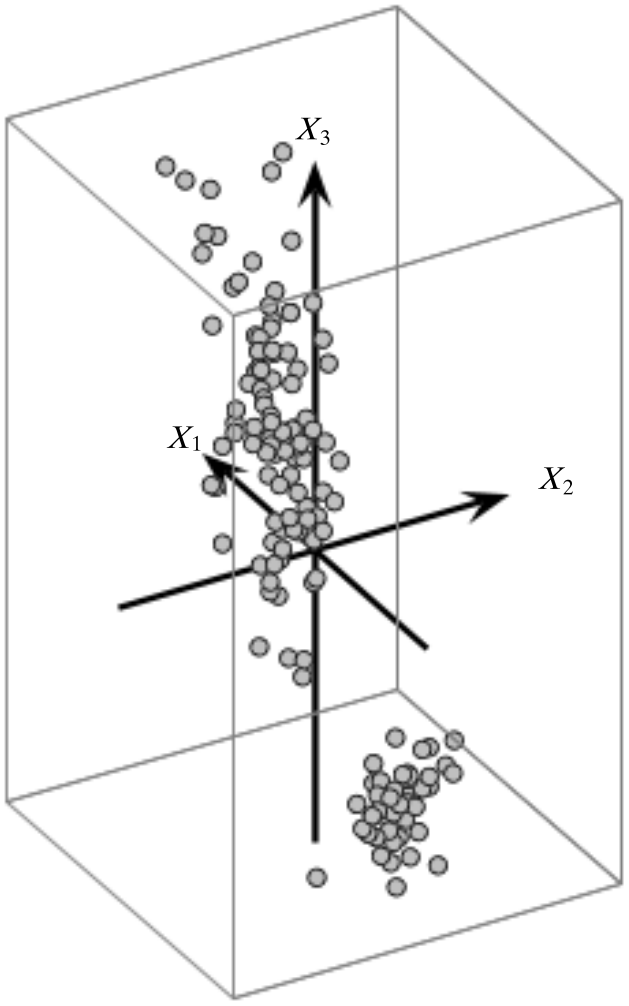
1. Years (e.g., 2014, 2015, 2016 . . .).
   - What type of attribute?
2. Years or time is computationally recorded as UNIX epochs (i.e., number of milliseconds elapsed since 01-January-1970).
   - Example: 1612813881000 milliseconds ≡ Monday, February 8, 2021 7:51:21 PM
   - What type of attribute is this?
3. Consider two timestamps recorded as UNIX epochs t_1 and t_2, where $t_2 \geq t_1$. Consider a column consisting of values that contain: $t_2 - t_1$.
   - What type of attribute is this?

Data Analysis Perspectives — Database Perspective

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	**Sepal width** X_2	**Petal length** X_3	**Petal width** X_4	**Class** X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Data Analysis Perspectives — Geometric and Algebraic Perspective



(a) Original Basis

Data Analysis Perspectives — Probabilistic Perspective

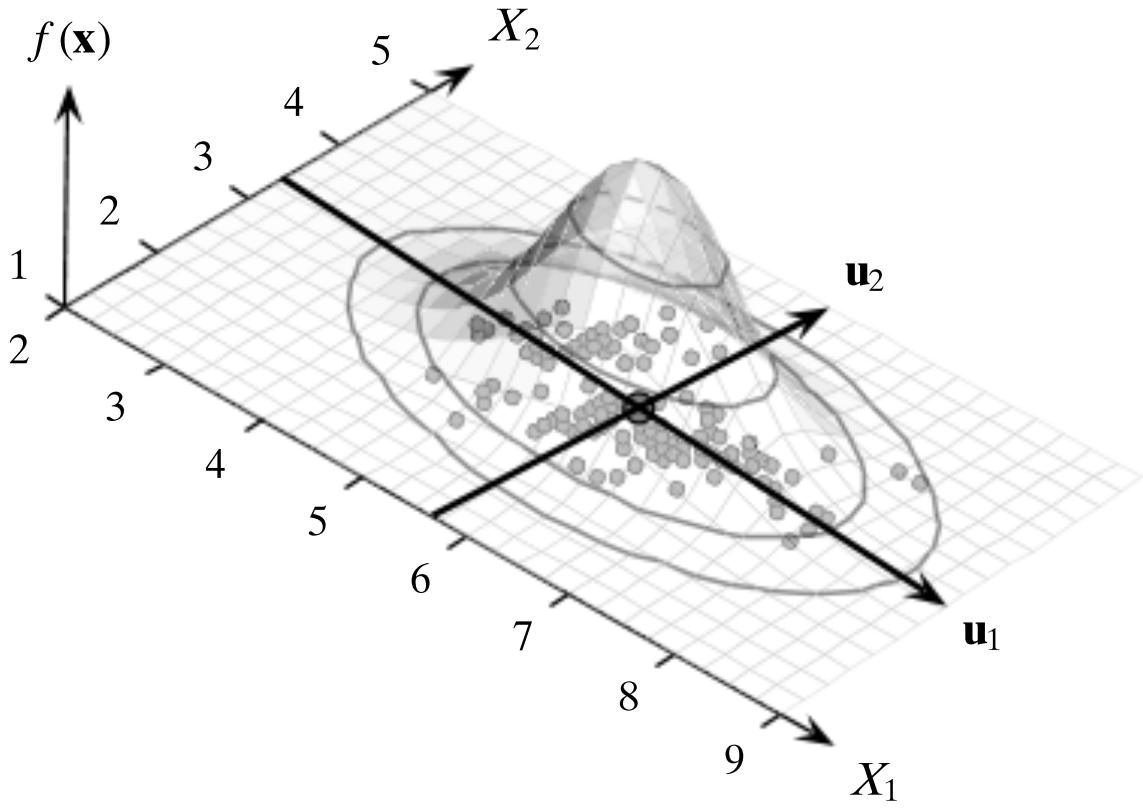


Figure 2.8. Iris: `sepal length` and `sepal width` , bivariate normal density and contours.

Data Quality

- Poor data quality negatively affects many data processing efforts.
- Data mining example:
  - A classification model for detecting people who are loan risks is built using poor data.
  - Some credit-worthy candidates are denied loans.
  - More loans are given to individuals that default.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

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
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Quality Issues — Noise

- Noise is the random component of a measurement error.
- It may involve the distortion of a value or the addition of spurious objects.
- The term noise is often used in connection with data that has a spatial or temporal component.
- In such cases, techniques from signal or image processing can frequently be used to reduce noise and thus, help to discover patterns (signals) that might be "lost in the noise."
- Robust algorithms: data mining methods that produce acceptable results even when noise is present.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Quality Issues — Noise

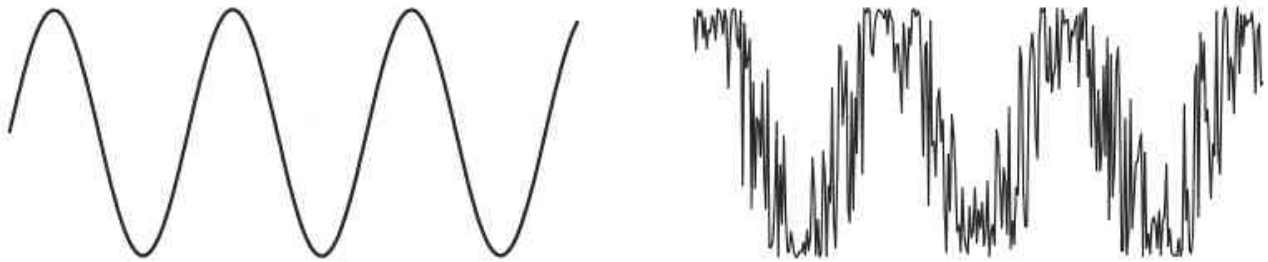


Figure 2.5. Noise in a time series context.

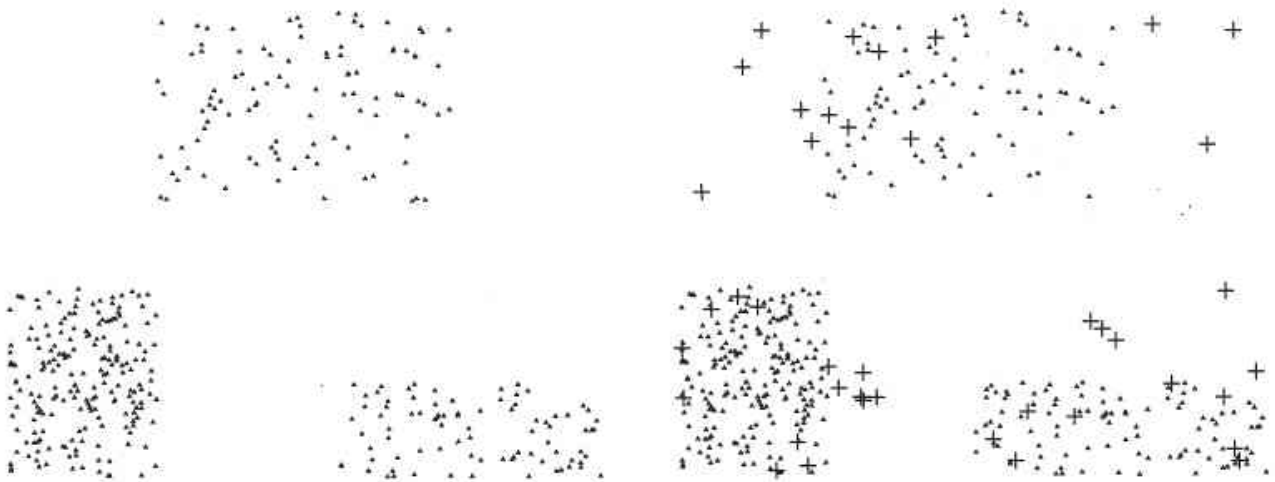
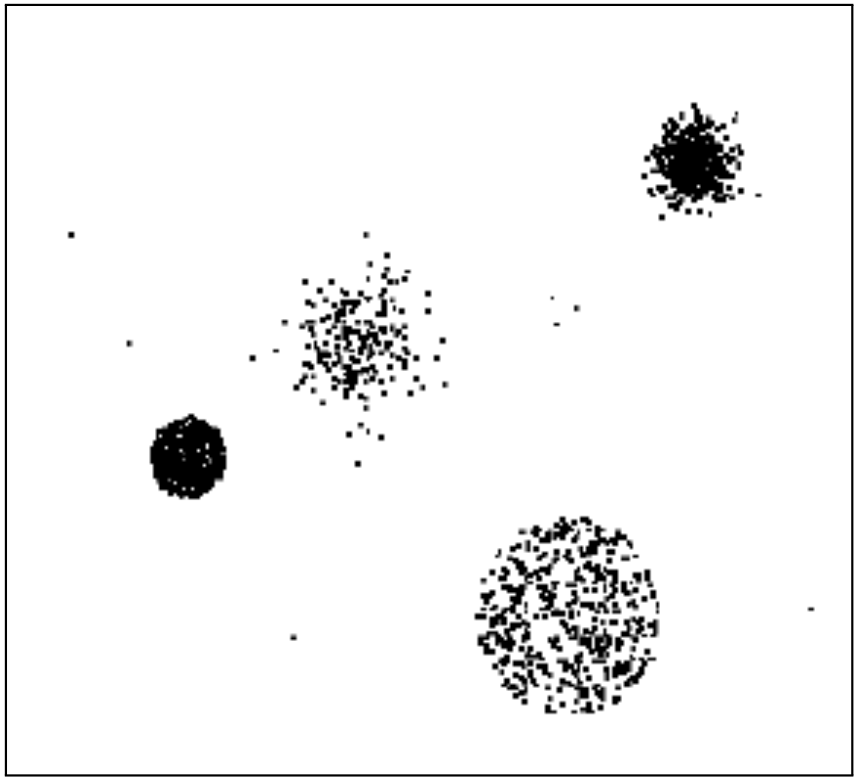


Figure 2.6. Noise in spatial context.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Quality Issues — Outliers

- Outliers are:
  1. Data objects that have characteristics different from most other objects in the data set.
  2. Values of an attribute that are unusual with respect to the typical values for that attribute.
- Outliers vs. Noise: Outliers can be legitimate data objects or values and may sometimes be of interest.



All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Quality Issues — 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 or variables.
    - Estimate missing values (e.g., time series of temperature or census results).
    - Ignore the missing value during analysis.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Quality Issues — Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another.
  - Major issue when merging data from heterogeneous sources.
- Example: same person with multiple email addresses.
- Data Cleaning: process of dealing with duplicate data issues.
- When should duplicate data not be removed?

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Data Preprocessing

- Data Preprocessing: processing the data so as to improve the data mining analysis with respect to time, cost, and quality.
- This done by: selecting data objects and attributes for the analysis or creating/changing the attributes.
- Typical data preprocessing steps:
  - Aggregation
  - Sampling
  - Discretization and Binarization
  - Attribute Transformation
  - Dimensionality Reduction
  - Feature Subset Selection
  - Feature Creation

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object).
- Purpose:
  - Data Reduction: reduce the number of attributes or objects.
  - Change of scale:
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years.
  - More stable data: aggregated data tends to have less variability.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Sampling

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

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Sampling

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative.
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data.
- Sampling Methods:
  - Simple Random Sampling.
  - Stratified Sampling.
  - Progressive Sampling.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Sampling — Simple Random Sampling (SRS)

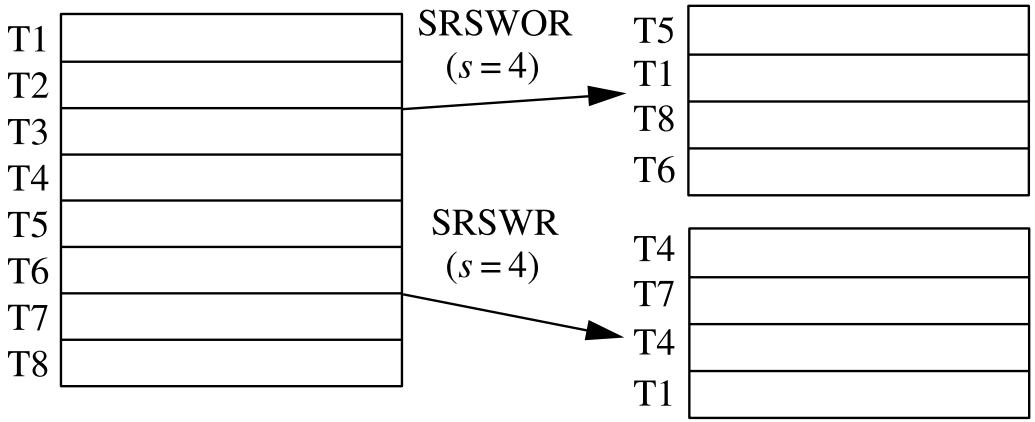
- For this type of sampling, there is an equal probability of selecting any particular item.
- Two variations on SRS:
  1. Sampling without replacement as each item is selected, it is removed from the set of all objects that together constitute the population.
  2. Sampling with replacement objects are not removed from the population as they are selected for the sample.
- In sampling with replacement, the same object can be picked more than once.
- The samples produced by the two methods are not much different when samples are relatively small compared to the data set size, but sampling with replacement is simpler to analyze since the probability of selecting any object remains constant during the sampling process.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

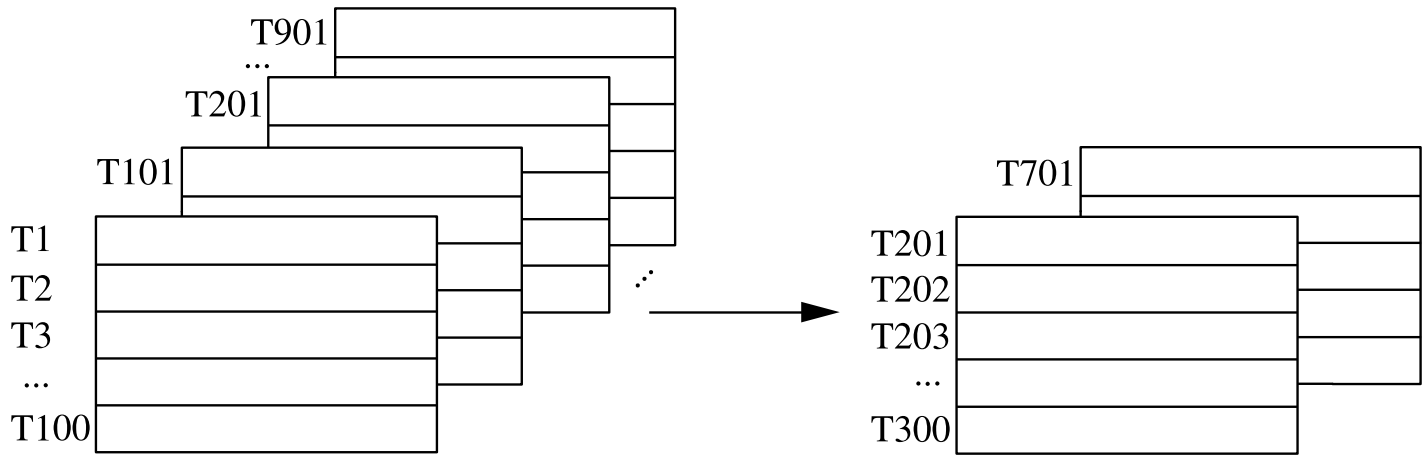
Sampling — Stratified Sampling

- Problems with SRS:
  - When the population consists of different classes of objects, with widely different numbers of objects per class. Then simple random sampling can fail to adequately represent those types of objects that are less frequent.
  - This can cause problems when the analysis requires proper representation of all object classes.
  - For example, when building classification models for rare classes, it is critical that the rare classes be adequately represented in the sample.
- Hence, a sampling scheme that can accommodate differing frequencies for the items of interest is needed.
- Stratified Sampling: which starts with pre-specified groups of objects, is such an approach.
- In the simplest version, equal numbers of objects are drawn from each group even though the groups are of different sizes. In another variation, the number of objects drawn from each group is proportional to the size of that group.

Sampling — Examples



Cluster sample
$(s = 2)$



All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Sampling — Examples

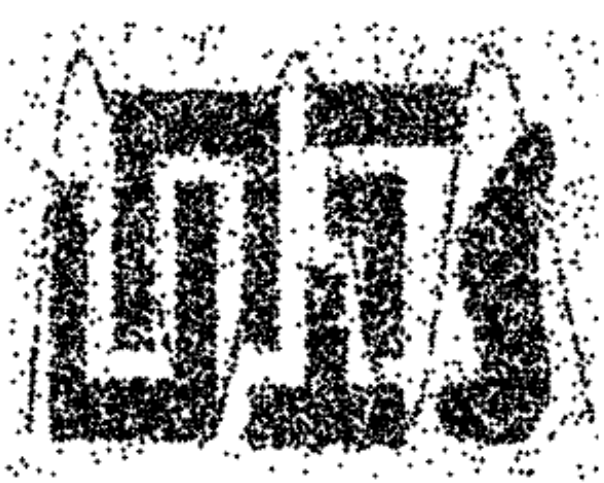
Startified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

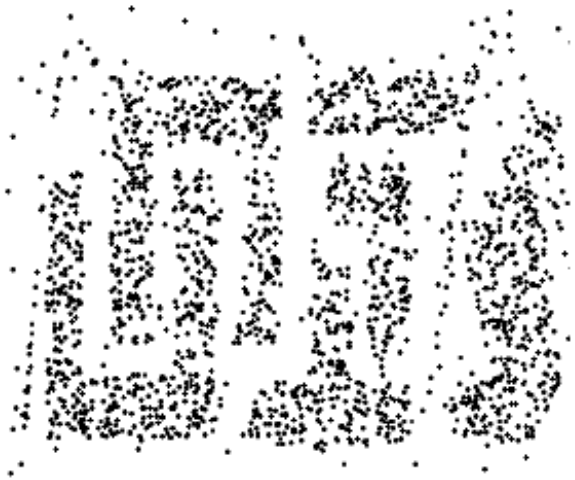
T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

All from: Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Sampling — Sample Size



8000 points



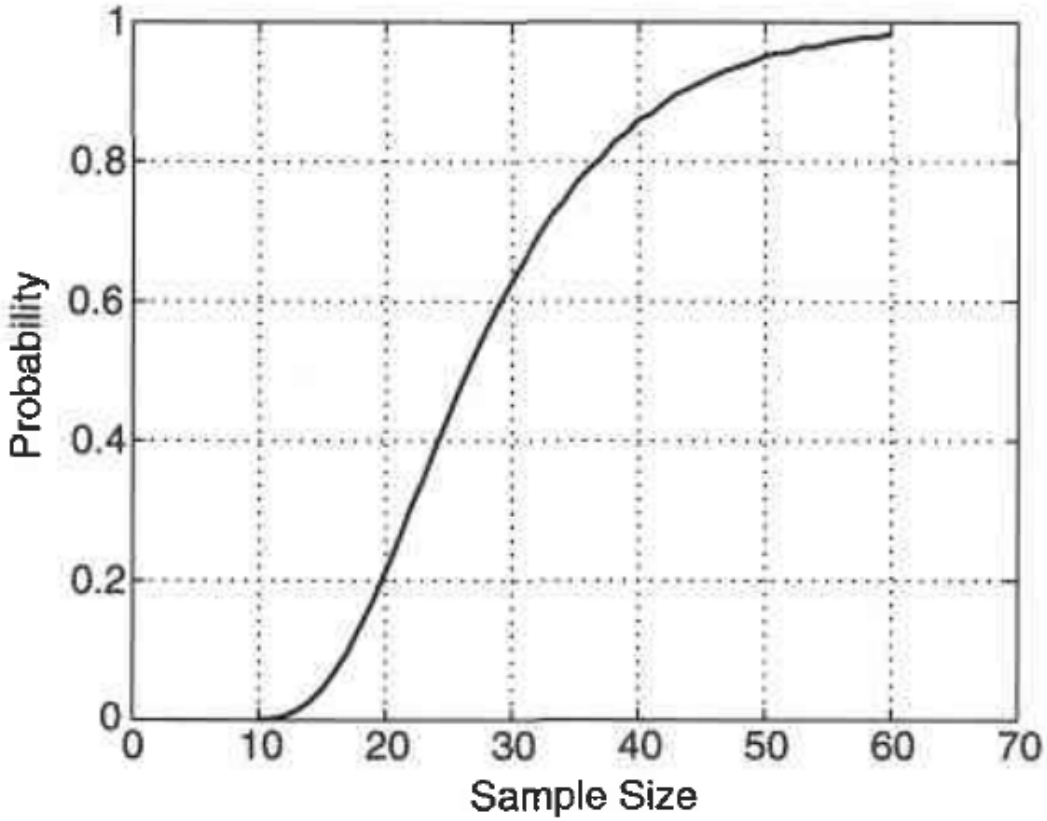
2000 Points

500 Points

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Sampling — Sample Size

(a) Ten groups of points.



(b) Probability a sample contains points from each of 10 groups.

Figure 2.10. Finding representative points from 10 groups.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Sampling — Progressive Sampling

- Progressive Sampling Schemes: these approaches start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.
- Eliminates the need to determine the correct sample size initially.
- However, it requires that there be a way to evaluate the sample to judge if it is large enough.
- This can be done with respect to the application at hand.
- Example: a progressive sampling is used to learn a predictive model.
- The accuracy of predictive models increases as the sample size increases up to some point.
- After which the increase in accuracy levels off.
- We want to stop increasing the sample size at this leveling-off point

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Curse of Dimensionality

The term curse of dimensionality was coined by Bellman (1961) to indicate that the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with the number of variables, that is, with the dimensionality of the function.

All from: Cichocki et al. "Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Part 1 Low-Rank Tensor Decompositions", Vol. 9 No. 4–5 (2016), Foundations and Trends in Machine Learning.

Curse of Dimensionality

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	**Sepal width** X_2	**Petal length** X_3	**Petal width** X_4	**Class** X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

Curse of Dimensionality

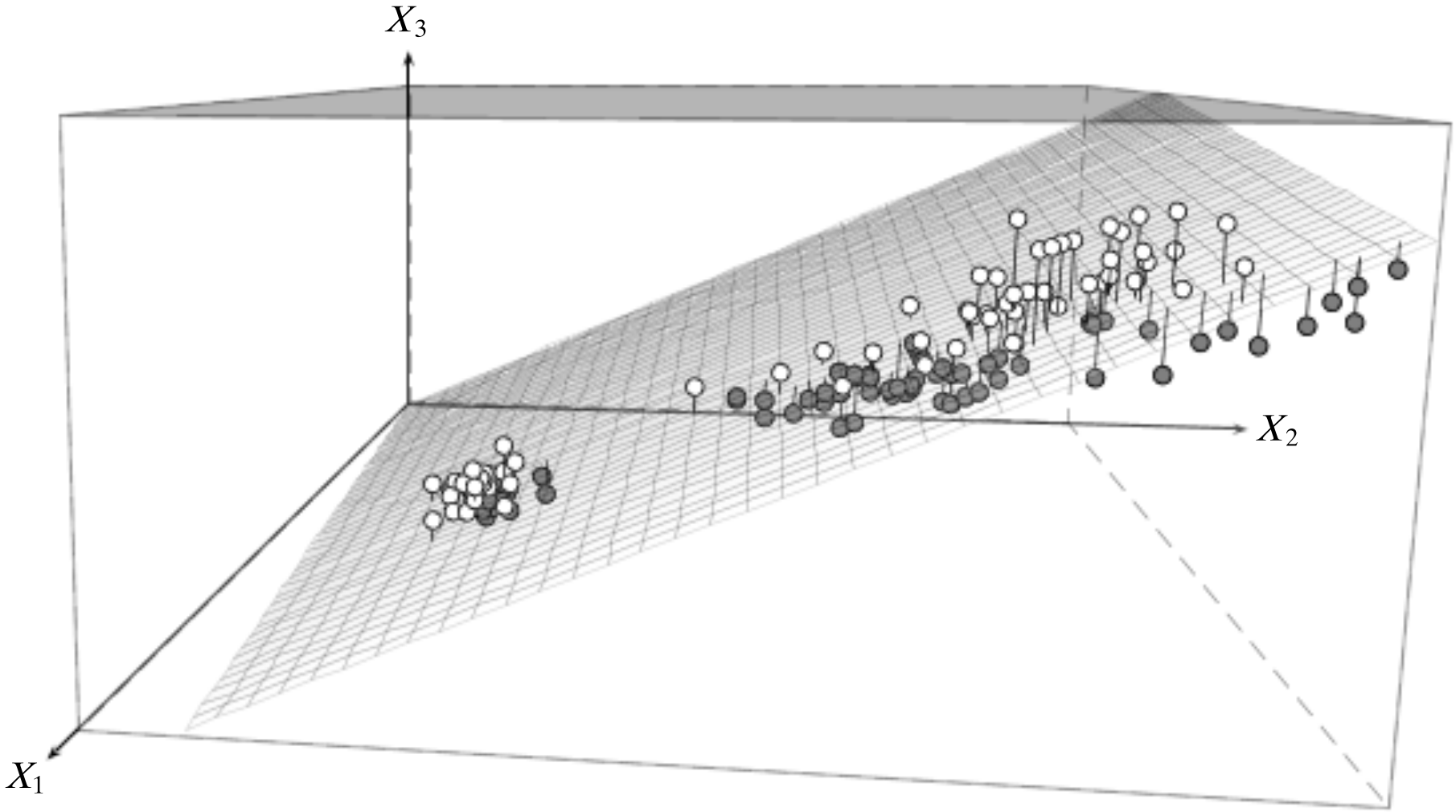


Figure 6.2. 3-dimensional Iris data hyperspace with a 2-dimensional hyperplane. Points in white are above the plane, whereas points in gray are below the plane.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

Curse of Dimensionality

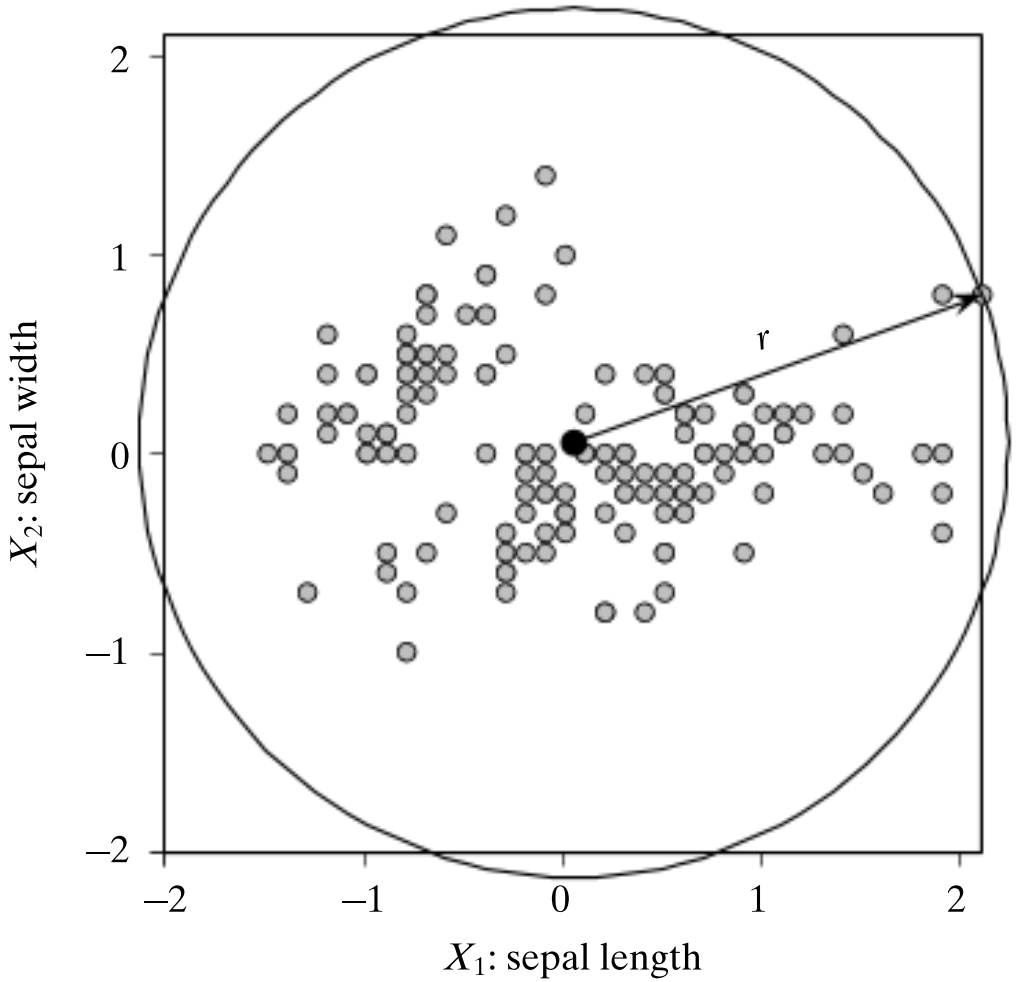


Figure 6.1. Iris data hyperspace: hypercube (solid; with $l = 4.12$) and hypersphere (dashed; with $r = 2.19$).

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

Curse of Dimensionality

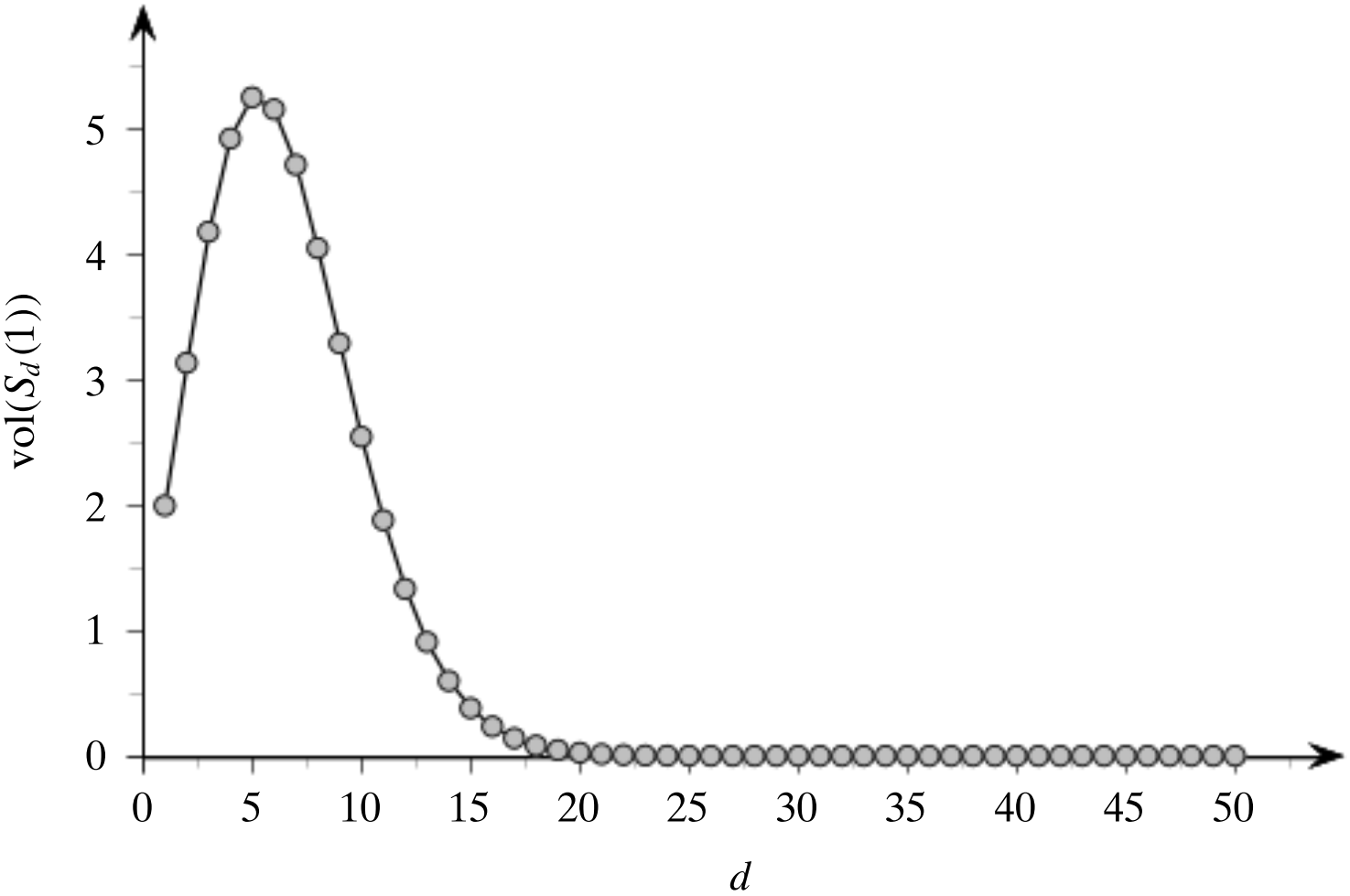


Figure 6.3. Volume of a unit hypersphere.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

Curse of Dimensionality

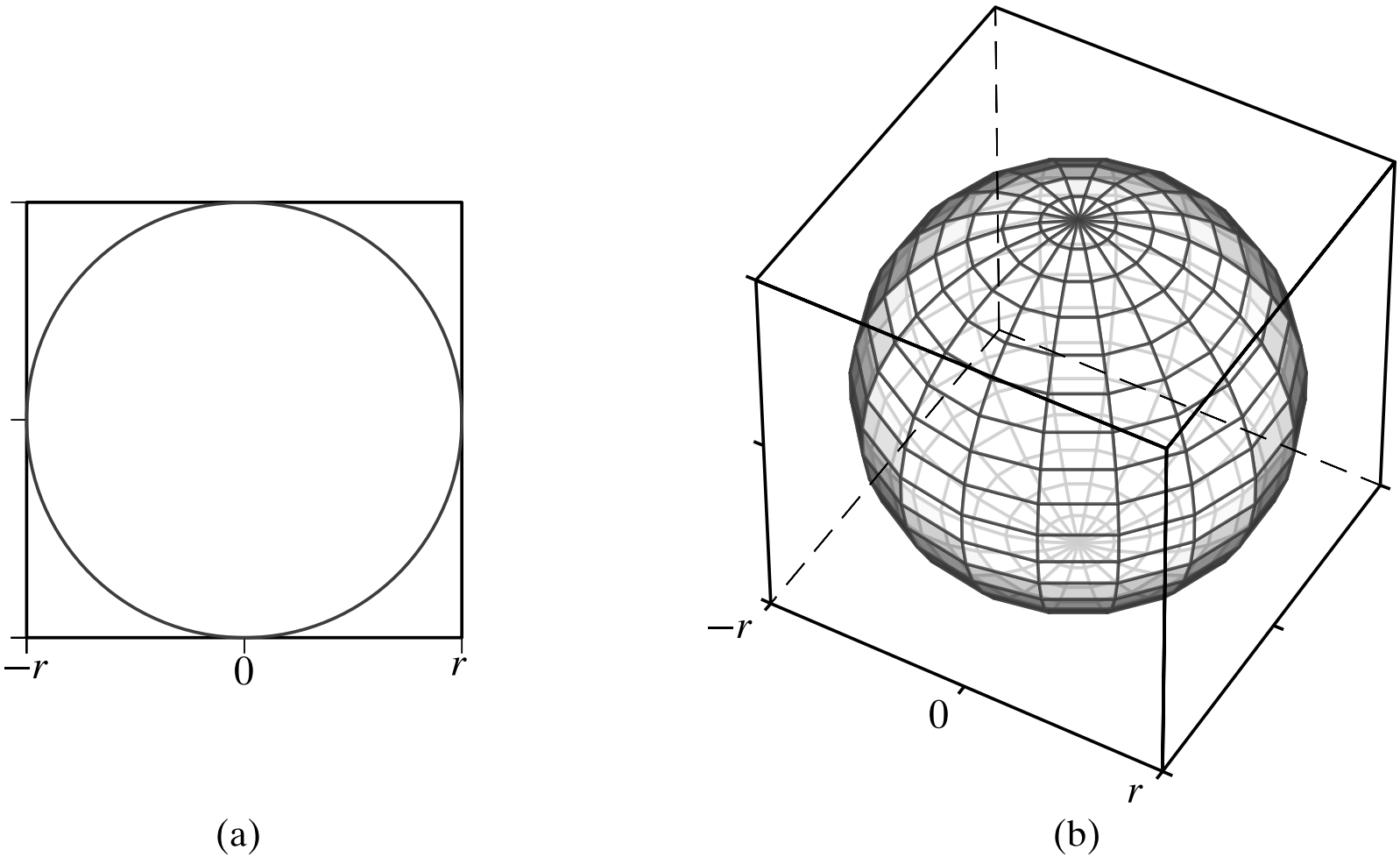


Figure 6.4. Hypersphere inscribed inside a hypercube: in (a) two and (b) three dimensions.

Curse of Dimensionality

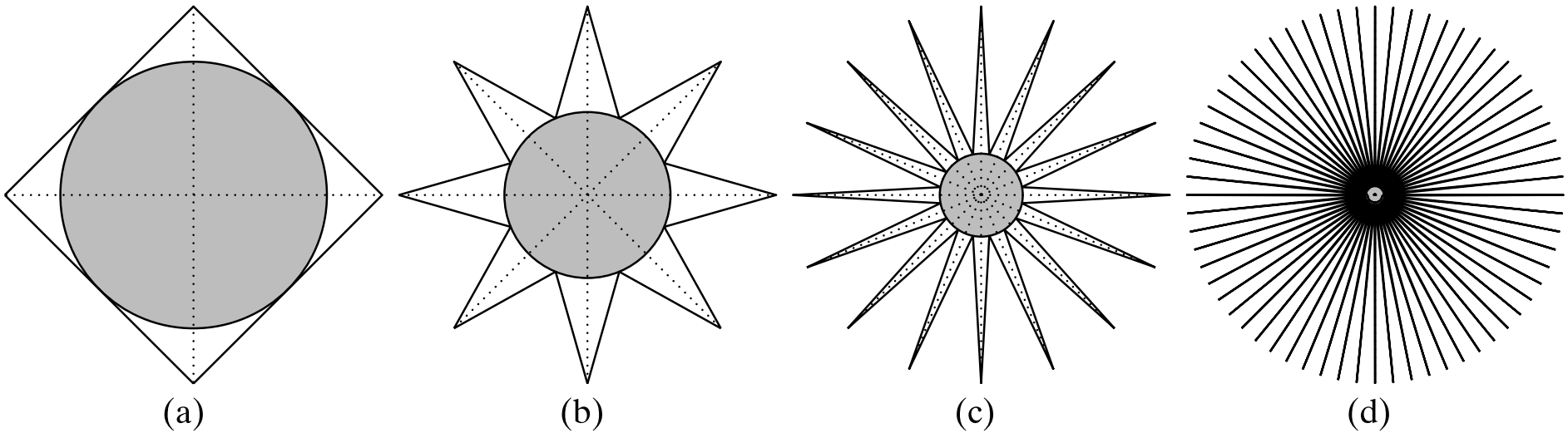
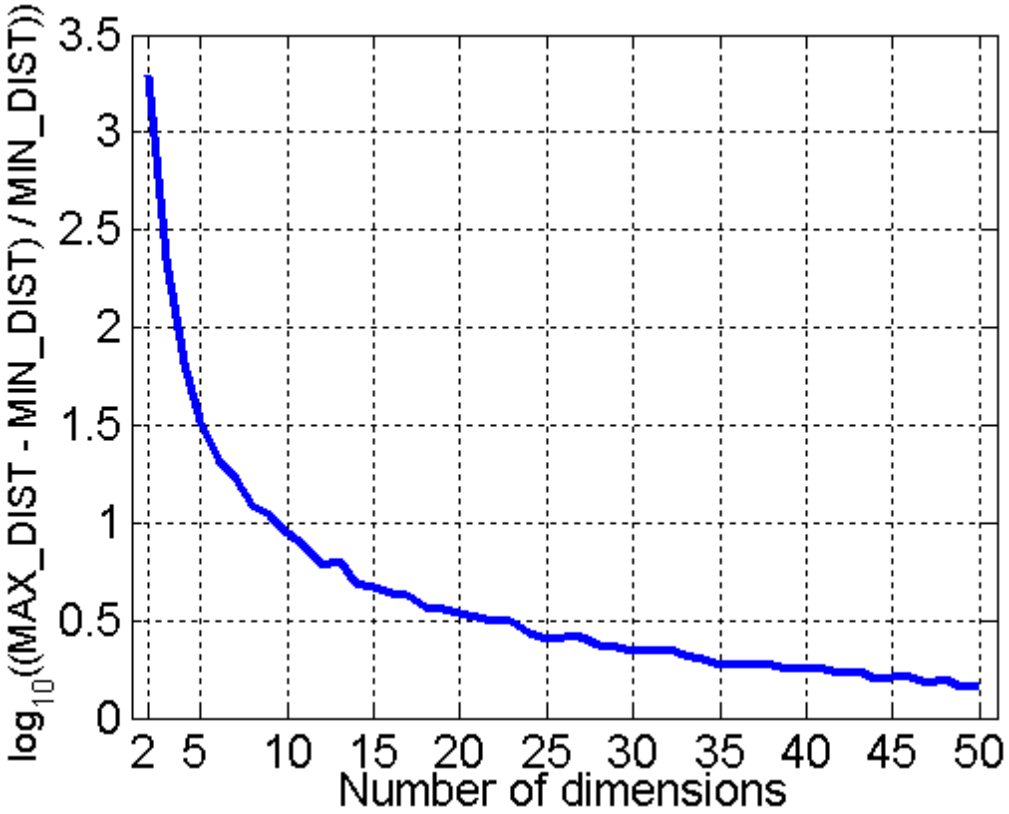


Figure 6.5. Conceptual view of high-dimensional space: (a) two, (b) three, (c) four, and (d) higher dimensions. In d dimensions there are 2^d "corners" and 2^{d-1} diagonals. The radius of the inscribed circle accurately reflects the difference between the volume of the hypercube and the inscribed hypersphere in d dimensions.

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.
- Definitions of density and distance between points, which are 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

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

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.
- Techniques:
  - Principal Components Analysis (PCA).
  - Singular Value Decomposition.
  - Others: supervised and non-linear techniques.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Principal Component Analysis (PCA)

- Principal Components Analysis (PCA) is a linear algebra technique for continuous attributes that finds new attributes (principal components) that are:
  1. are linear combinations of the original attributes,
  2. are orthogonal (perpendicular) to each other, and
  3. capture the maximum amount of variation in the data.

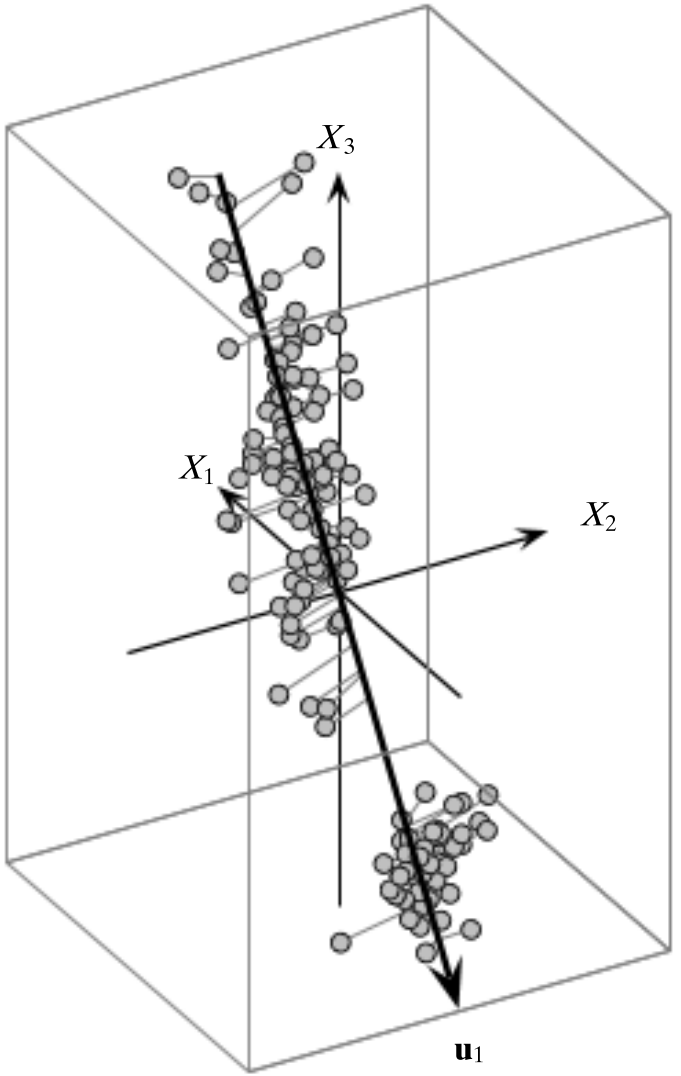
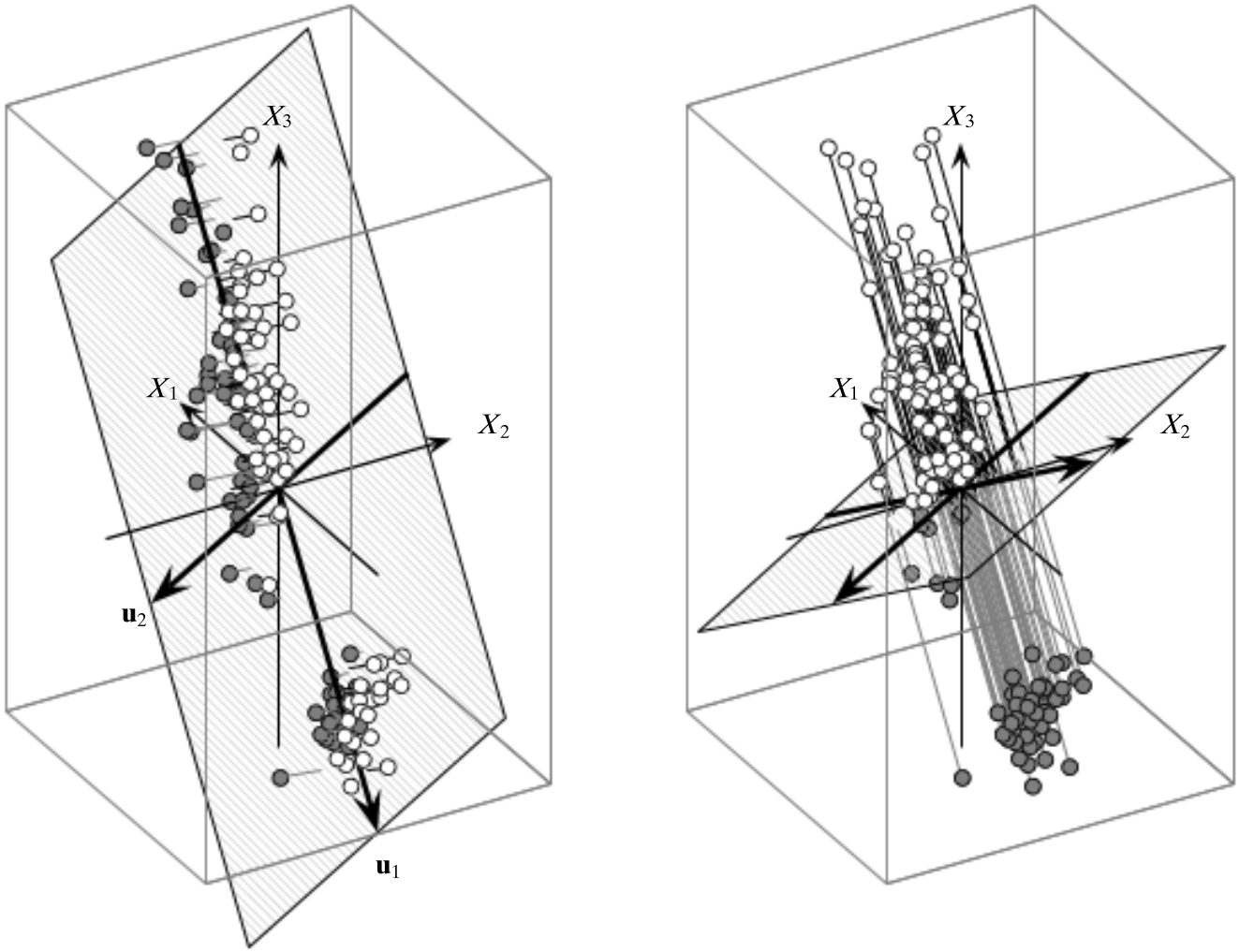


Figure 7.2. Best one-dimensional or line approximation.

All from: 1. Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pea
2. Zaki and Meira, "Data Mining and Machine Learning: Fundamental Cor
Press.

)

Principal Component Analysis (PCA)



(a) Optimal basis (b) Nonoptimal basis

Figure 7.3. Best two-dimensional approximation.

All from: Zaki and Meira, "Data Mining and Machine Learning: Fundamental Concepts and Algorithms", 2nd Edition, 2020, Cambridge University Press.

Linear Discriminant Analysis (LDA)

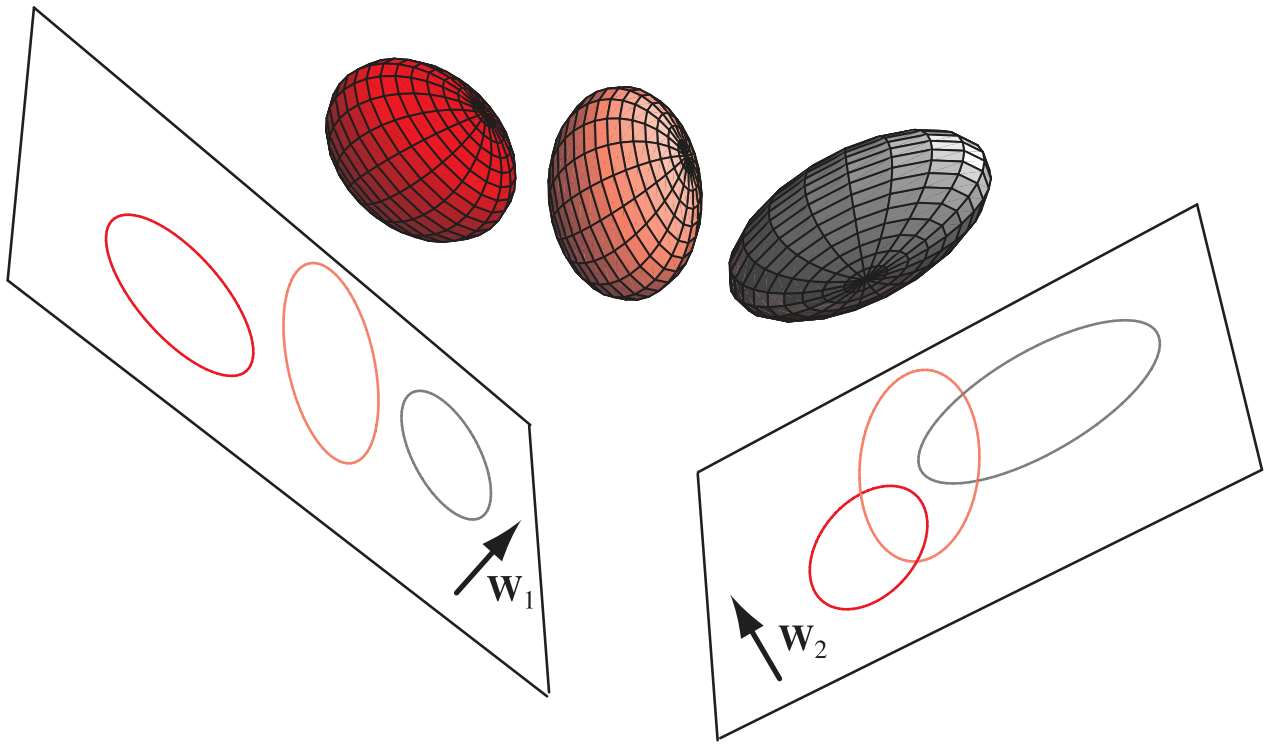


Figure 4.28: Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. Informally, multiple discriminant methods seek the optimum such subspace, i.e., the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{w}_1$.

All from: Duda et al. "Pattern Classification", 2nd Edition, 2001, John Wiley & Sons, Inc.

Feature 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.
  - Many techniques developed, especially for classification.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Feature Selection — Ideal Approach

- Irrelevant and redundant attributes can be eliminated by using common sense domain knowledge.
- A more systematic approach is however required.
- Ideal Approach:
  - Enumerate all possible subsets of features.
  - For each subset obtain performance results form the data mining algorithm being used.
- Pros: reflects the objective and bias of the data mining algorithm that will eventually be used.
- Cons: number of subsets involving n attributes is 2^n

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Feature Selection — Other Approaches

- Embedded Approaches: algorithm itself decides which attributes to use and which to ignore (e.g., decision trees).
- Filter Approaches: features are selected independently and separately from the algorithm being run (e.g., where pairwise correlation is as low as possible).
- Wrapper Approaches: same as Ideal Approach but without enumerating all possible subsets.

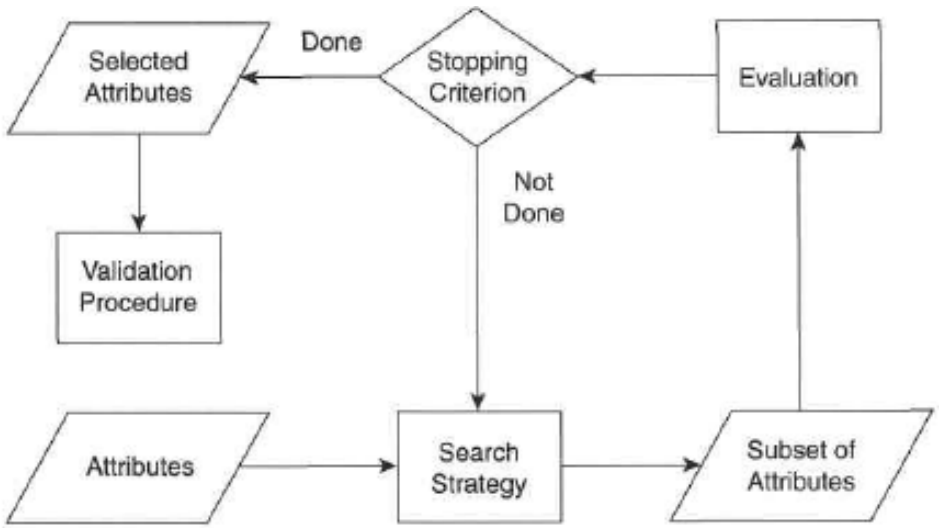


Figure 2.11. Flowchart of a feature subset selection process.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

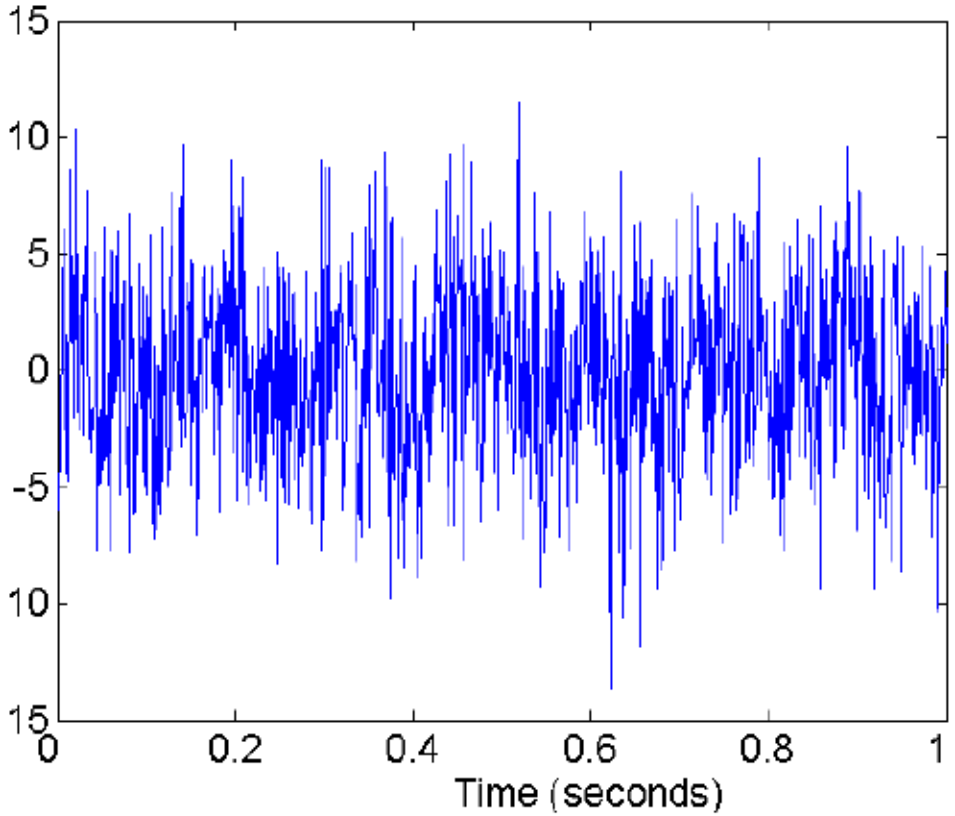
Feature Creation

- 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.

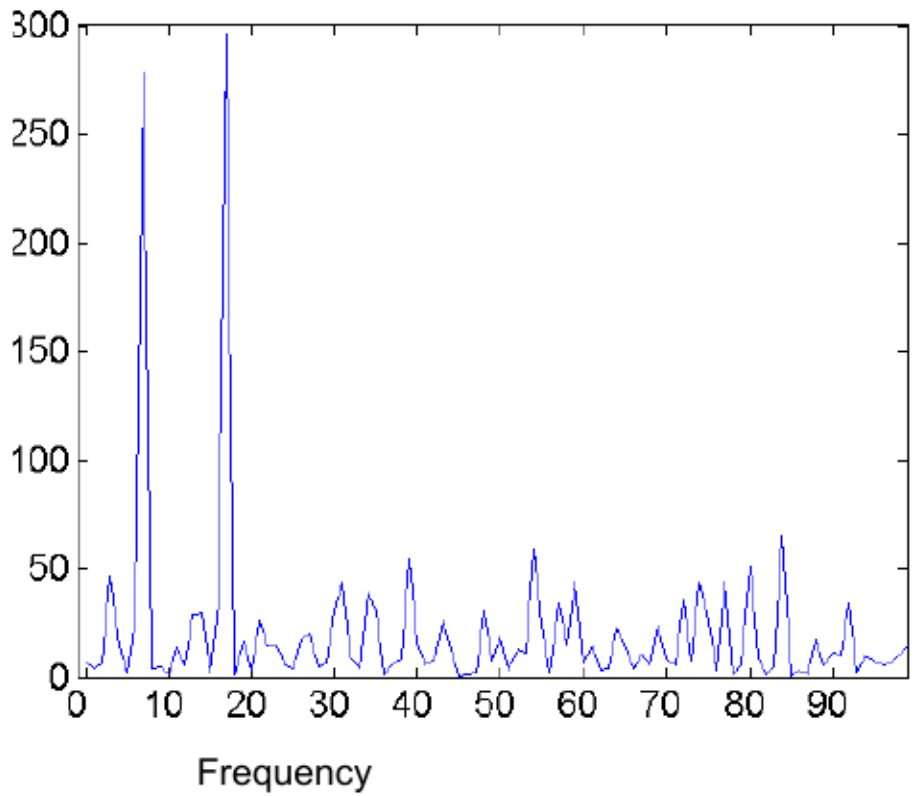
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Feature Creation

- Example of mapping data to new space via Fourier Transform.



Two Sine Waves + Noise



Frequency

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Discretization

- Certain classification algorithms, require data in the form of categorical attributes.
- Algorithms that find association patterns require that the data be in the form of binary attributes.
- Discretization: transform a continuous attribute into a categorical attribute.
- Binarization: Both continuous and discrete attributes may need to be transformed into one or more binary attributes.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Binarization

- A simple approach to binarize m categorical values:
  1. Assign a unique integer in the interval $[0, m-1]$. Maintain order during assignment if the attribute is ordinal.
  2. Convert each of these m integers to a binary number.
  3. $n = \lceil \log_2(m) \rceil$ binary attributes are required to represent these integers.
- Problems: unintended correlations introduced.

Categorical Value	**Integer Value**	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
okay	2	0	1	0
good	3	0	1	1
great	4	1	0	0

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Binarization

- For association problems asymmetrical attributes are required.
- Therefore necessary to introduce one binary attribute for each categorical value.

Categorical Value	**Integer Value**	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
okay	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

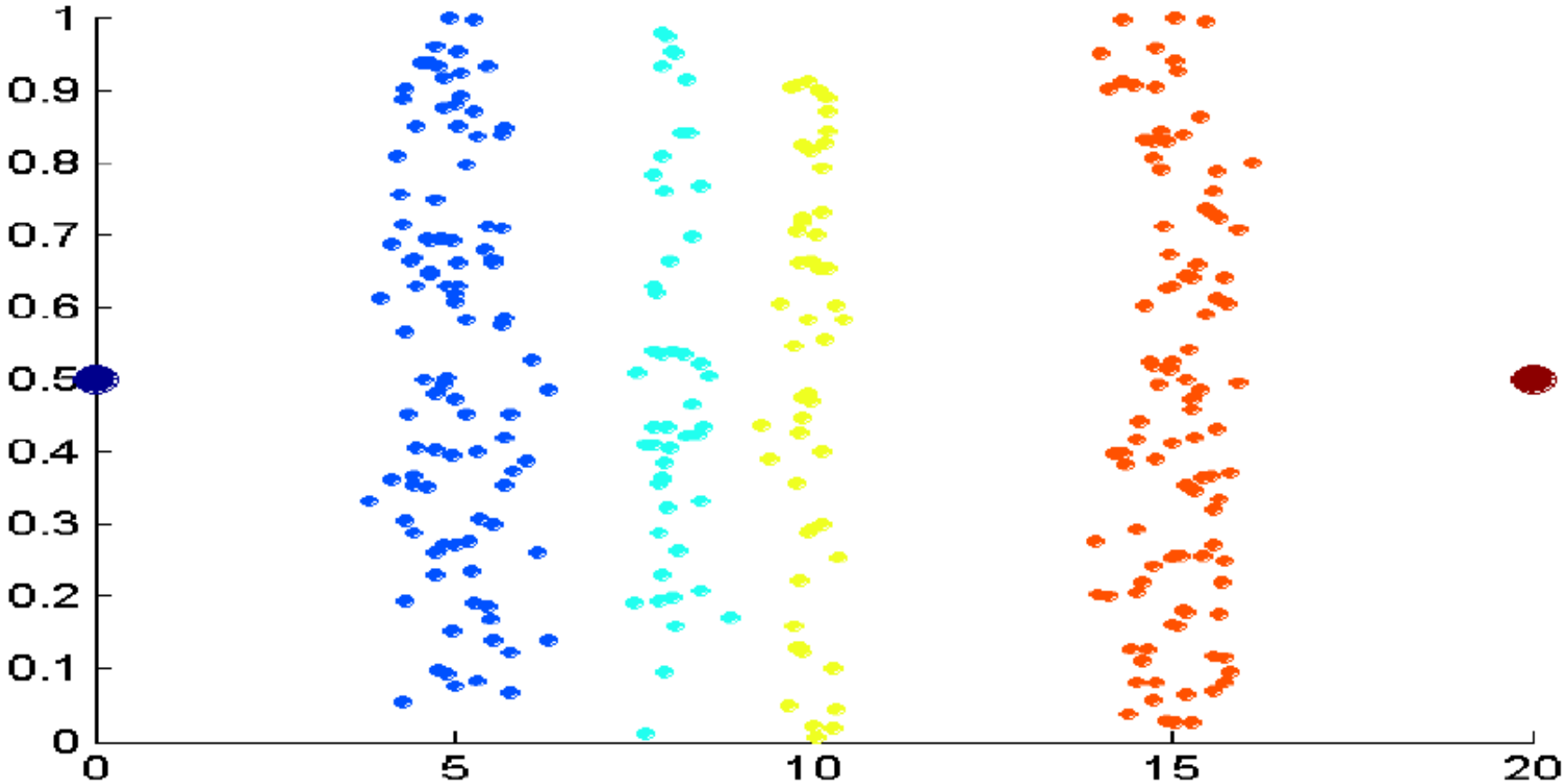
All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Discretization

- Transformation of a continuous attribute to a categorical attribute involves two subtasks:
  1. First step: continuous attribute are sorted and then are divided into n intervals by specifying $n-1$ split points.
  2. Second step: all the values in one interval are mapped to the same categorical value.
- Unsupervised Approaches: label associated with the objects are not used.
  - Example: equal width, equal frequency / equal depth, and k-means.
- Supervised Approaches: labels associated with the objects are used.
  - Example: a simple approach for partitioning a continuous attribute starts by bisecting the initial values so that the resulting two intervals give minimum entropy.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

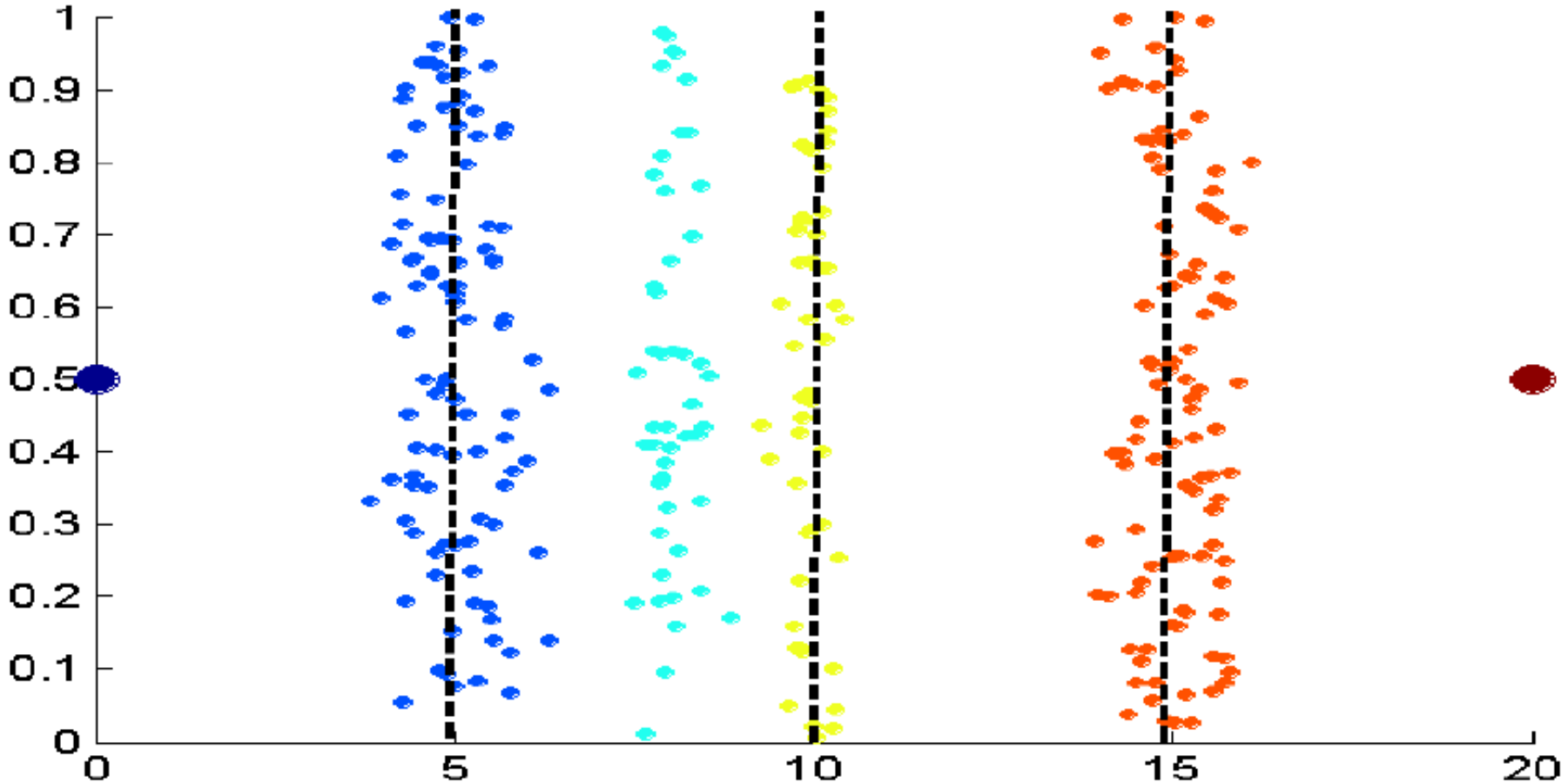
Discretization



Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

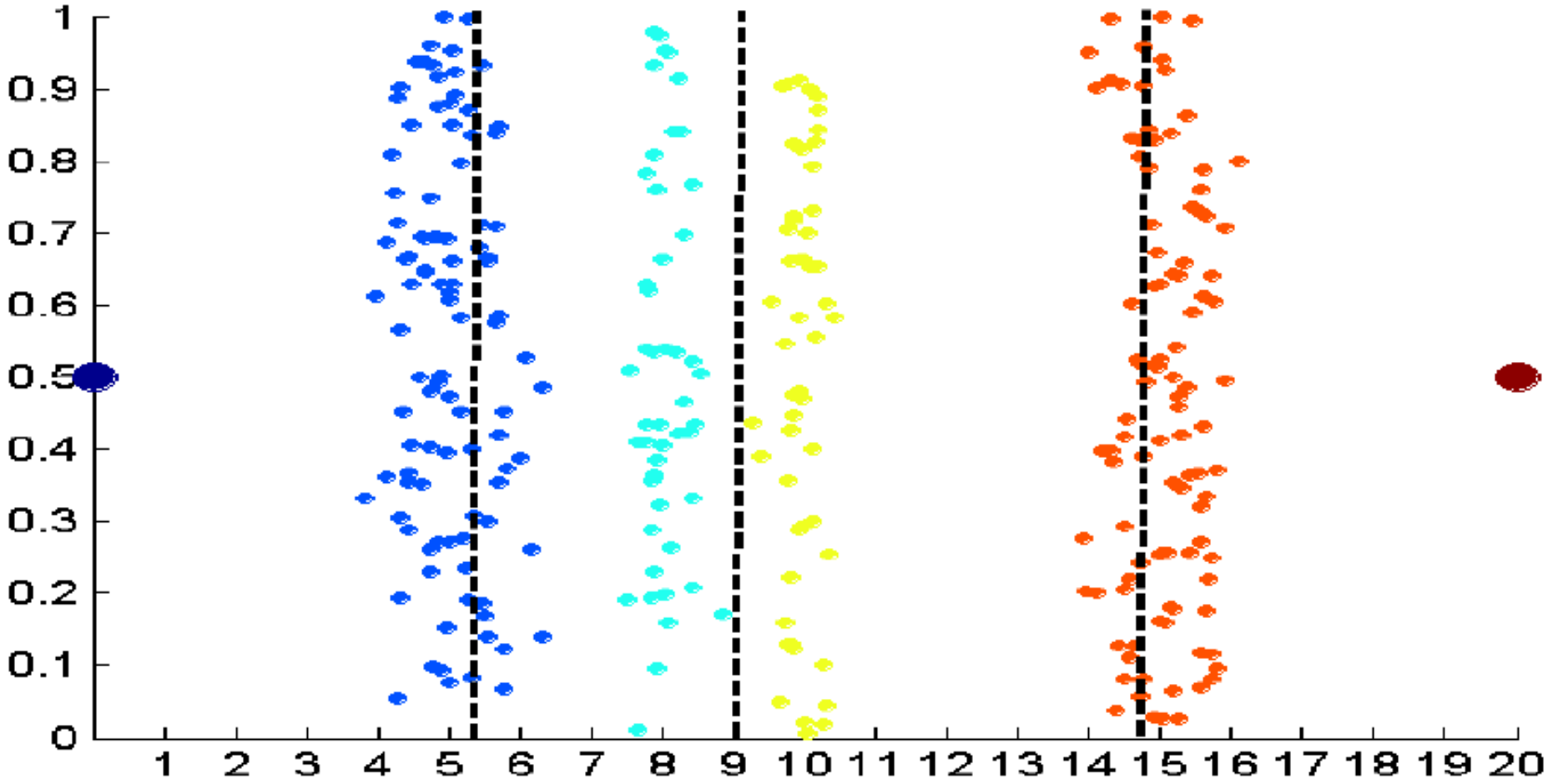
Discretization



Equal interval width approach used to obtain 4 values.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

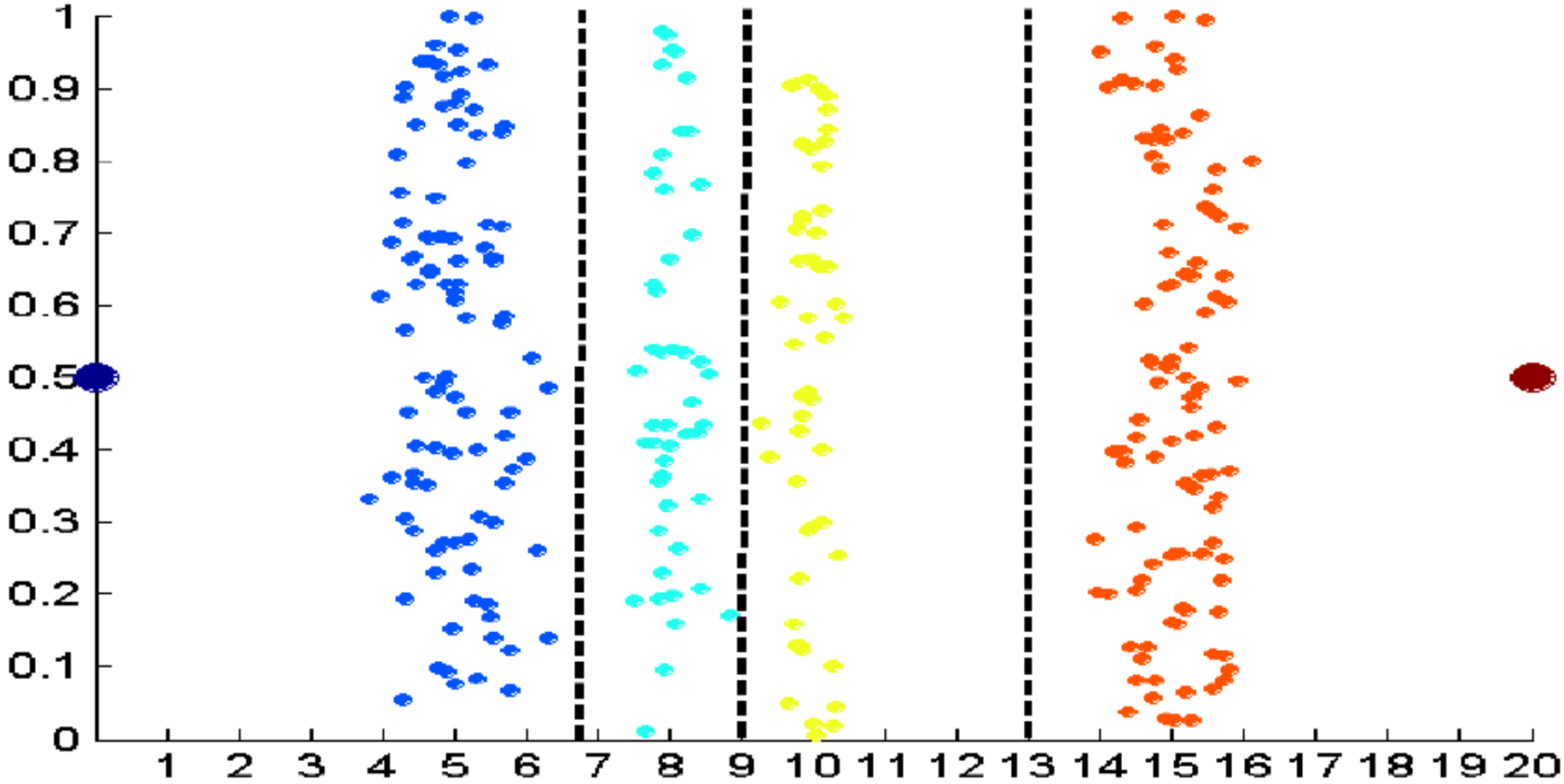
Discretization



Equal frequency approach used to obtain 4 values.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Discretization



K-means approach to obtain 4 values.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Simple Functions

- Attribute Transform: 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.
- Using Simple Functions:
  - Using transformations such as x^k, $\log(x)$, e^x, $\sqrt{x}$, $1/x$, or $|x|$.
  - Variable transformations should be applied with caution since they change the nature of the data.
  - Does the transformation apply to all values?
  - Especially negative values and 0?
  - What is the effect of the transformation on the values between 0 and 1?

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Normalization or Standardization

- Normalization refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, and range.
- Take out unwanted, common signal, e.g., seasonality.
- In statistics, standardization refers to subtracting off the means and dividing by the standard deviation.

$$x' = \frac{x - \bar{x}}{s_x} \tag{1}$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Similarity and Dissimilarity

- Similarity between two objects:
  - A numerical measure indicating degree to which the two objects are alike.
  - Usually non-negative and are often between 0 (no similarity) and 1 (complete similarity).
- Dissimilarity between two objects:
  - A numerical measure of the degree to which the two objects are different.
  - Dissimilarities are lower for more similar pairs of objects.
  - Dissimilarities sometimes fall in the interval [0,1], but it is also common for them to range from 0 to ∞.
  - Distances, which are dissimilarities with certain properties.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Similarity and Dissimilarity 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	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Figure: The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Distances (Dissimilarities with Certain Properties)

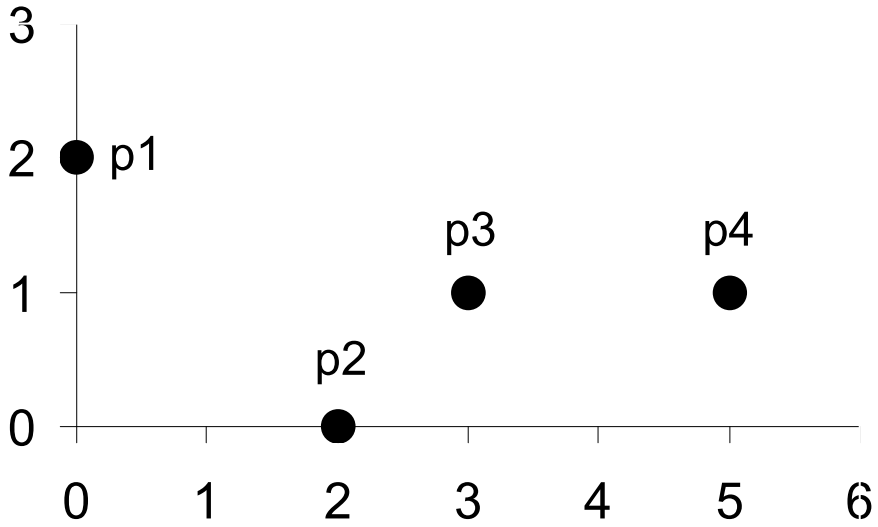
- Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{2}$$

- where, n is the number of dimensions (attributes) and x_k and y_k are respectively the k^{th} attributes (components) or data objects x and y.
- Standardization is necessary if scales differ.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Minkowski Distance

- Minkowski Distance — generalization of Euclidean distance:

$$d(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r} \tag{3}$$

- where, r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L 1 norm) distance.
  - A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance.
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm, Uniform norm, and Chebyshev) distance.
  - This is the maximum difference between any component of the vectors.
  - That is,

$$d(x, y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r} = \max_{k}^{n} |x_k - y_k|. \tag{4}$$

- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

All from: 1. Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.
2. Han et al. "Data Mining Concepts and Techniques", 3rd Edition, 2012, Morgan Kaufmann Publishers.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Common Properties of Distance

- Distances, such as the Euclidean distance, have some well known properties.
  - Positivity: $d(x, y) \geq 0$ for all x and y and $d(x, y) = 0$ if and only if $x = y$.
  - Symmetry: $d(x, y) = d(y, x)$ for all x and y.
  - Triangle Inequality: $d(x, y) \leq d(x, y) + d(y, z)$ for all points x, y, and z.
- A distance that satisfies these properties is a metric.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Similarity between Binary Vectors

- Common situation is that objects, x and y, have only binary attributes.
- Compute similarities using the following quantities:
  - f_{00} = the number of attributes where x was 0 and y was 0.
  - f_{01} = the number of attributes where x was 0 and y was 1.
  - f_{10} = the number of attributes where x was 1 and y was 0.
  - f_{11} = the number of attributes where x was 1 and y was 1.
- Simple Matching Coefficient:

$$SMC = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}}. \tag{5}$$

- Jaccard Coefficient:

$$J = \frac{\text{number of 11 matches}}{\text{number of non-zero attributes}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}. \tag{6}$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Simple Matching and Jaccard Coefficient Example

- $x = \langle 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 \rangle$
- $y = \langle 0, 0, 0, 0, 0, 0, 1, 0, 0, 1 \rangle$
  - $f_{00} = 7$ the number of attributes where x was 0 and y was 0.
  - $f_{01} = 2$ the number of attributes where x was 0 and y was 1.
  - $f_{10} = 1$ the number of attributes where x was 1 and y was 0.
  - $f_{11} = 0$ the number of attributes where x was 1 and y was 1.
- Simple Matching Coefficient:

$$SMC = \frac{f_{11} + f_{00}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{0 + 7}{7 + 2 + 1 + 0} = \frac{7}{10}. \tag{7}$$

- Jaccard Coefficient:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2 + 1 + 0} = 0. \tag{8}$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Recall — 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.
- Example of sparse data matrix.

	$\mathbf{word}_1$	$\mathbf{word}_2$	$\mathbf{word}_3$	...	$\mathbf{word}_{\lvert\mathcal{V}\rvert}$
$\mathbf{document}_1$	10	0	0	...	5
$\mathbf{document}_2$	1	2	0	...	0
⋮	⋮	⋮	⋮	...	⋮
$\mathbf{document}_n$	0	1	0	...	0

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Cosine Similarity

- If x and y are two document vectors, then:

$$\cos(x, y) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} = \frac{\sum_{k=1}^{n} x_k \cdot y_k}{\sqrt{\sum_{k=1}^{n} x_k^2} \cdot \sqrt{\sum_{k=1}^{n} y_k^2}}. \tag{9}$$

- where, $\langle x, y \rangle$ indicates the inner product or vector dot product of vectors, x and y, and $\|x\|$ is the length of vector x.
- Example:

$$d_1 = \langle 3, 2, 0, 5, 0, 0, 0, 2, 0, 0 \rangle,$$
$$d_2 = \langle 1, 0, 0, 0, 0, 0, 0, 1, 0, 2 \rangle.$$

$$\langle d_1, d_2 \rangle = 3 \cdot 1 + 2 \cdot 0 + 0 \cdot 0 + 5 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 2 = 5$$

$$\|d_1\| = \sqrt{(3 \cdot 3 + 2 \cdot 2 + 0 \cdot 0 + 5 \cdot 5 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 2 + 0 \cdot 0 + 0 \cdot 0)} = \sqrt{42} = 6.481$$

$$\|d_2\| = \sqrt{(1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 2 \cdot 2)} = \sqrt{6} = 2.449$$

$$\cos(d_1, d_2) = \frac{5}{6.481 \cdot 2.449} = 0.3150$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Correlation

- Correlation between two data objects that have binary or continuous variables is a measure of the linear relationship between the attributes of the objects.
- Perfect Correlation: correlation is always in the range -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship.
- That is, $x_k = a \cdot y_k + b$, where a, and b are constants.
- Example 1:

$$x = \langle -3, 6, 0, 3, -6 \rangle \text{ and } y = \langle 1, -2, 0, -1, 2 \rangle.$$

- where, $x = -3 \cdot y$ implies correlation is -1.
- Example 2:

$$x = \langle 3, 6, 0, 3, 6 \rangle \text{ and } y = \langle 1, 2, 0, 1, 2 \rangle.$$

- where, $x = 3 \cdot y$ implies correlation is 1.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Pearson's Correlation Coefficient

- Pearson's Correlation Coefficient between two data objects, x and y is defined by the following equation:

$$\text{corr}(x, y) = \frac{\text{covariance(x,y)}}{\text{std\_dev}(x) \cdot \text{std\_dev}(y)} = \frac{s_{xy}}{s_x \cdot s_y}. \tag{10}$$

- where, the standard statistical notation and definitions are:

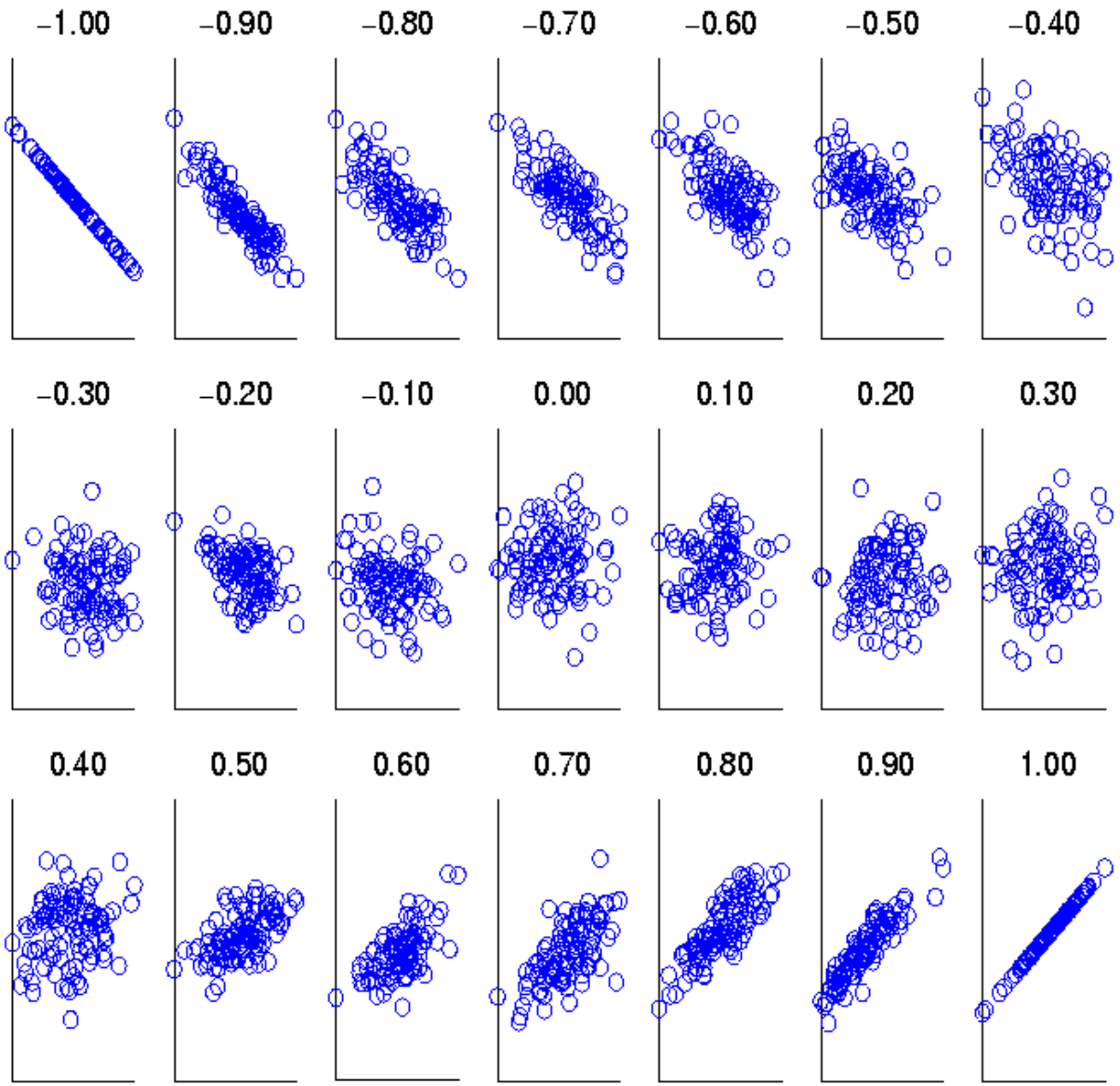
$$\text{covariance}(x, y) = s_{xy} = \frac{1}{n-1} \cdot \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y}). \tag{11}$$

$$\text{std\_dev}(x) = s_x = \sqrt{\frac{1}{n-1} \cdot \sum_{k=1}^{n} (x_k - \bar{x})^2}. \tag{12}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } x. \tag{13}$$

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Visualizing Correlations



Scatter plots showing the similarity from –1 to 1.

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

1. Announcements and References
   - Administrative
   - References for Today's Lecture

2. Data
   - Data Analysis Perspectives
   - Data Quality
   - Data Preprocessing
     - Aggregation
     - Sampling
     - Curse of Dimensionality
     - Feature Selection
     - Feature Creation
     - Discretization and Binarization
     - Attribute Transformations
     - Similarity and Dissimilarity
     - Correlation

3. Summary

Summary — 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
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Summary — Data Preprocessing

- Data Preprocessing: processing the data so as to improve the data mining analysis with respect to time, cost, and quality.
- This done by: selecting data objects and attributes for the analysis or creating/changing the attributes.
- Typical data preprocessing steps:
  - Aggregation
  - Sampling
  - Discretization and Binarization
  - Attribute Transformation
  - Dimensionality Reduction
  - Feature Subset Selection
  - Feature Creation

All from: Tan et al. "Introduction to Data Mining", 1st Edition, 2006, Pearson Education Inc.

Summary — Administrative

1. First Assignment
   - Available and due by 09.February.2023.
2. Volunteers for feedback regarding course
   - Interested? Please contact me by email!