Chapter 2: Getting to Know Your Data

Dong-Kyu Chae

PI of the Data Intelligence Lab @HYU
Department of Computer Science & Data Science
Hanyang University

Data
Intelligence
LAB

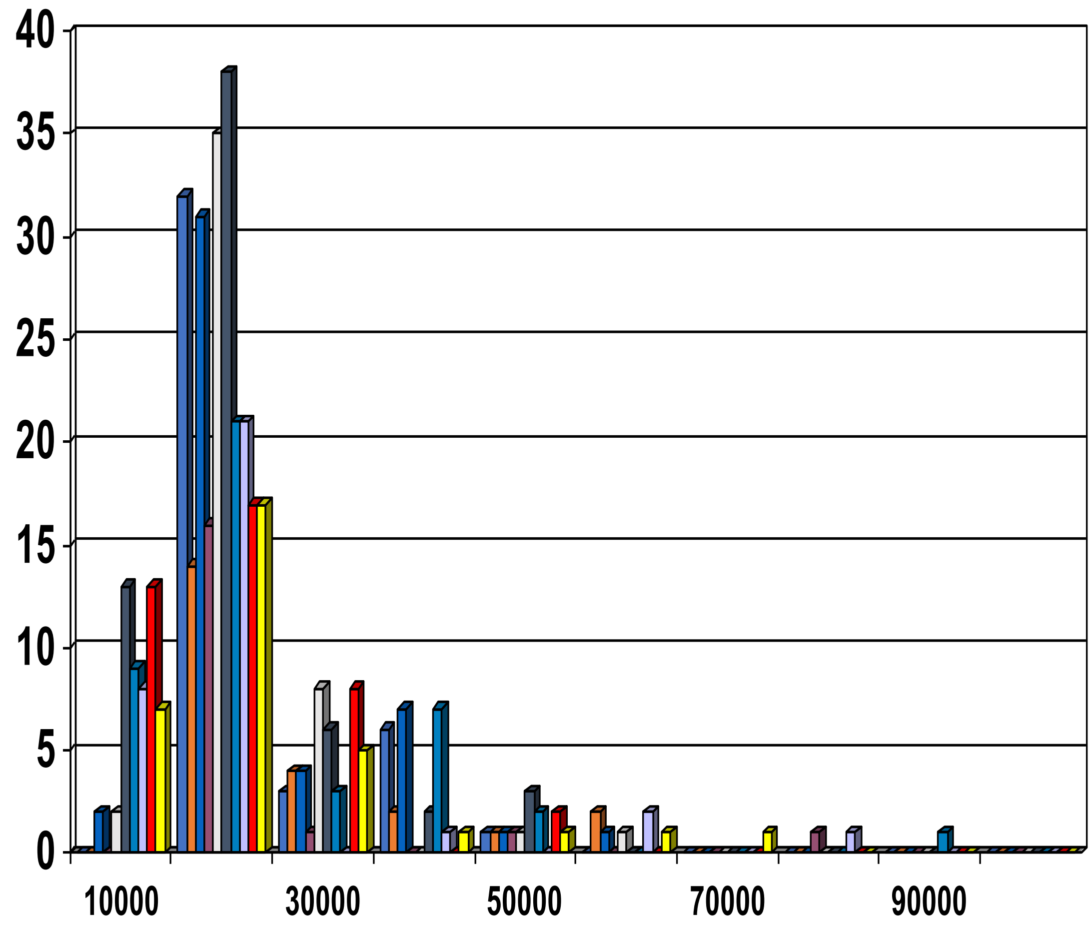
Contents

Graphic Displays of Basic Statistical Descriptions

- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis: values/ranges, y-axis: frequencies
- **Quantile plot:** each value x_i is paired with f_i indicating that approximately f_i of data are $\leq x_i$
- **Quantile-quantile (q-q) plot:**
  The quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot:** each pair of two feature values is plotted as points in the 2D space

Histogram Analysis

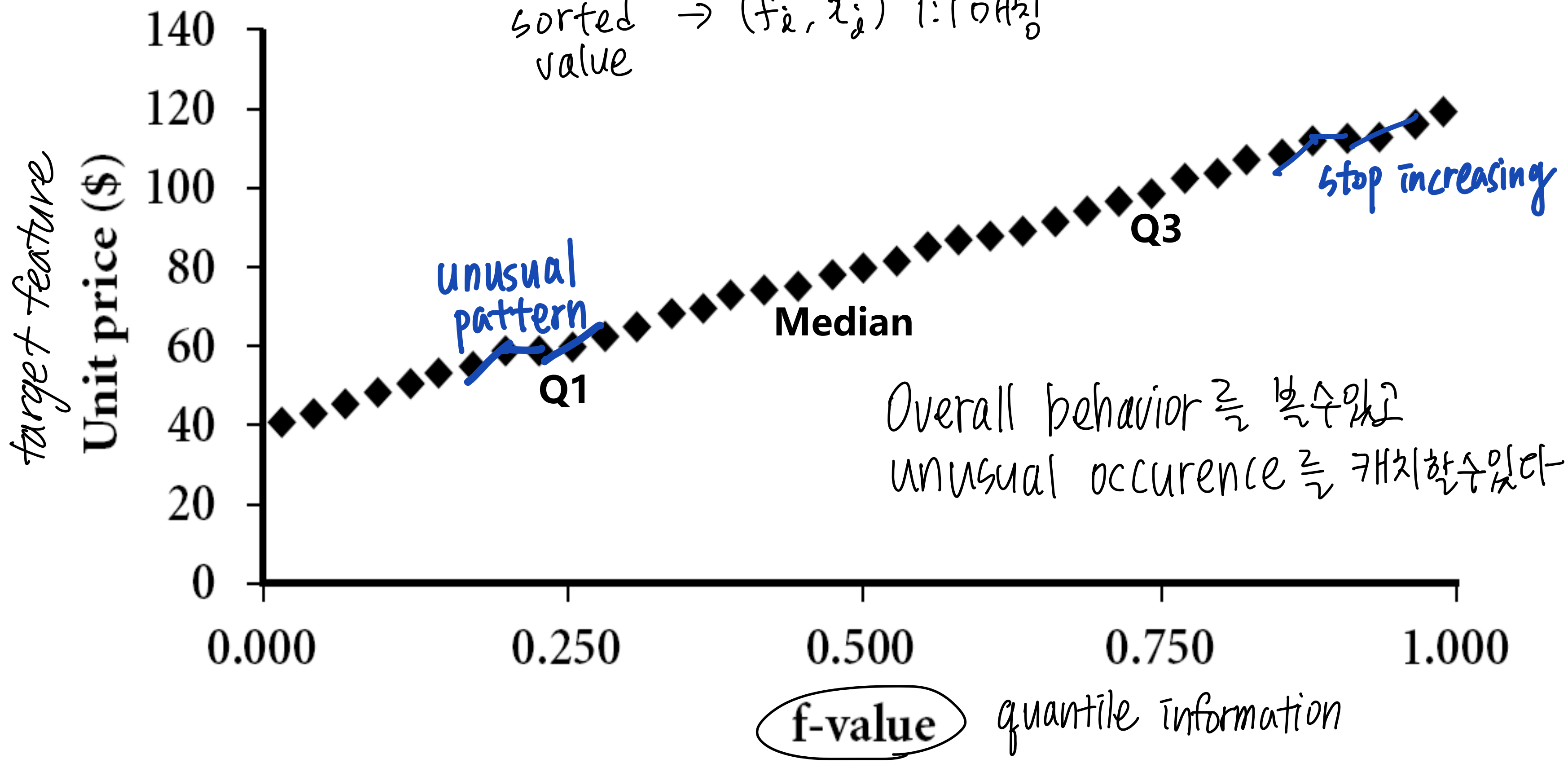
- **Histogram:** Graph display of frequencies shown as bars
- **It shows what proportion of cases fall into each of several categories**
  - The categories are usually specified as non-overlapping intervals of some variable
  - The categories (bars) must be adjacent



Quantile Plot

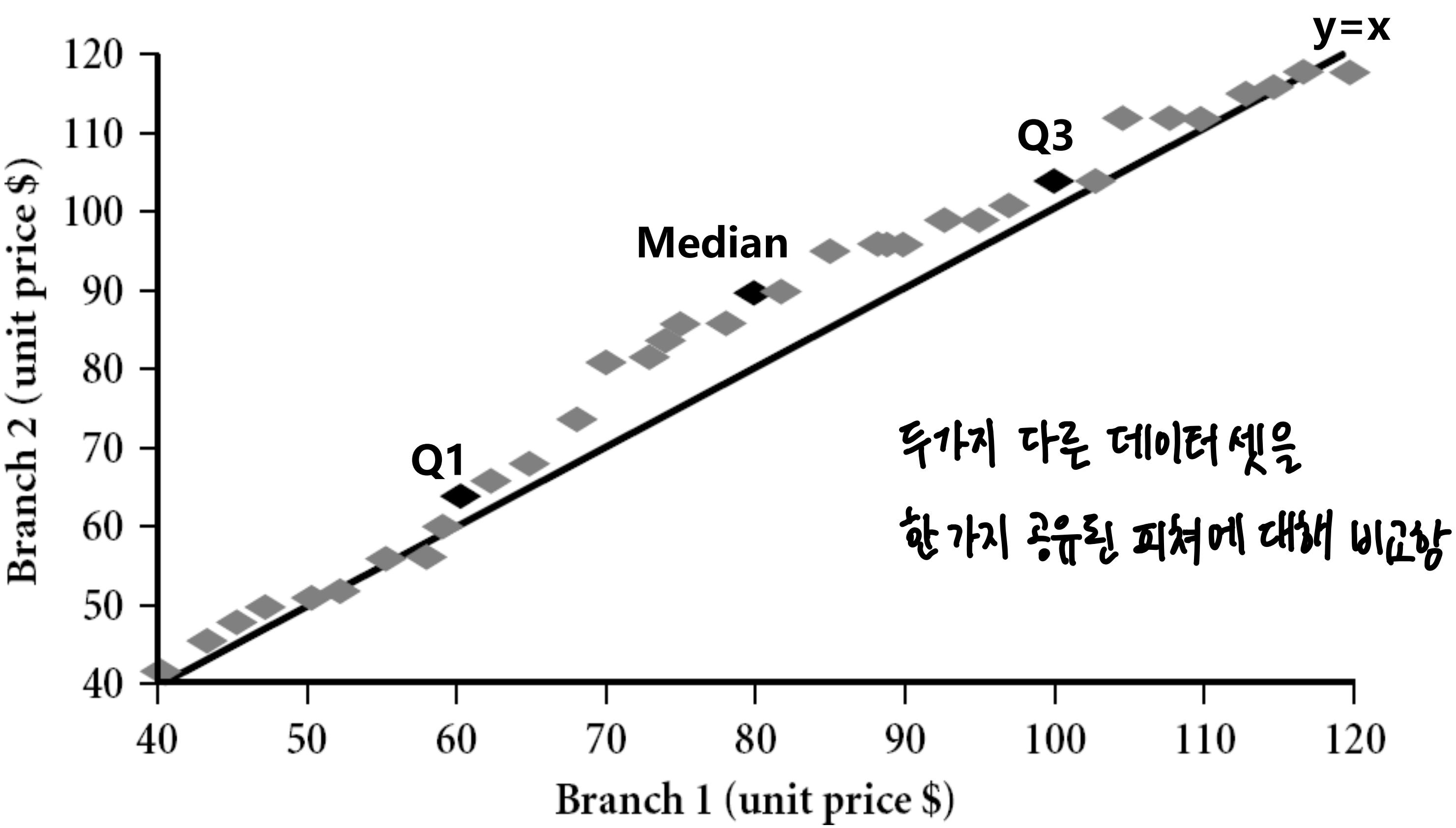
- **Sort data in increasing order, and display all the data points**
  - Allowing users to assess both the overall behavior and unusual occurrences
  - f_i indicates that approximately (100 x f_i)% of the data are below or equal to the value x_i

sorted value → (f_i, x_i) 1:1 매칭



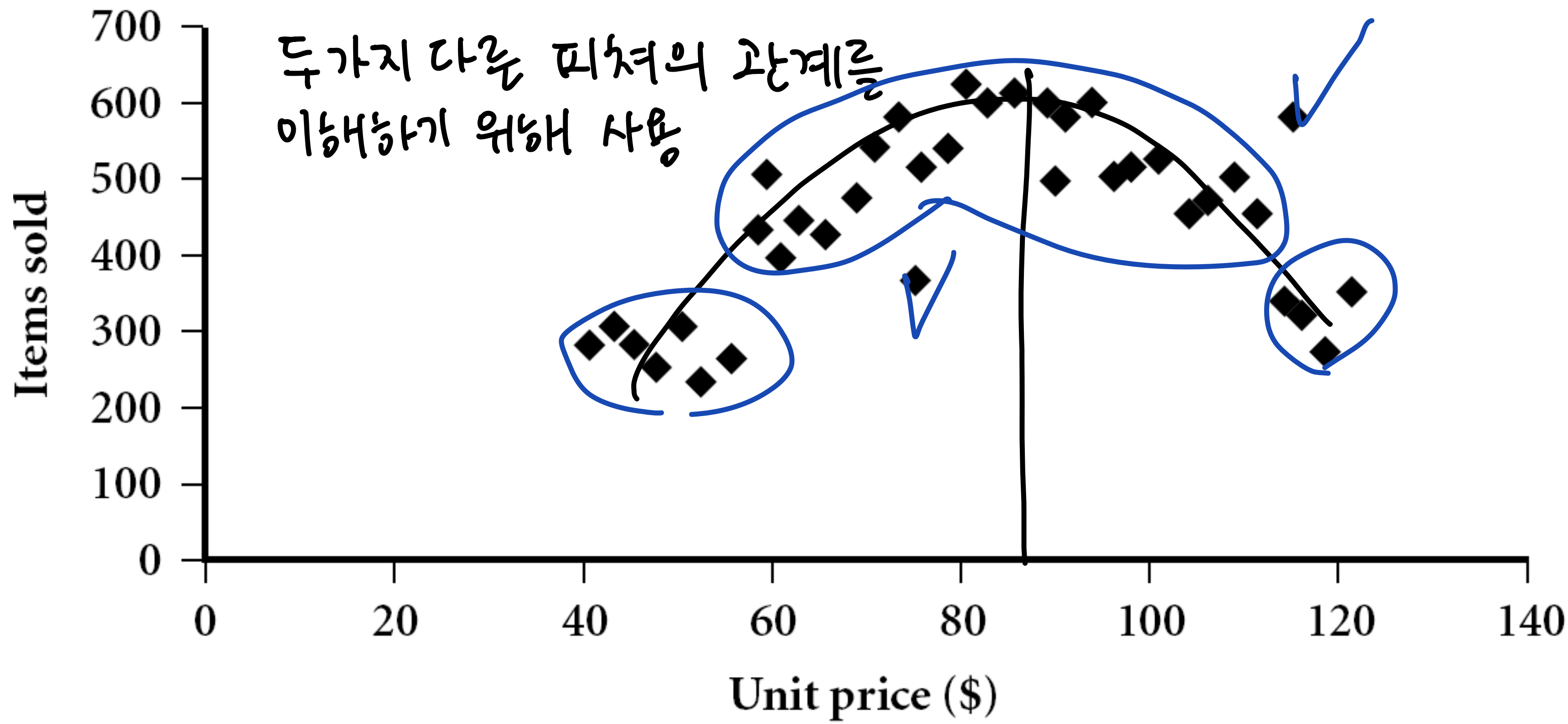
Quantile-Quantile (Q-Q) Plot

- **Displays the quantiles of one univariate distribution of a dataset against the corresponding quantiles of another dataset**
  - Example: Shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile.
  - Unit prices of items sold at Branch 1 tend to be cheaper than those at Branch 2.



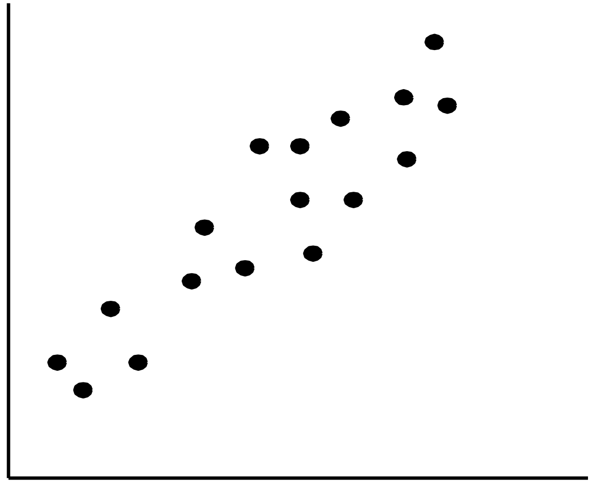
Scatter Plot

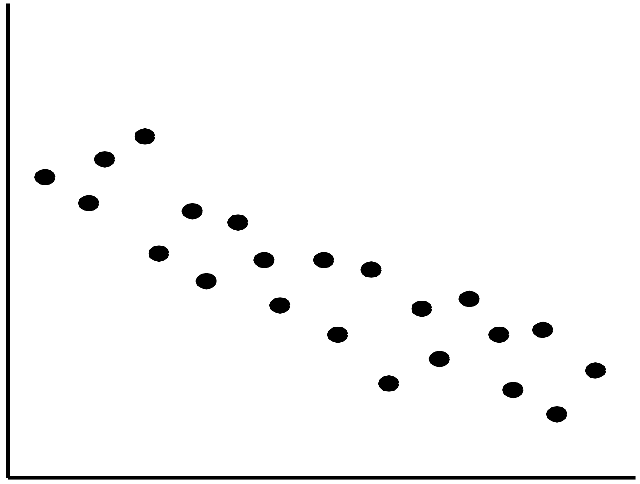
- Each pair of feature values is treated as the coordinates and plotted as a point in the 2D space
- It shows correlation between two features
- Also provides an overview to see clusters of points, outliers, etc



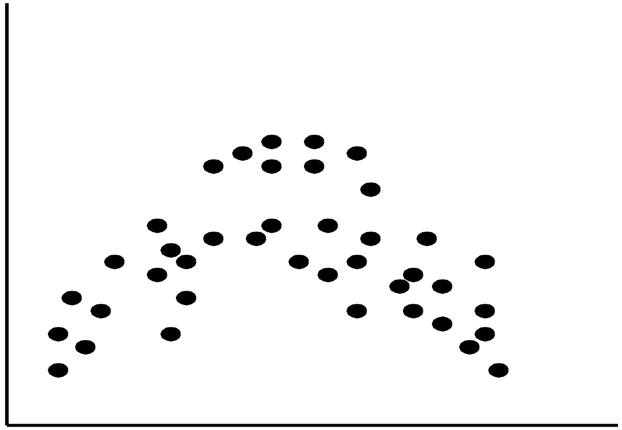
Scatter Plot

- Examples of positively and negatively correlated data



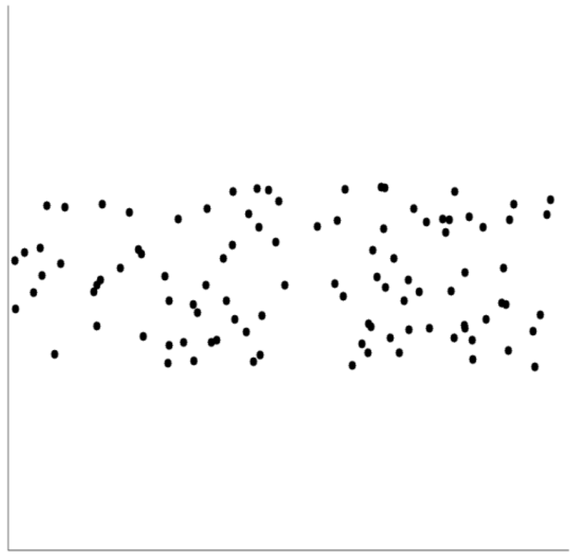


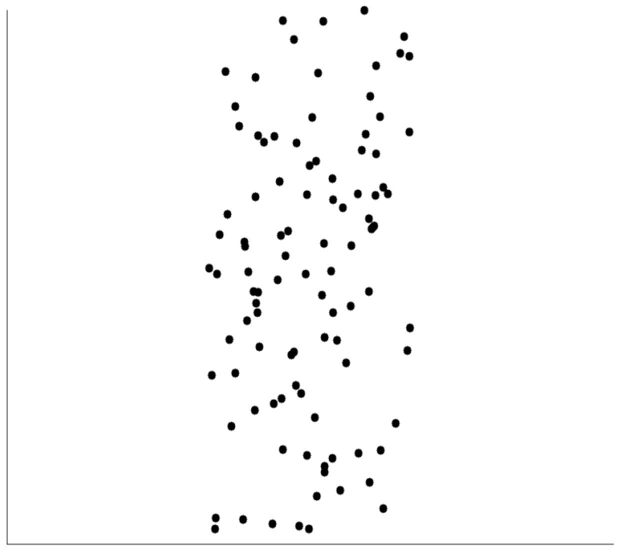
- In the example below, the left half fragment is positively correlated, and the right half is negative correlated



Scatter Plot

❑ Uncorrelated data





Contents

Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how much alike **two data objects** are
  - This value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity (e.g., distance)**
  - Numerical measure of how much different **two data objects** are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0 $[0,\infty]$
- **Proximity refers to a similarity or dissimilarity**

Matrix for Data

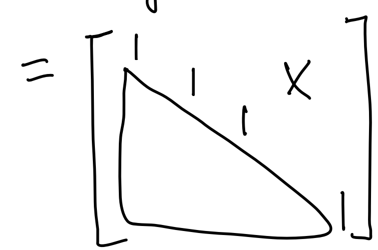
❑ Data matrix (n-by-p)

- ❑ n data points with p dimensions (features)
- ❑ Two modes

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

❑ Dissimilarity (distance) matrix (n-by-n)

Similarity matrix



- ❑ n data points, but registers only the distance
- ❑ A triangular matrix
- ❑ Single mode

dissimilarity (or distance)
is symmetric

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

Proximity Measure for Nominal Features

- **Nominal features can take 2 or more states**
  - E.g., COLOR: [red, yellow, blue, green, etc…]
- **Method 1: Simple matching**
  - m : # of matches, p : total # of nominal features

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

- **Method 2: Use multiple binary features to express one nominal feature**
  - Creating a new binary feature for each of the M nominal states
  - E.g., a nominal feature **COLOR**: [red, yellow, blue, green] can be expressed by four binary features:

red => [0, 1]; **yellow** => [0, 1]; **blue** => [0, 1]; **green** => [0, 1]

Proximity Measure for Binary Features

- **A contingency (분할표) table for two different objects:**

		Obj j		
		1	0	sum
Obj i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p — 전체 binary feature 개수

- **Proximity measure for symmetric binary features:**

$$d(i, j) = \frac{r+s}{q+r+s+t = p}$$ # of mismatch features

$$sim(i,j) = \frac{q+t}{q+r+s+t = p}$$ # of match features

  - Used when two results (0, 1) are equally important (similar with Method 1)

- **Proximity measure for asymmetric binary features :**

t (# of match on 0 which is out of our interest) 가 없어짐

$$d(i, j) = \frac{r+s}{q+r+s = p-t}$$

$$sim(i,j) = \frac{q}{q+r+s}$$

ex) 암인지 아닌지.
Covid 인지아닌지

  - Used when one is much more important than the other.

1 0

Dissimilarity between Asymmetric Binary Features

Example

matches on 0

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

mismatch

- Gender is a symmetric feature (ignored in this case)
- Assume that the remaining features are **asymmetric**
- Let the values Y and P be 1, and the value N 0

much more interested

Not interested

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

$= p - t = 6 - $ (# of matches on uninterested value) $= 6 - 3 = 3$

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

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

Proximity Measure for Numeric Features

- **First of all, we need to $\boxed{\text{standardize / normalize}}$ numeric data (will be introduced in the next chapter)**
  - To match scale of each feature

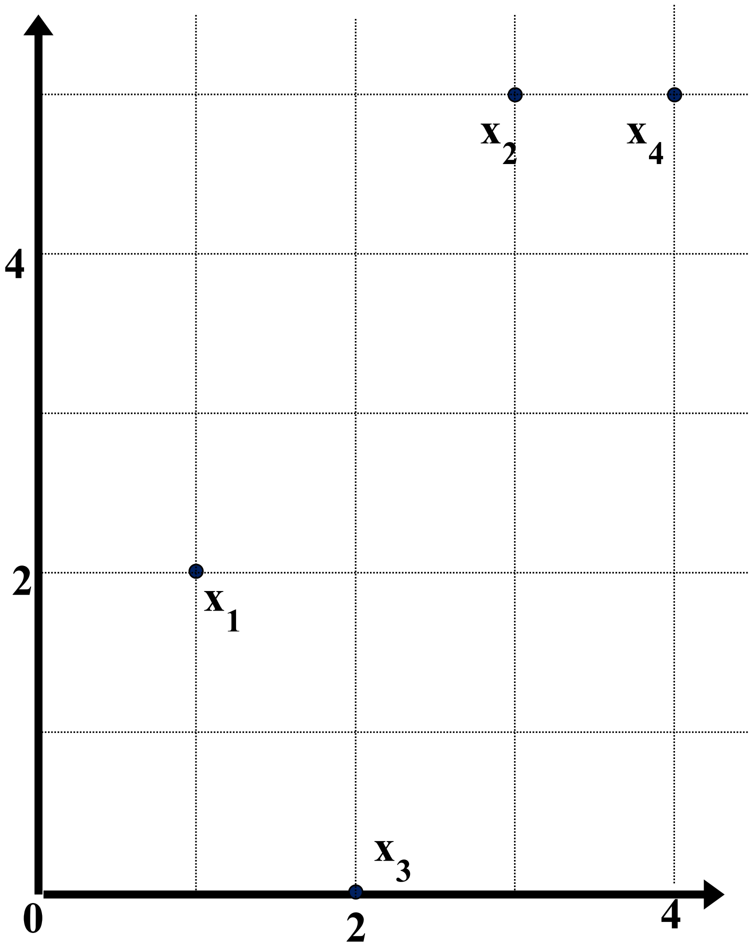
Numeric 피쳐는
각각 다른 scale 가지고 있으므로

$$z = \frac{x - \mu}{\sigma}$$

다음챕터

- Z-score: (normalized method)
  - x: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - Meaning: the distance between the raw score and the population mean in units of the standard deviation
    - "–" when the raw score is below the mean
    - "+" when the raw score is above the mean

Matrix for Numeric Data



Data Matrix

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

Distance 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

$$\mathrm{d(i,j)} = \sqrt{\sum_{f=1}^{p} \left(x_{if} - x_{jf}\right)^2}$$

Distance on Numeric Data: Minkowski Distance

Minkowski distance : A popular distance measure

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

h is hyperparameter

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

If **h=2**, it is equal to **Euclidean Distance (L-2 norm)**

Special Cases of Minkowski Distance

- **h = 1: Manhattan (city block, L_1 norm) distance**

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

- **h = 2: (L_2 norm) Euclidean distance**

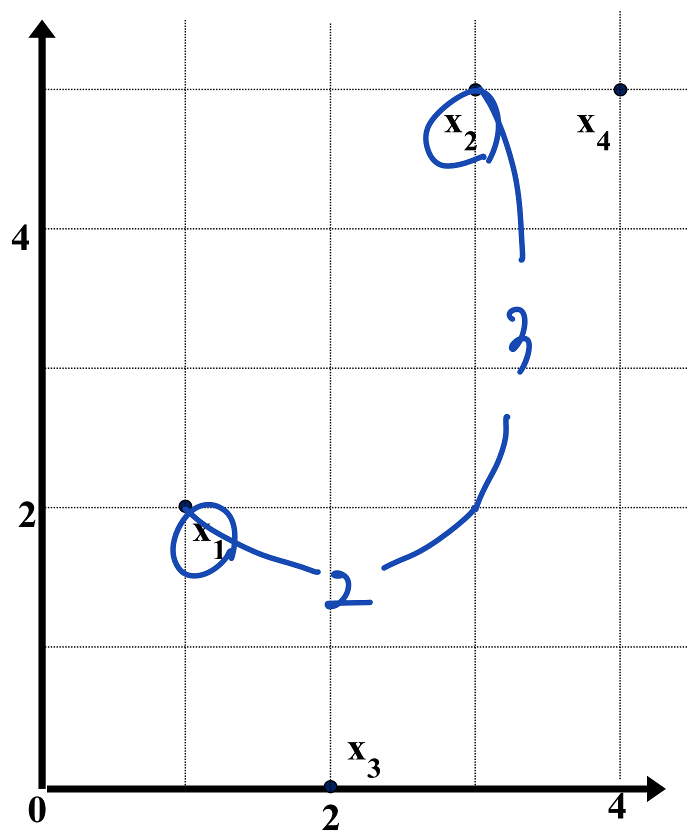
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- **$h \to \infty$. Supremum (L_{max} norm, L_∞ norm) distance**
  - This is the maximum difference between any component (feature) 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: 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

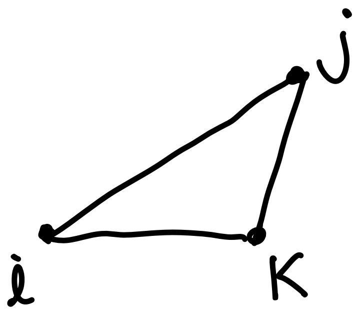
L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Distance on Numeric Data: Minkowski Distance

- **Minkowski distance : A popular distance measure**

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

- **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**)
    삼각 부등식



- **A distance that satisfies these properties is a *metric***

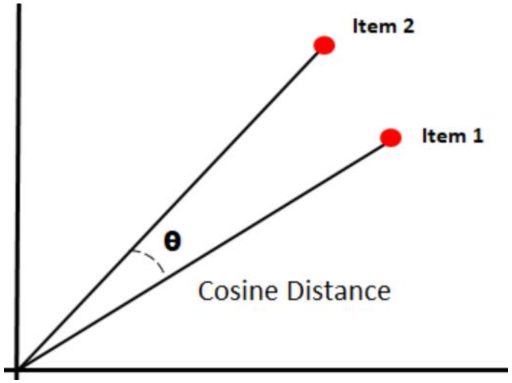
In other words, Metric is the distance function satisfying all these three properties
- Minkowski distance is Metric

Cosine Similarity

- A document can be represented by thousands of words, each recording the *frequency* of a particular word in the document.

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

  - Documents are represented as n-dimensional vectors



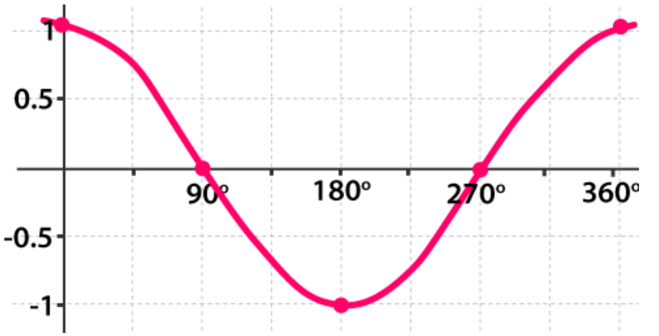
- **Cosine measure:** If *A* and *B* are two vectors, then:

Inner product $A \cdot B = \|A\| \|B\| \cos\theta$

$$\text{cosine similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

각도에 기반.

→ 두 데이터의 tendency 파악



where • indicates dot product, ||*A* || is the length of vector *A*

Example: Cosine Similarity

- **$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$**
  - It focuses more on **vector direction**, rather than coordinate distance
- **Ex: Find the similarity between documents 1 and 2.**

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = \mathbf{25}$$

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

$$= \mathbf{6.481}$$

$$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{\mathbf{0.5}} = (17)^{\mathbf{0.5}}$$

$$= \mathbf{4.12}$$

$$\cos(d_1, d_2) = \mathbf{0.94}$$

Distance for Ordinal Features

- **Order is important,** e.g., grade, size, etc…
- **Such values can be expressed by rank (integers)**
  - Replace *i*-th object in the *f*-th feature x_{if} by their *rank:* r_{if}

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

- **Finally the rank values are mapped onto [0, 1] by:**

Normalization $r_{if} \rightarrow$ $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$ start point를 0로 만들기 위해

M_f : highest rank value

- **Then we can use any distance/dissimilarity measures for numeric data**

Features of Mixed Type

- **A database may contain various feature types**
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- **One may use a weighted average to combine their effects**

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

  - $\delta_{ij}^{(f)}$: importance weight on each feature type f

Summary

- **Data feature types: nominal, binary, ordinal, numerical (interval-scaled, ratio-scaled)**
- **Gain insight into the data by:**
  - Basic statistical data description: central tendency, dispersion
  - 5 numbers summary, visualized by a boxplot.
- **Visualizations**
  - Scatter plot, QQ plot, histogram, etc…
- **Measure data similarity**
  - Minkowski Distance and its special cases
  - Cosine similarity
  - ….

Thank You