Contents

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- **Measuring Data Similarity and Dissimilarity**

Similarity and Dissimilarity

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

Data Matrix and Dissimilarity Matrix

- **Data matrix**
  - n data points with p dimensions

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- **Dissimilarity matrix**
  - n data points, but registers only the distance
  - A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Proximity Measure for Nominal Attributes

- Nominal attributes can take 2 or more states, e.g., red, yellow, blue (generalization of a binary attribute)

Method 1: Simple matching

- m: # of matches, p: total # of variables

$$d(i,j) = \frac{p - m}{p}$$

Method 2: Use a large number of binary attributes

- creating a new binary attribute for each of the M nominal states

Color
red
yellow
blue
yellow

Color_red	Color_yellow	Color_blue
1	0	0
0	1	0
0	0	1
0	1	0

Proximity Measure for Binary Attributes

A **contingency table** (분할표) for binary data

Object *j*

Object *i*

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Distance measure for **symmetric** binary attributes: $$d(i, j) = \frac{r+s}{q+r+s+t}$$
- Distance measure for **asymmetric** binary attributes: $$d(i, j) = \frac{r+s}{q+r+s}$$
- **Jaccard coefficient** (similarity measure for asymmetric binary variables): $$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$

Dissimilarity between Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute, and the remaining attributes are asymmetric binary
- Let's calculate dissimilarity for the **asymmetric** binary attributes
- Let the values Y and P be 1, and the value N 0

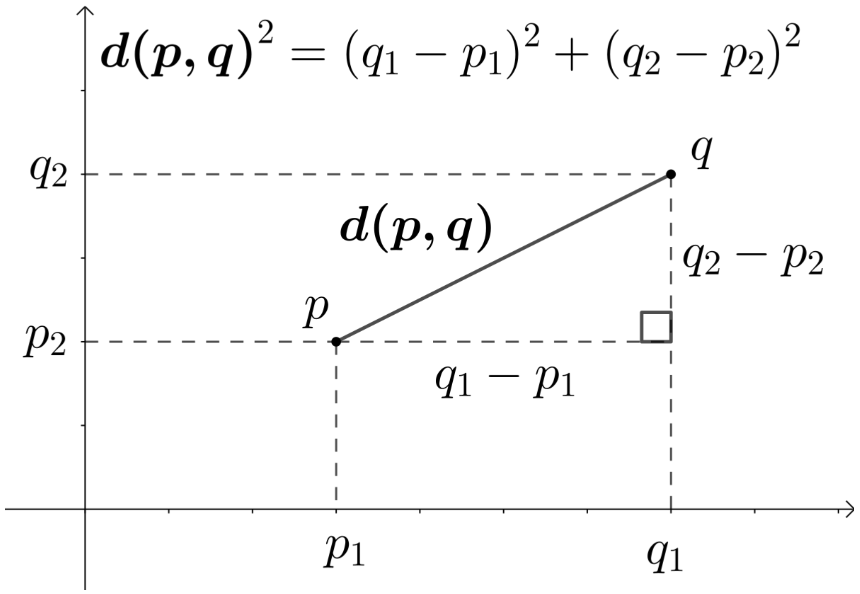
$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

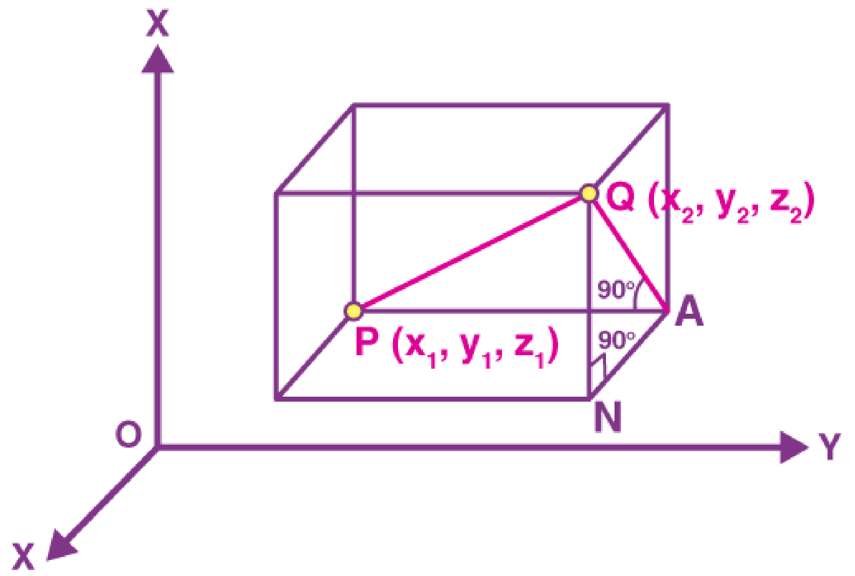
$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Dissimilarity on Numeric Data

- Euclidean distance (유클리드 거리)





$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

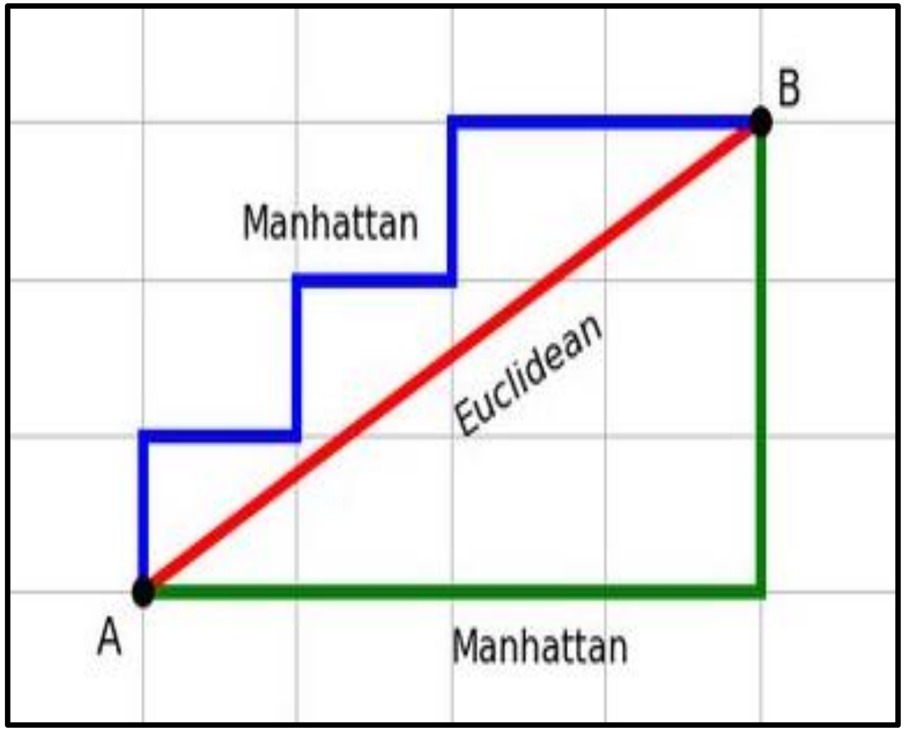
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

Source: Wikipedia, Byju's

Dissimilarity on Numeric Data

- Manhattan distance (맨하탄 거리)

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$



Source: Omni Calculator, Naver

Distance on Numeric Data

- **Minkowski distance**: a general form of distance measure for numeric data

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \cdots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L_h norm)

- Properties
  - $d(i,j) > 0$ if $i \neq j$, and $d(i,i) = 0$ (Positive definiteness)
  - $d(i,j) = d(j,i)$ (Symmetry)
  - $d(i,j) \leq d(i,k) + d(k,j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

- $h = 1$; **Manhattan** (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

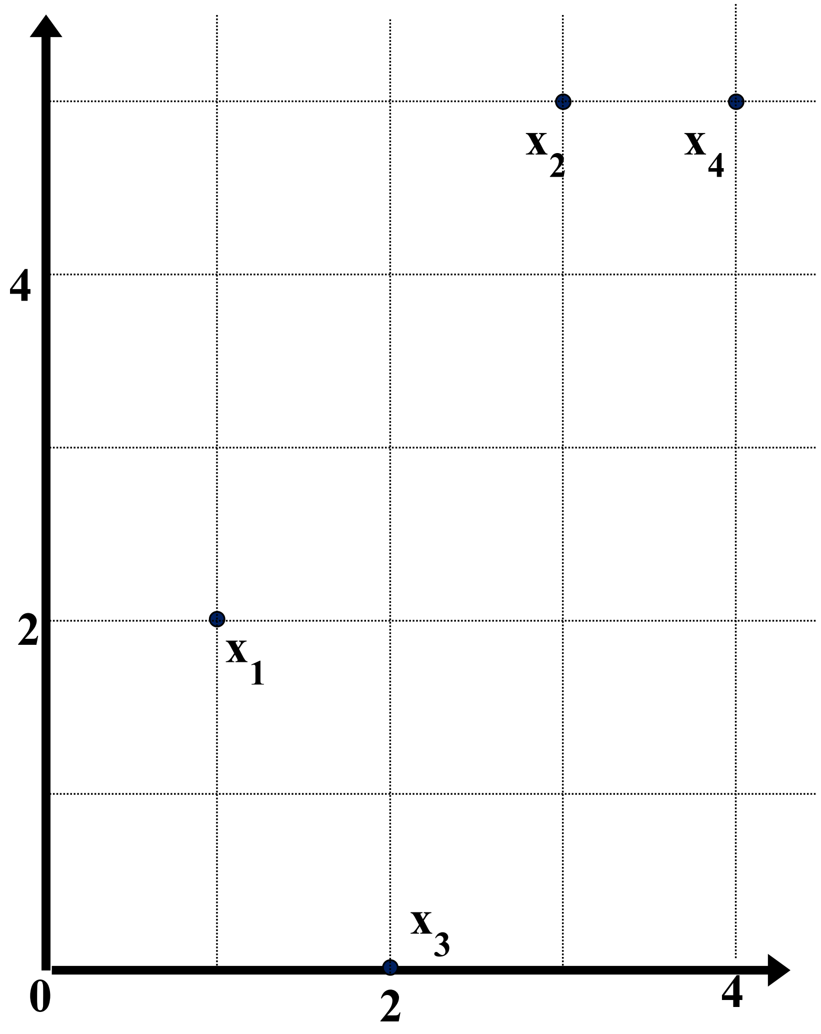
- $h = 2$; (L_2 norm) **Euclidean** distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$; "**supremum**" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

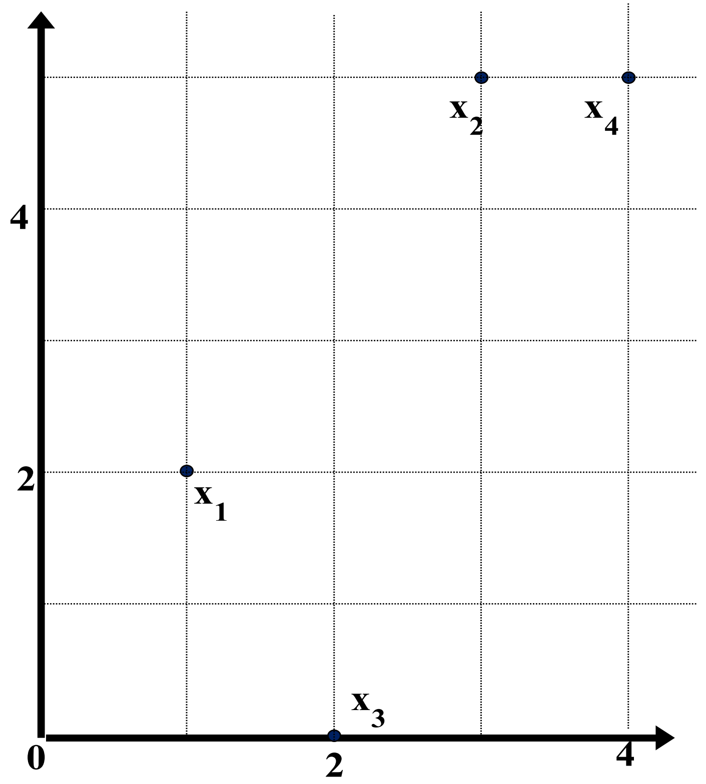
Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Numeric Data normalization

- Z-score (standardization) $z = \frac{x - \mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - negative when the raw score is below the mean, "+" when above

- Min-max normalization $x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$

- Recommended to normalize (standardize) numeric attributes for measuring distance using multiple attributes

Distance on Ordinal Variables

- An ordinal variable can be discrete or continuous
- Since rank is important, ordinal variables can be treated like interval-scaled attributes
  - replace x_{if} by their rank r_{if}

$$r_{if} \in \{1, \ldots, M_f\}$$

  - map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Ex.

Low
Medium
Medium
High

➡

1
0.5
0.5
0

Attributes of Mixed Type

- A dataset may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- We use the (weighted) mean to measure dissimilarity based on mixed attributes
  - f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all non-missing objects for attribute f
  - f is ordinal: compute ranks r_{if} and treat z_{if} as interval-scaled

$$d(i,j) = \frac{\sum_{f=1}^{p} d_{ij}^{(f)}}{p}$$

가중치

$$d(i,j) = \frac{\sum_{f=1}^{p} \boxed{\delta_{ij}^{(f)}} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

An example of multi-attribute distance

Object ID	Attribute 1	Attribute 2	Attribute 3	Attribute 4
X1	1	2	M	Low
X2	3	5	F	Medium
X3	2	0	F	Medium
X4	4	5	M	High

Normalization

Transformation

Object ID	Attribute 1	Attribute 2	Attribute 3	Attribute 4
X1	0	0.4	M	1
X2	0.67	1	F	0.5
X3	0.33	0	F	0.5
X4	1	1	M	0

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

if $\delta^{(1)} = \delta^{(2)} = \delta^{(3)} = \delta^{(4)}$

$d^{(1)}$	X1	X2	X3	X4
X1	0			
X2	0.67	0		
X3	0.33	0.34	0	
X4	1	0.33	0.67	0

$d^{(2)}$	X1	X2	X3	X4
X1	0			
X2	0.6	0		
X3	0.4	1	0	
X4	0.6	0	1	0

$d^{(3)}$	X1	X2	X3	X4
X1	0			
X2	1	0		
X3	1	0	0	
X4	0	1	1	0

$d^{(4)}$	X1	X2	X3	X4
X1	0			
X2	0.5	0		
X3	0.5	0	0	
X4	1	0.5	0.5	0

$d^{(4)}$	X1	X2	X3	X4
X1	0			
X2	0.69	0		
X3	0.56	0.34	0	
X4	0.65	0.46	0.79	0

Summary

- A dataset consists of data objects, which are described by attributes
  - Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Gain insight into the data by
  - Basic statistical description: central tendency, dispersion, graphical displays
  - Data visualization
  - Measure data similarity
- The above steps help know the data better, allowing for effective knowledge discovery in the later KD process