Data Mining: Data

Lecture Notes for Chapter 2

2.1 Types of Data - 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

Measurement of Length

- **The way you measure an attribute is somewhat may not match the attributes properties.**
  - 두번째 선분은 첫번째 선분 **2**개 합쳐진것, 세번째 선분은 첫번째 선분 **3**개가 합쳐진 것…
  - 오른쪽은 이와 같은 **2**배, **3**배,**4**배 등의 배수 특성이 반영되어 있지만, 왼쪽은 길이 속성의 순서 속성만 반영되어 있음

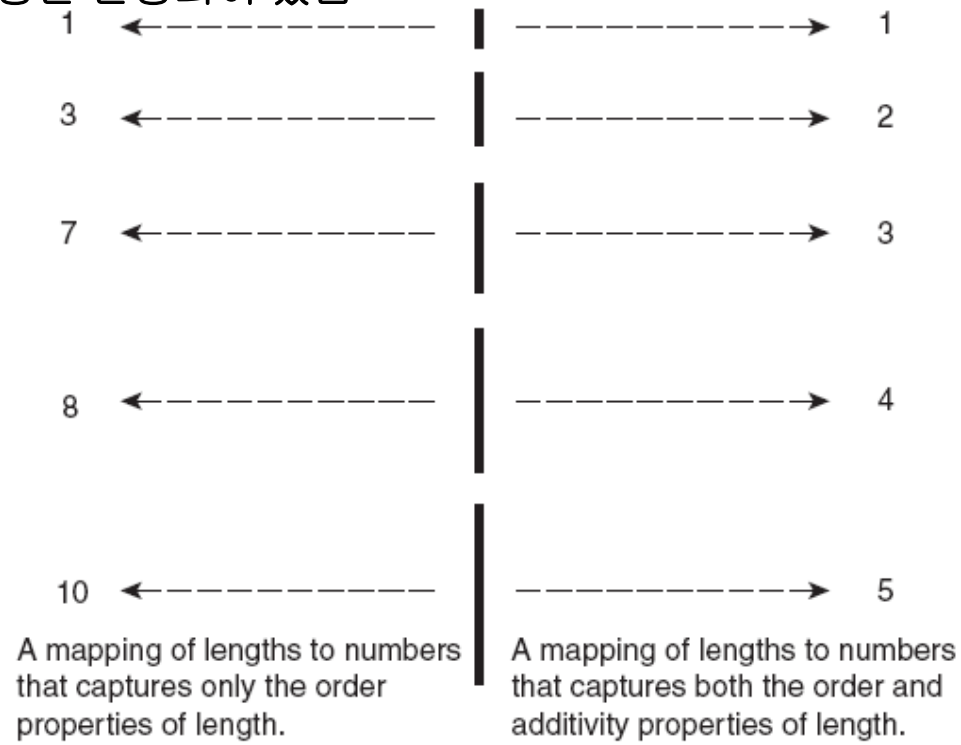


Figure 2.1. The measurement of the length of line segments on two different scales of measurement.

Types of Attributes

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

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness(유일성): = ≠
  - Order(순서): < >
  - Addition(덧셈): + -
  - Multiplication(곱셈): * /

  - Nominal attribute(명목형): distinctness
  - Ordinal attribute(서열형): distinctness & order
  - Interval attribute(구간형): distinctness, order & addition
  - Ratio attribute(비율형): all 4 properties

Attribute Type	Description	Examples	Operations
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. (=, ≠) **(명목형은 단순히 서로 다른 이름뿐 → 서로 구별 정보만 제공)**	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >) **(서열형은 서로 순서를 정하는데 필요한 정보 제공)**	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	Median(중간값), percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -) **(구간형은 값들 사이의 차이가 의미가 있음)**	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. (*, /) **(비율형은 차이와 비율이 모두 의미가 있음)**	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation(퍼센트 편차)

< 속성 수준을 정의한 변환예 >

Attribute Level	Transformation	Comments
Nominal	Any permutation of values (순서 변환: 순열)	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. (값의 순서 보존형 변환)	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 (새값 = a*예전값 + b)	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* (새값 = a * 예전값)	Length can be measured in meters or feet.

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.

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

Important Characteristics of Structured Data

- **Dimensionality**
  - 데이터 집합의 객체들이 갖는 속성의 수/변수의 수
  - **Curse of Dimensionality(차원의 저주) 문제 존재**
  - **Dimensionality reduction(차원 감소) 필요**
- **Sparsity**
  - 객체 대부분의 속성이 **0**이며, 일부**(1%** 이내**)** 객체만 속성이 **0**이 아닌 경우 → **0**이 아닌 값만 처리하면 됨**(**이 때, 희소성은 장점이 됨**)**
  - **Only presence counts**
- **Resolution**
  - 데이터는 상이한 수준의 해상도로 얻는 경우 많음
  - **Patterns depend on the measurement scale(척도)**
  - 예) 지구 표면을 **km**단위로 보는 경우**(**평평함**) vs. m** 단위로 보는 경우

Record Data

- Data that consists of a collection of records, each of which consists of 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

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

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

Transaction Data

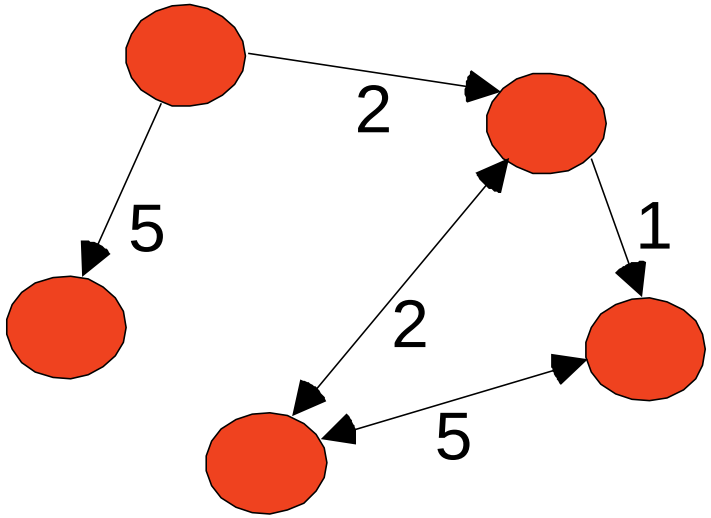
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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 (그래프 기반 데이터)

- 그래프 데이터:
  - 데이터 객체간의 관계를 나타내는 그래프
  - 데이터 객체 자체가 그래프로 표현될 수 있음

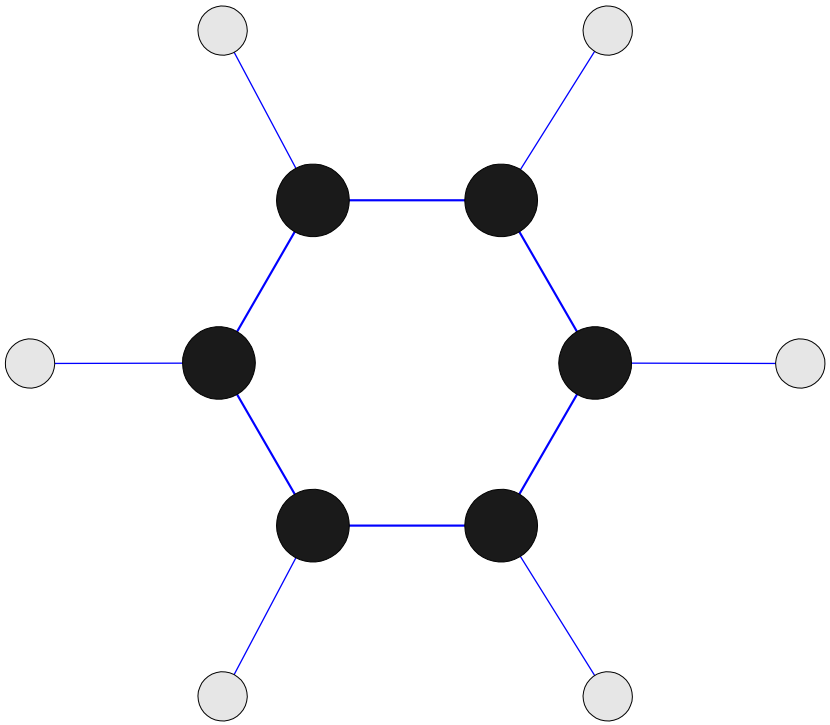
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Graph Data - Chemical Data

- Examples: Benzene Molecule: C_6H_6



Ordered Data (서열형 데이터)

- 서열형 데이터
  - 데이터 속성이 시간/공간 순서와 관련된 관계 가짐
  - Temporal data(시간 데이터)
  - Sequence data(서열 데이터)
  - Time series data(시계열 데이터)
  - Spatial data(공간 데이터)

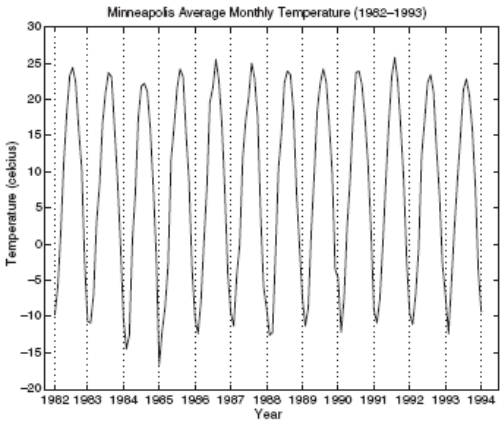
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

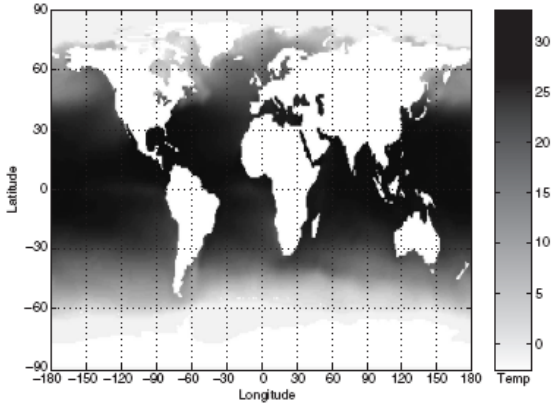
(a) Sequential transaction data.

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

(b) Genomic sequence data.



(c) Temperature time series.



(d) Spatial temperature data.

Figure 2.4. Different variations of ordered data.

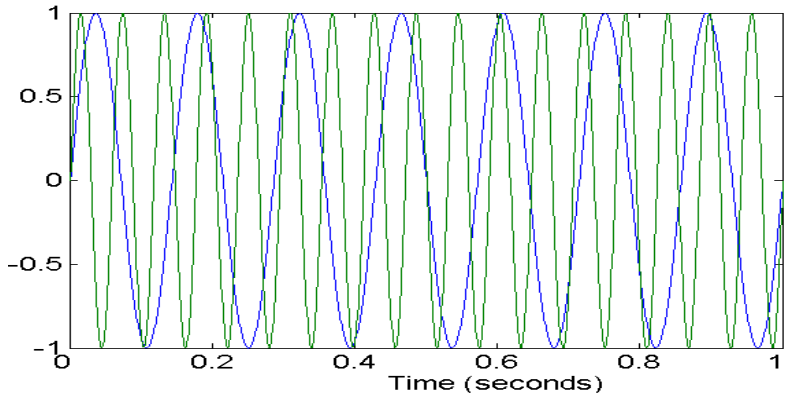
2.2 Data Quality (데이터 품질)

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

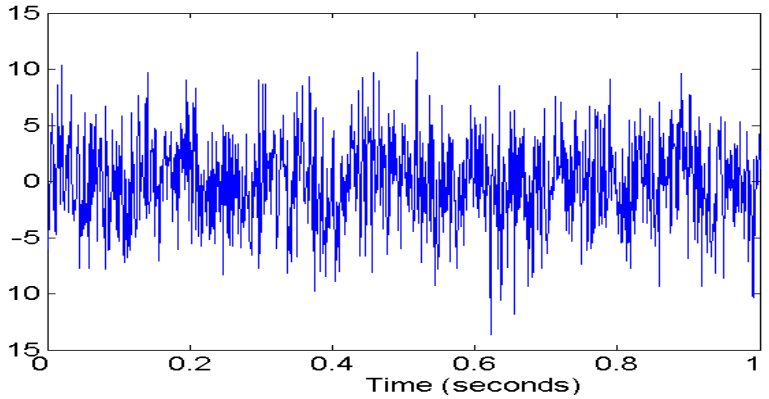
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- Noise is the random component of a measurement error
  - It may involve the distortion of a value or the addition of spurious objects
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



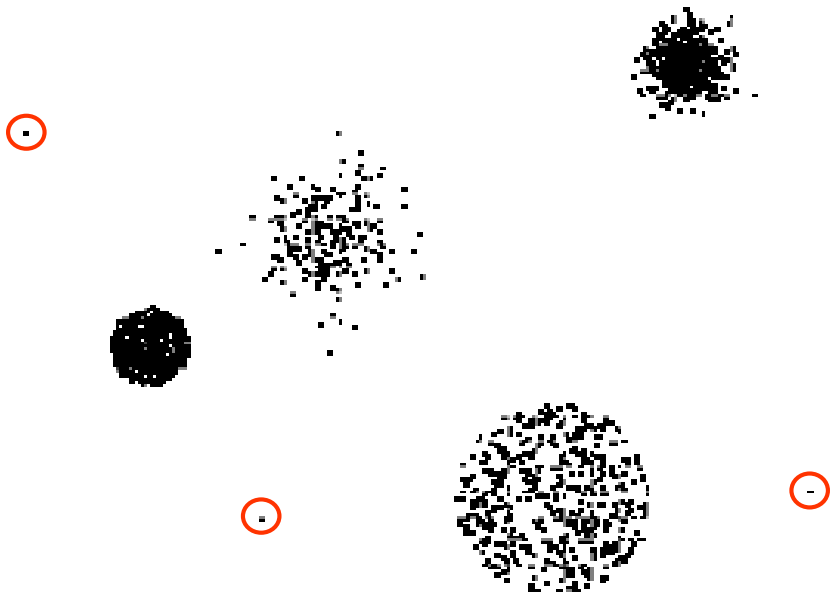
Two Sine Waves



Two Sine Waves + random Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- It is usual for an object to be missing one or more attribute 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
  - Estimate Missing Values (누락값 추정)
  - Eliminate Data Objects (누락값 제거)
  - Ignore the Missing Value During Analysis (분석 과정에서 누락값 무시)
  - Replace with all possible values (weighted by their probabilities)
    (누락값을 가능성있는 값으로 대체)

Duplicate Data (중복 데이터)

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning (데이터 정제)
  - Process of dealing with duplicate data issues

2.3 Data Preprocessing(데이터 전처리)

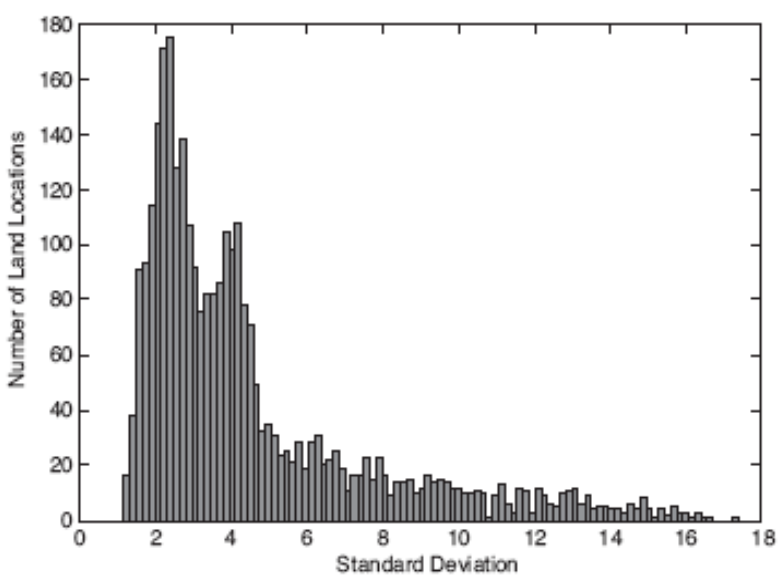
- Aggregation(집합/집계/총계)
- Sampling(표본 추출)
- Dimensionality Reduction(차원 축소)
- Feature subset selection(특징 부분집합 선택)
- Feature creation(특징 생성)
- Discretization and Binarization(이산화와 이진화)
- Attribute Transformation(속성 변환)

Aggregation

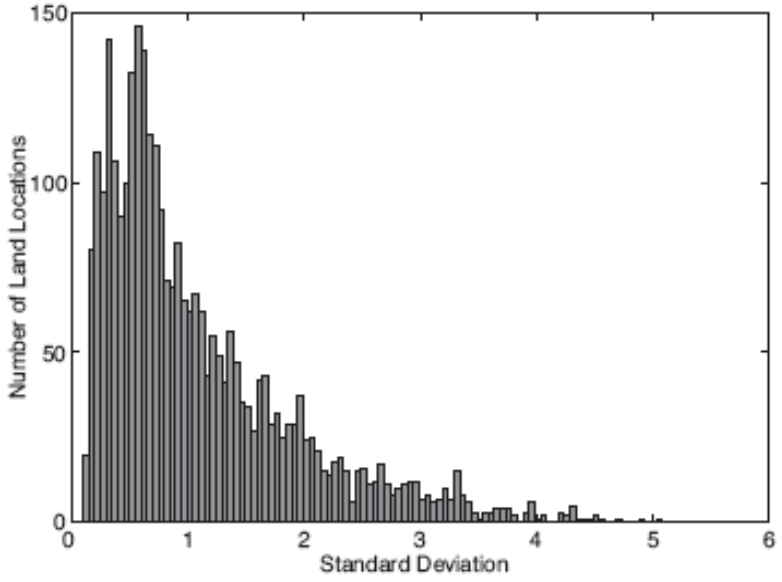
- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

Aggregation



(a) Histogram of standard deviation of average monthly precipitation



(b) Histogram of standard deviation of average yearly precipitation

Figure 2.8. Histograms of standard deviation for monthly and yearly precipitation in Australia for the period 1982 to 1993.

(평균 월별 강수량의 표준편차) **(평균 연별 강수량의 표준 편차)**

평균 연별 강수량(b)이 더 적은 가변성을 가지고 있음

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 …

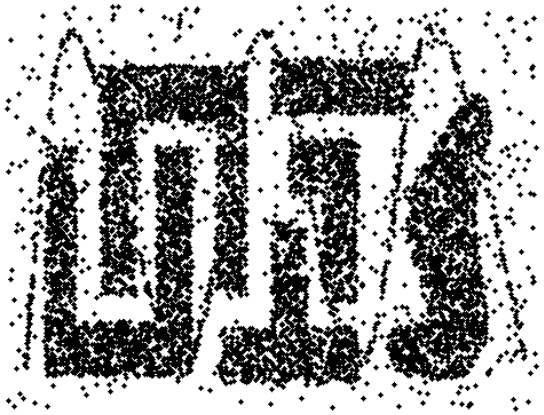
- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

Types of Sampling

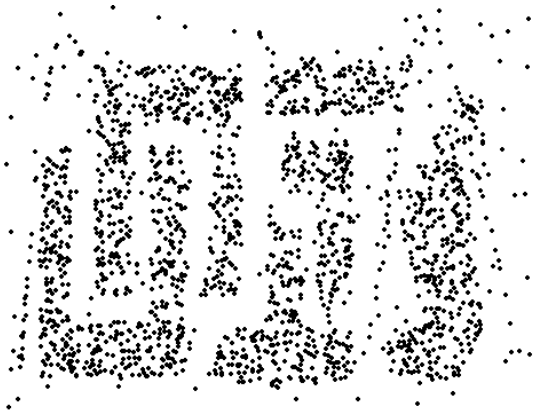
- Simple Random Sampling(단순 임의 표본추출)
  - There is an equal probability of selecting any particular item
- Sampling without replacement(무대체 표본추출)
  - As each item is selected, it is removed from the population
- Sampling with replacement(대체 표본추출)
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling(층화 표본추출)
  - Split the data into several partitions; then draw random samples from each partition (모집단을 층으로 나눈 후, 각 층에서 샘플링)
  - 층내에서는 동질적, 층간은 이질적 특성을 가지도록 하면 적은 비용으로 더 정확한 추정 가능

Sample Size

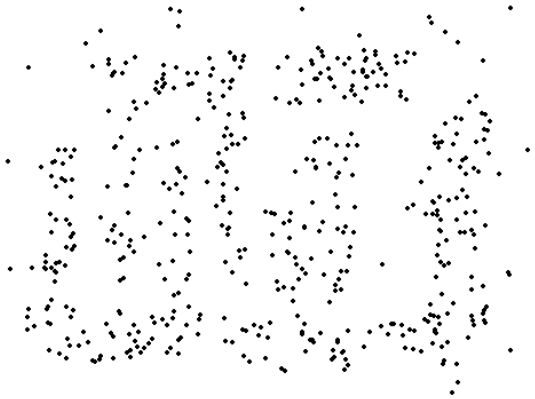
- 표본 추출과 정보 손실
  - 표본 크기가 커지면 표본이 전체 데이터의 대표성을 가질 확률이 높아짐(하지만, 표본추출의 장점이 퇴색됨)
  - 표본 크기가 작으면 패턴이 누락되거나 잘못된 패턴이 감지될 수 있음
  - 예:



8000 points



2000 Points

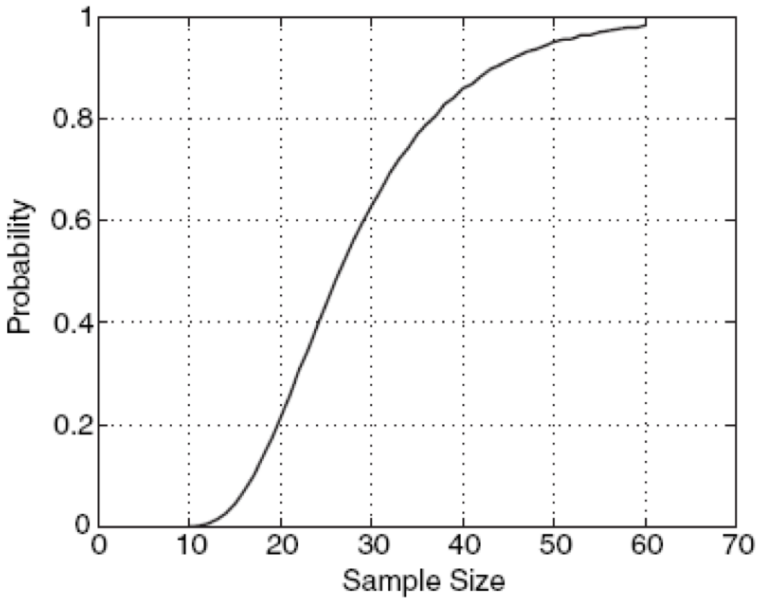


500 Points

Sample Size

- 적절한 표본 크기 결정
  - What sample size is necessary to get at least one object from each of 10 groups.

(a) Ten groups of points.



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

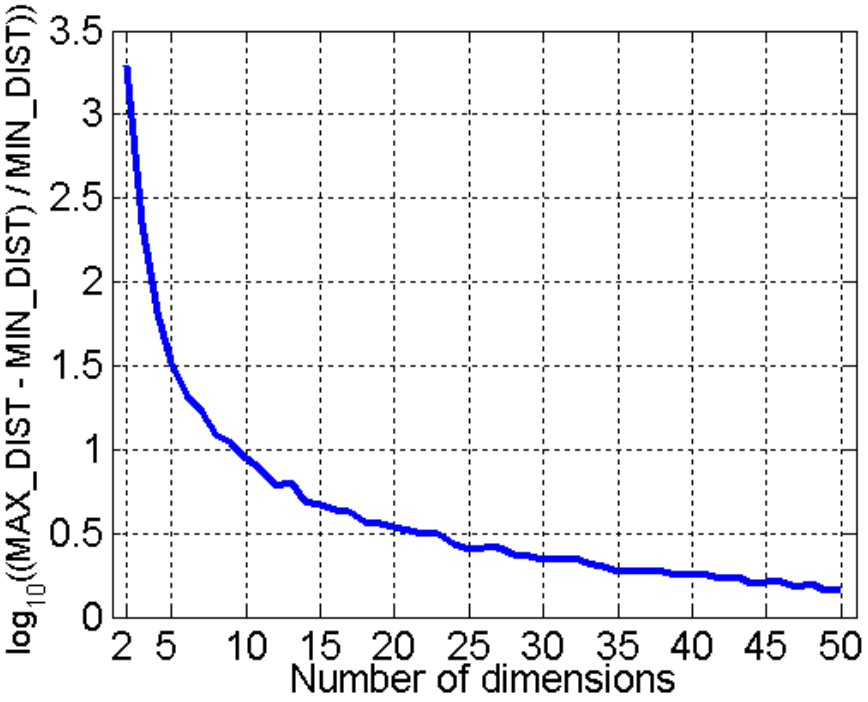
(b)는 표본 크기가 10에서 60까지 변할때, (a)이 10개 그룹에서 각각 하나의 객체를 선택할 확률임

샘플 크기가 커질 수록 10개의 모든 그룹에서 객체를 얻을 확률이 높아짐

Figure 2.10. Finding representative points from 10 groups.

Curse of Dimensionality

- 차원이 증가하면 데이터는 그 차원이 차지하는 공간상에서 점점더 sparse하게 됨 → 분석이 어려워짐
- Classification에선 충분한 데이터 객체가 존재하지 않아 모델 생성이 어려워짐
- Clustering에선 clustering의 핵심인 density와 두점간의 distance 정보가 작아서 군집화가 어려워짐



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

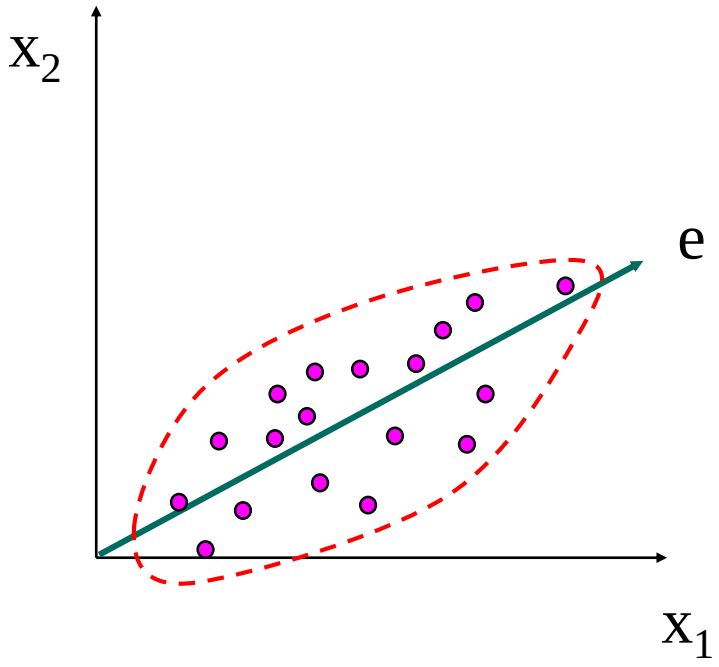
Dimensionality Reduction 기법

- 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
  - Principle Component Analysis (주성분 분석)
  - Singular Value Decomposition(특이값 분해)
  - Others: supervised and non-linear techniques

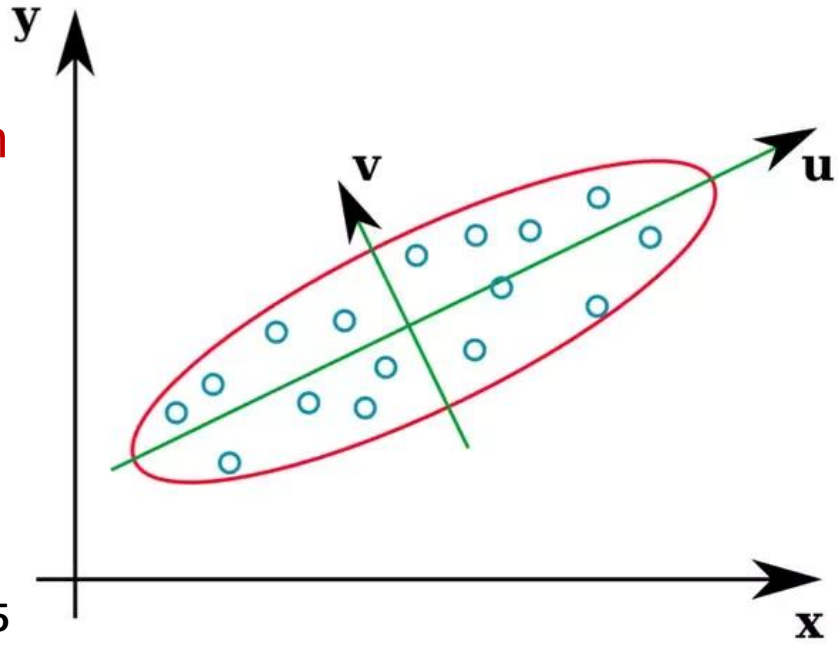
Dimensionality Reduction: PCA

- 최대한의 데이터 변이를 얻기 위함
  - The Goal is to find a projection that captures the largest amount of variation in data



Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
  - One of the most intuitive explanations of eigenvectors of a covariance matrix is that they are **the directions in which the data varies the most.**
  - Example) Each data sample is a 2 dimensional point with coordinates x, y. The eigenvectors of the covariance matrix of these data samples are the vectors u and v;
  - u : first eigenvector, v, the shorter arrow, is the second eigenvector



- The first eigenvector is the direction in which the data varies the most, the second eigenvector is the direction of greatest variance among those that are orthogonal (perpendicular) to the first eigenvector.
- The eigenvalues are the length of the arrows.

Feature Subset Selection(특징 부분집합 선택)

- Another way to reduce dimensionality of data
- Redundant features(중복 특징)
  - 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(비관련 특징)
  - 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

Feature Subset Selection

- 특징 부분집합 선택 기법:
  - 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:
    - Features are selected **before** data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm **as a black box** to find best subset of attributes

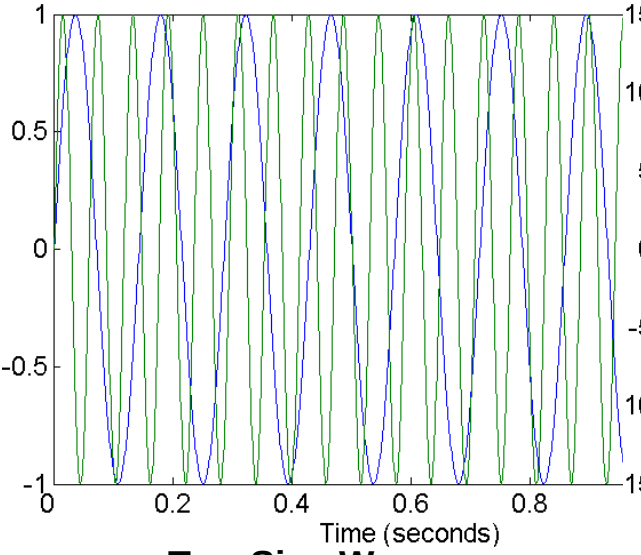
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

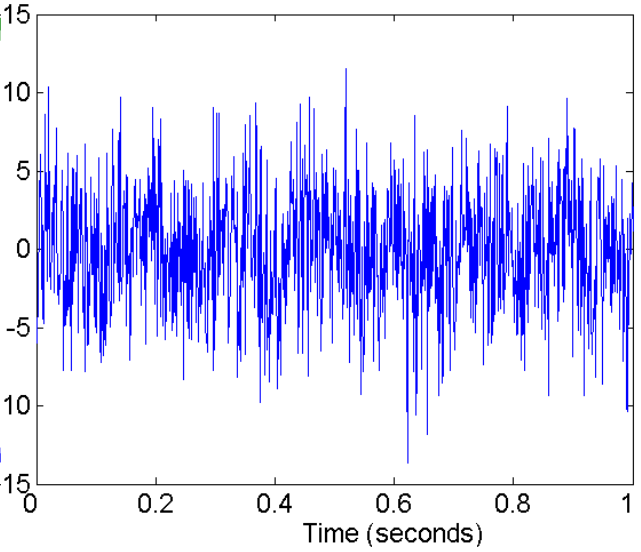
- Three general methodologies:
  - Feature Extraction(특징 추출)
    - domain-specific
  - Mapping Data to New Space(새로운 공간으로 데이터 매핑)
  - Feature Construction(특징 구축)
    - combining features

Mapping Data to a New Space

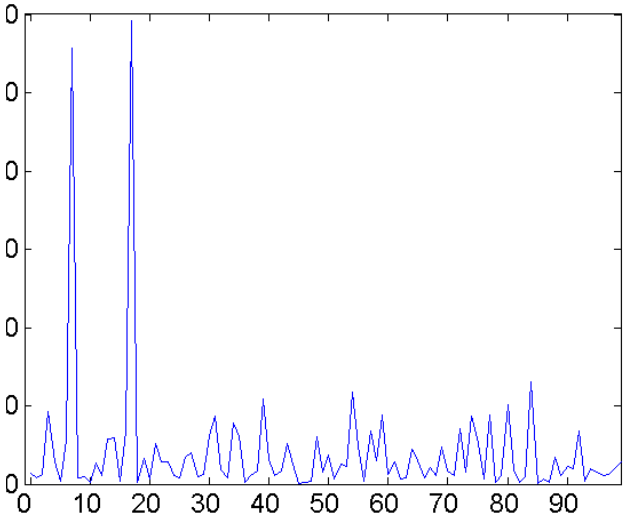
- **Fourier transform : sin/cos 함수로 time domain 신호→ frequency domain으로 변환**
- **Wavelet transform : sin/cos함수뿐만 아니라, 다양한 wavelet 모함수를 사용하여 변환함**



Two Sine Waves



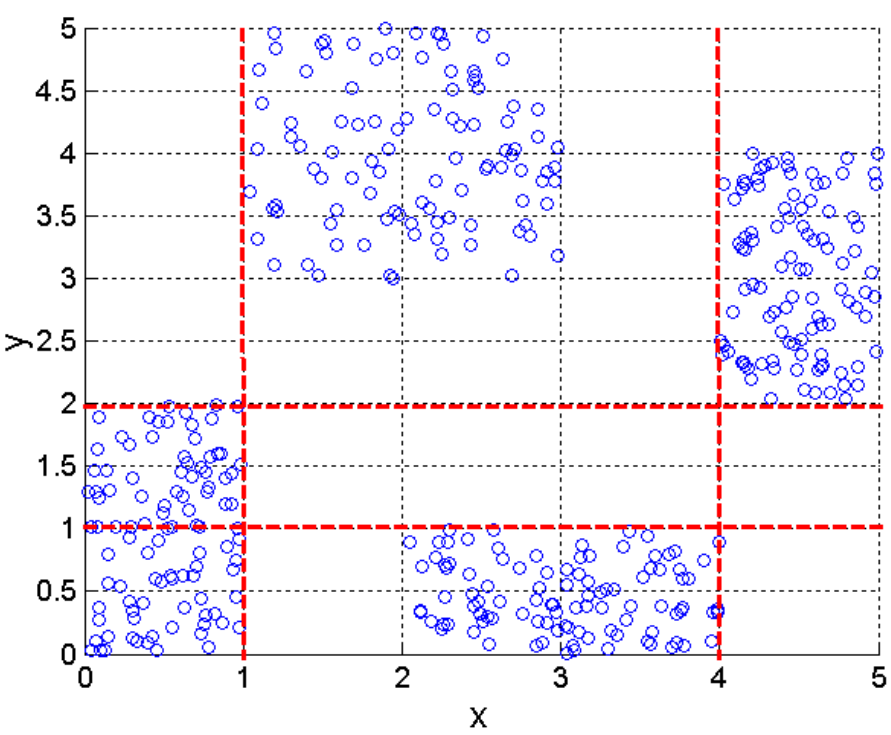
Two Sine Waves + Noise



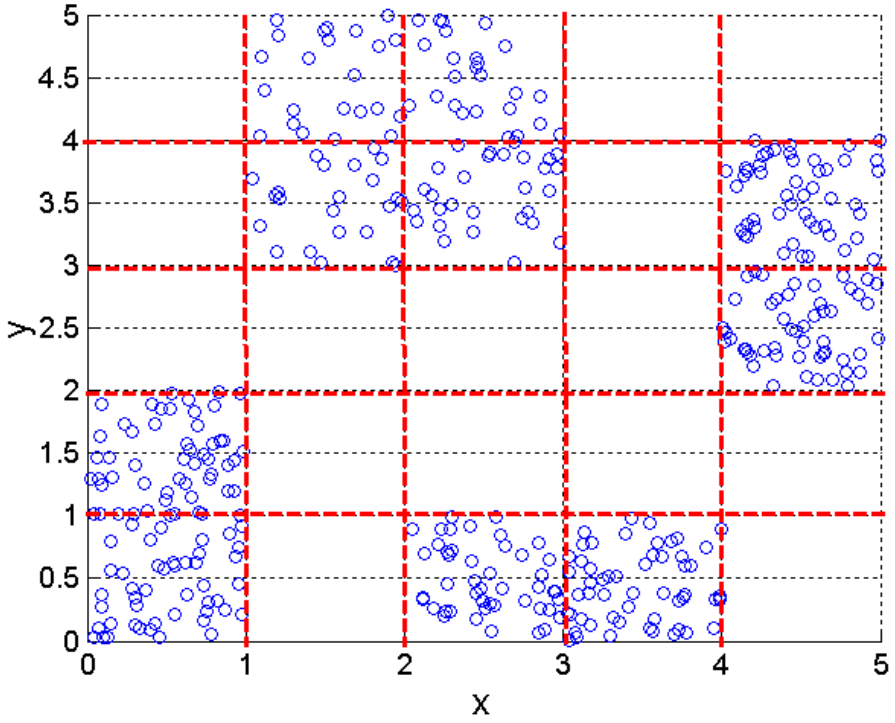
Frequency

Discretization Using Class Labels (이산화)

- **Discretization of Two Attributes**
  - (a) x와 y 속성을 세개의 구간으로 이산화함. (b)는 5개의 구간으로 이산화함 → (a),(b) 모두, 2차원에서는 잘 분리가 됨. 1차원 관점에서는 그러지 못함
  - (b)의 5개 구간으로 분리한 것이 더욱 잘 분리했음



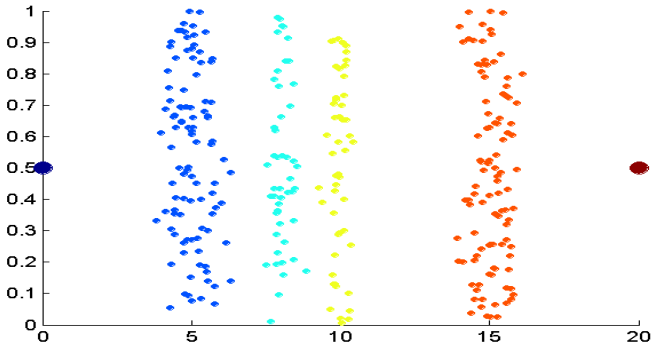
(a) 3 categories for both x and y



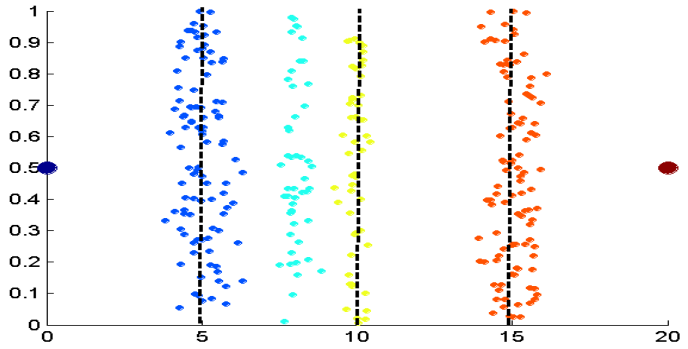
(b) 5 categories for both x and y

Discretization Without Using Class Labels

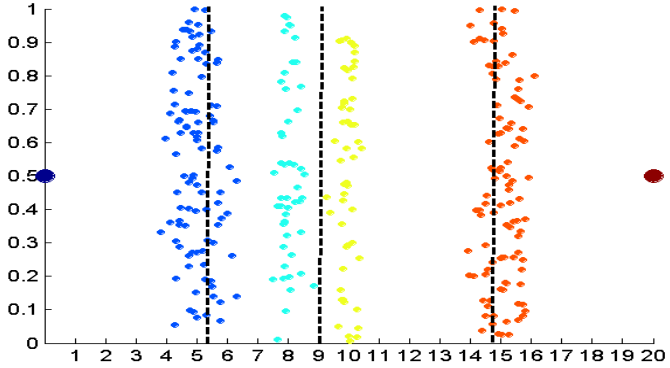
- 4개의 상이한 그룹에 속하는 데이터 점 **(a)**를 분리함(이산화)
- **(d) K-means** 기반 이산화가 최상임. 그 다음은 **(c)** 동등 주파수 기반 이산화, 그 다음이 **(b)** 동등폭 이산화 수행 결과



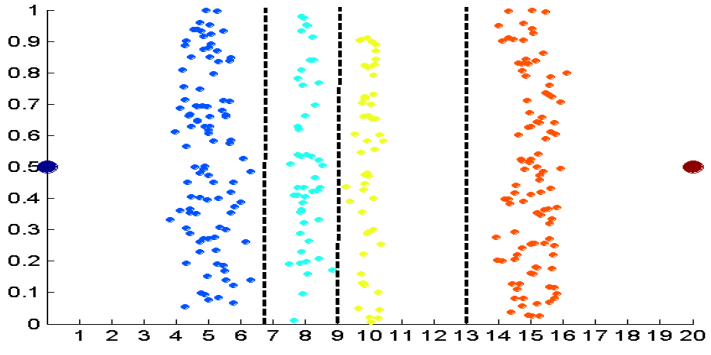
(a) 원 데이터



(b) 동등폭 이산화 수행



(c) 동등 주파수 이산화 수행



(d) K-means 이산화 수행

Attribute Transformation

- It alters the data by replacing a selected attribute by one or more new attributes
- 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, etc.

2.4 Similarity and Dissimilarity(유사도와 비유사도)

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

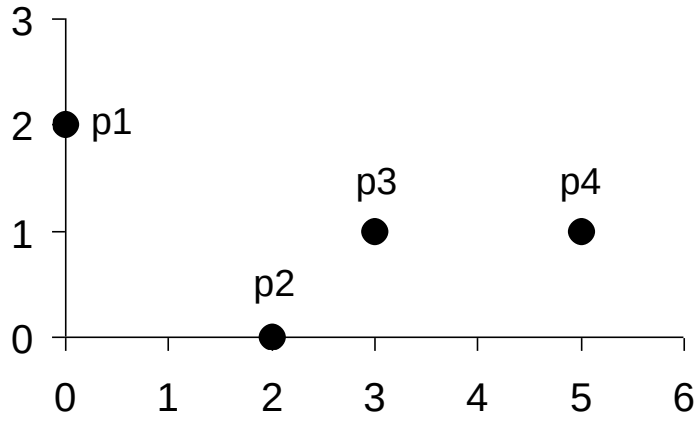
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.

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

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance: Examples

- *r* = 1. City block :Manhattan norm, taxicab norm, L_1 norm
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

- *r* = 2. Euclidean distance

Euclidean distance

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

L2 norm

$$\|\mathbf{p}\| = \sqrt{p_1^2 + p_2^2 + \cdots + p_n^2} = \sqrt{\mathbf{p} \cdot \mathbf{p}},$$

- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors

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

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

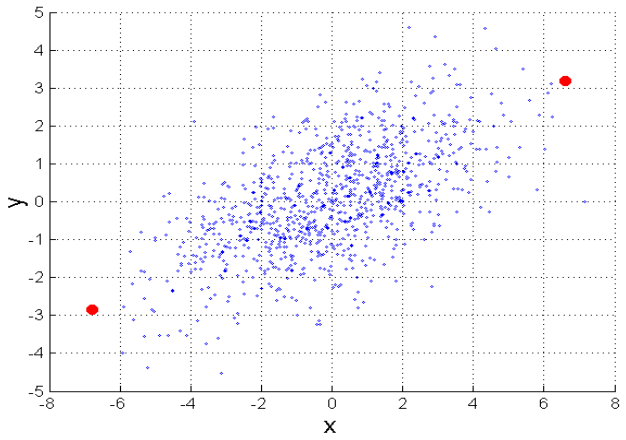
Distance Matrix

Mahalanobis Distance

- This is a measure of the distance between a point P and a distribution **D**
- P가 분포 **D**의 평균으로부터 표준편차의 몇 배 떨어져 있는지를 알 수 있음

$$D^2 = (\mathbf{x} - \mathbf{m})^{\mathbf{T}} \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$$

D^2 = Mahalanobis distance
$\mathbf{x}$ = Vector of data
$\mathbf{m}$ = Vector of mean values of independent variables
$\mathbf{C}^{-1}$ = Inverse Covariance matrix of independent variables
$\mathbf{T}$ = Indicates vector should be transposed



For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance (example)

$$D^2 = (\mathbf{x} - \mathbf{m})^{\mathbf{T}} \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$$

D^2 = Mahalanobis distance
$\mathbf{x}$ = Vector of data
$\mathbf{m}$ = Vector of mean values of independent variables
$\mathbf{C}^{-1}$ = Inverse Covariance matrix of independent variables
$\mathbf{T}$ = Indicates vector should be transposed

If, in our single observation, X = 410 and Y = 400 (other data is not shown in this example), we would calculate the Mahalanobis distance for that single value as:

< Calculation of Mahalanobis distance>

Given that Mahalanobis Distance $D^2 = (\mathbf{x} - \mathbf{m})^{\mathbf{T}} \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$

$$(\mathbf{x} - \mathbf{m}) = \begin{pmatrix} 410 - 500 \\ 400 - 500 \end{pmatrix} = \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$\mathbf{C}^{-1} = \begin{pmatrix} 6291.55737 & 3754.32851 \\ 3754.32851 & 6280.77066 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix}$$

$$\text{Therefore } D^2 = \begin{pmatrix} -90 & -100 \end{pmatrix} \times \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix} \times \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$= 1.825$$

< covariance matrix >

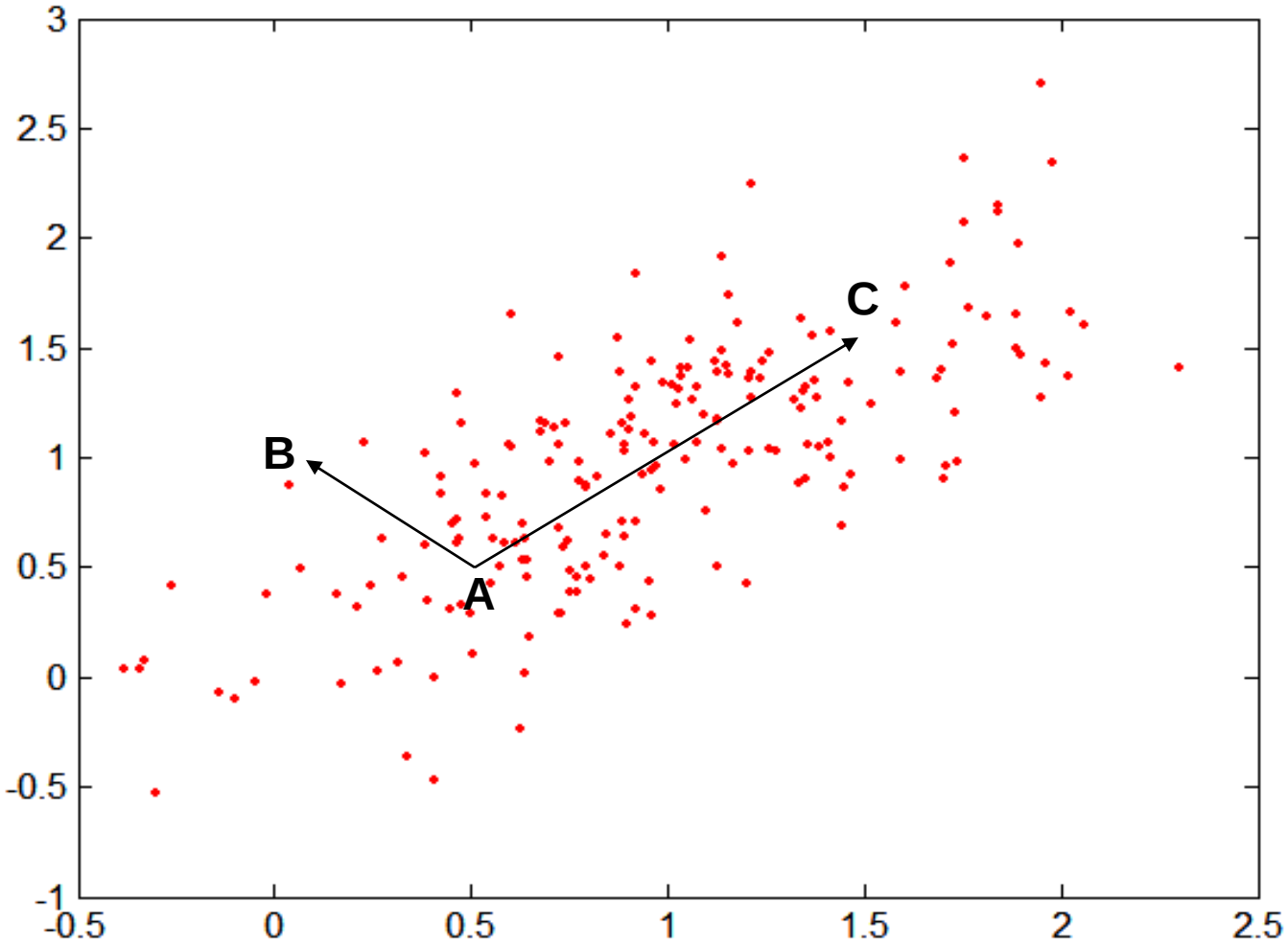
Variable X: mean = 500, SD = 79.32
Variable Y: mean = 500, SD = 79.25

and

Variance/Covariance Matrix		
	X	Y
X	6291.55737	3754.32851
Y	3754.32851	6280.77066

Var(X)=std_dev(X)^2, Cov(X,Y)=E((X-u)(Y-v))
Cov(Y,X), Var(Y)

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q, and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1

  M_{10} = the number of attributes where p was 1 and q was 0

  M_{00} = the number of attributes where p was 0 and q was 0

  M_{11} = the number of attributes where p was 1 and q was 1

- **Simple Matching Coefficients** (단순 매칭 계수)

  SMC = 매칭 속성의 수**/**속성의 수

  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

- SMC는 p도 없고 q도 없는 속성까지 고려를 하고 있음
- 예를 들어, 마켓에서 판매하는 1000개의 물건에서, 두 고객(p,q)의 장바구니(구매 물품)의 **SMC 유사도를 본다면**,
  - **p = (salt, pepper),**
  - **q= (salt, sugar)**
  - M11 = 1개(salt), M00=(1000개 – 3), M01=1(sugar), M10=1(pepper)
  - **SMC = (1 + 997)/(1 + 1 + 1 + 997) = 0.998 → 의미없음**

Similarity Between Binary Vectors

- Jaccard Coefficients (자카드 계수)
  - SMC와 달리 M_{00}을 사용하지 않음
  - 비대칭 이진 속성으로 구성된 객체의 유사도 처리에 유용

J = number of 11 matches / number of not-both-zero attributes values

= $(M_{11}) / (M_{01} + M_{10} + M_{11})$

  - 예를 들어, 마켓에서 판매하는 1000개의 물건에서, 두 고객의 장바구니(구매 물품)의 Jaccard 유사도를 본다면,
    - **p = (salt, pepper),**
    - **q= (salt, sugar)**
    - M11 = 1개(salt), M01=1(sugar), M10=1(pepper)
    - **Jaccard 유사도 = (1)/(1 + 1 + 1) = 1/3 → 두 장바구니의 Jaccard 유사도는 의미있음**

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)

M_{10} = 1 (the number of attributes where p was 1 and q was 0)

M_{00} = 7 (the number of attributes where p was 0 and q was 0)

M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where • indicates vector dot product and || *d* || is the length of vector *d*.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

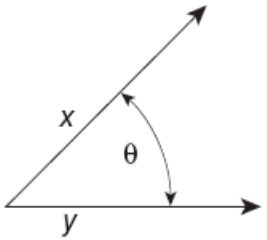


Figure 2.16. Geometric illustration of the cosine measure.

- Cosine 유사도가 0 이라는 것은 아무런 관련성 없음을 의미

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

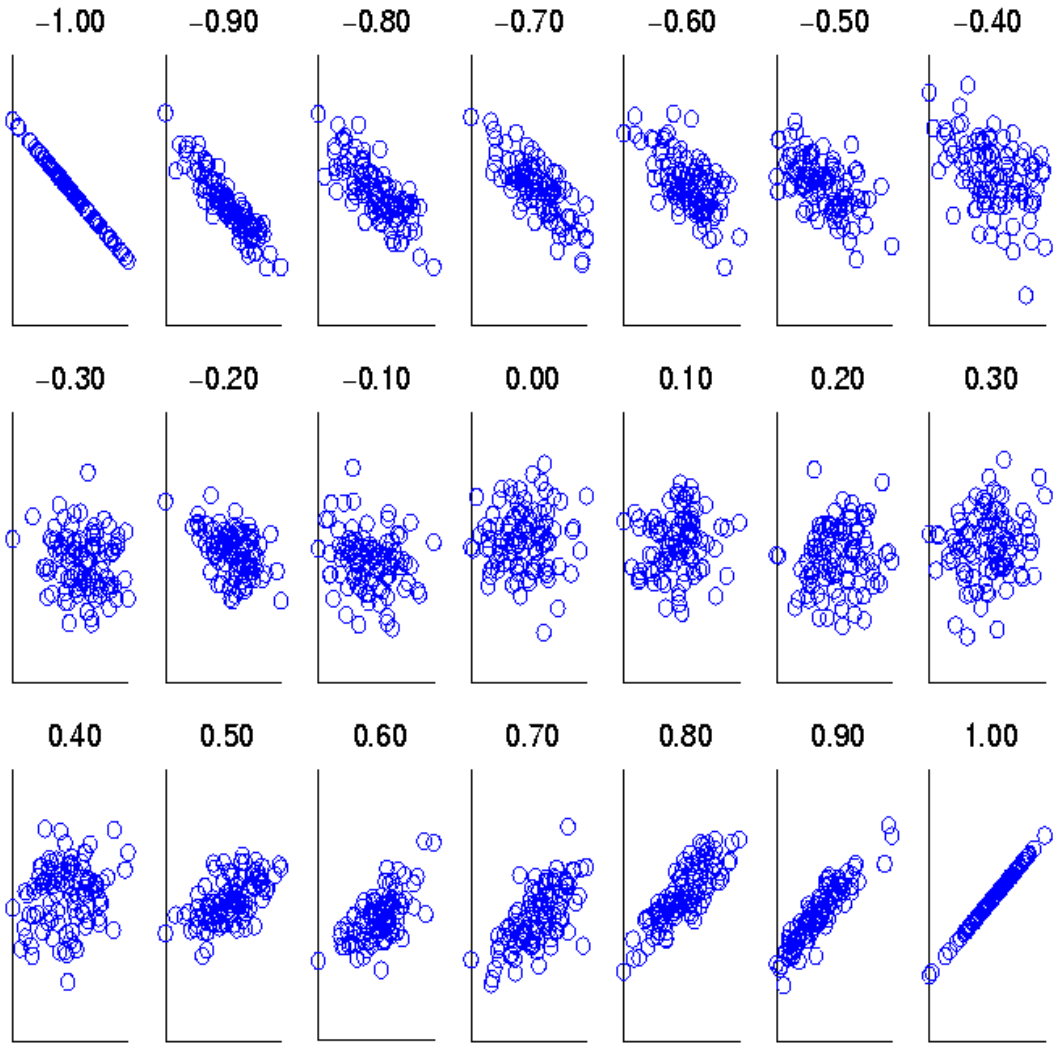
Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$
$$q'_k = (q_k - mean(q)) / std(q)$$
$$correlation(p, q) = p' \bullet q'$$

즉, $Correlation = \frac{Cov(p,q)}{stdDev(p) * stdDev(q)}$

Visually Evaluating Correlation



-1에서 1의 상관도를 보여주는 산포도(scatter plots)

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.

2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$

Density

- Density-based clustering require a notion of density

- Examples:
  - Euclidean density
    - Euclidean density = number of points per unit volume
  - Probability density
  - Graph-based density

Euclidean Density – Cell-based

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains

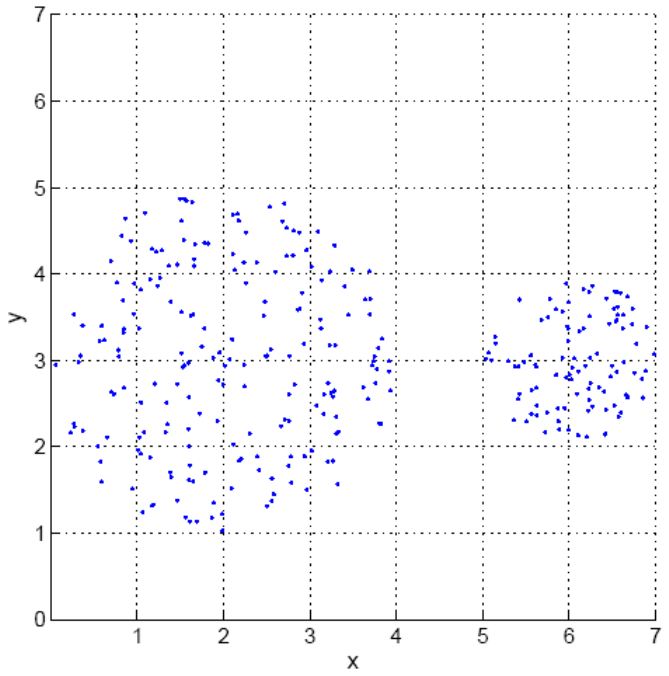


Figure 7.13. Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Table 7.6. Point counts for each grid cell.

Euclidean Density – Center-based

- Euclidean density is the number of points within a specified radius of the point

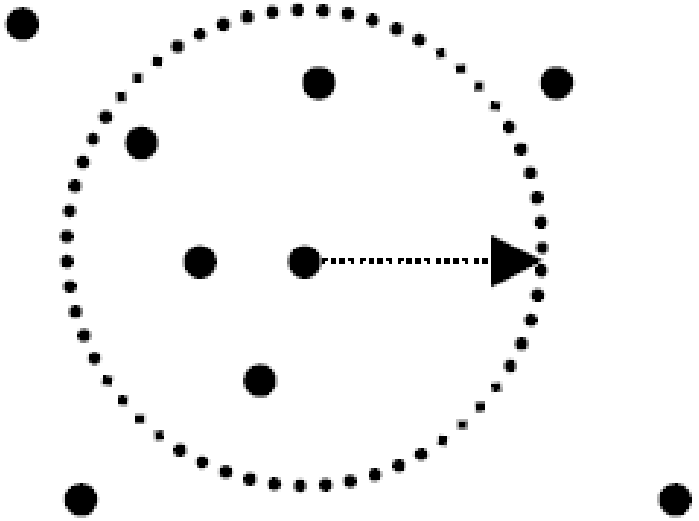


Figure 7.14. Illustration of center-based density.