Knowledge Discovery & Data Mining
— Similarity and Distance Measures —

Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Recall: data cleaning and Integration

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
  - Missing Values,
  - Noise(Denoising): Binning, Regression, Low-pass filter
  - Outliers,
  - Data Cleaning as a Process
- Data Integration
  - Schema integration, Entity identification problem, Detect and resolve data value conflicts
  - Handling Redundancy: At the tuple level; Between attributes

Recall: data transformation

- Data Transformation
  - Transformation functions
  - Data normalization
    - Min-max
    - Z-score
    - Decimal scaling
  - Data discretization: Binning, Clustering analysis, Histogram analysis

Recall: data reduction

- Data compression
  - Discrete wavelet transform (DWT)
- Sampling
  - Sampling without replacement
  - Sampling with replacement
  - Cluster or Stratified Sampling

Outline

- Similarity and distance measures
  - Proximity Measures for
    - Nominal Attributes
    - Binary Attributes
    - Numeric Attributes
    - Ordinal attributes
    - Mixed types
  - Cosine Similarity
  - Kullback-Leibler divergence
  - Entropy & Cross Entropy

Similarity and distance measures

- Similarity
  - 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)
  - 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

Dissimilarity matrix

Data matrix (or object-by-attribute structure):

- This structure stores the n data points with p dimensions (n objects ×p attributes)
- Two-mode

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

Dissimilarity matrix (or object-by-object structure):

- A triangular matrix
- d(i, j) is the measured dissimilarity or “difference” between objects i and j
- d(i, j)>=0, close to 0 when objects i and j are highly similar or “near” each other, and becomes larger the more they differ.
- d(i, j) = d(j, i).
- One-mode

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

Proximity Measures for Nominal Attributes

Simple matching

- Dissimilarity: m: # of matches, p: total # of variables, then dissimilarity between two objects i and j can be computed based on the ratio of mismatches

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

- Similarity: $$sim(i,j) = 1 - d(i,j) = \frac{m}{p}$$

Encoding: creating a new binary attribute for each of the M states of a nominal attribute.

Proximity Measures for Nominal Attributes

Example. Suppose that we have the sample data of following table, so the dissimilarity matrix is

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

Object Identifier	Test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

Proximity Measures for Nominal Attributes

Example. Suppose that we have the sample data of following table, so the dissimilarity matrix is

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

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

Object Identifier	Test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

Only one nominal attribute, so p = 1

Proximity Measures for Nominal Attributes

Example. Suppose that we have the sample data of following table, so the dissimilarity matrix is

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

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Object Identifier	Test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

Proximity Measures for Binary Attributes

If all binary attributes are thought of as having the same weight, we have the 2×2 contingency table, where q is the number of attributes that equal 1 for both objects i and j , r is the number of attributes that equal 1 for object i but equal 0 for object j , s is the number of attributes that equal 0 for object i but equal 1 for object j , and t is the number of attributes that equal 0 for both objects i and j . The total number of attributes is p, where p = q +r +s +t .

		Object j		
		1	**0**	**Sum (row)**
Object i	**1**	q	r	$q+r$
	0	s	t	$s+t$
	Sum(col.)	$q+s$	$r+t$	p

contingency table

Proximity Measures for Binary Attributes

Symmetric binary attributes: symmetric binary dissimilarity

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

Asymmetric binary attributes:

- the two states are not equally important,
- the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match).
- asymmetric binary dissimilarity

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

- asymmetric binary similarity:
  - is called the **Jaccard coefficient**

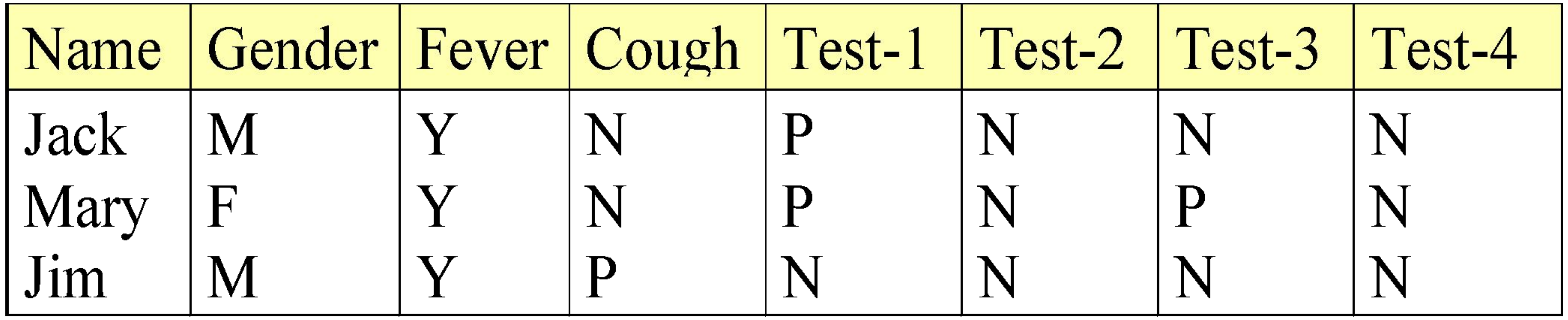
$$sim(i, j) = \frac{q}{q + r + s} = 1 - d(i, j)$$

		Object j		
		1	**0**	**Sum (row)**
Object i	**1**	*q*	*r*	***q+r***
	0	*s*	*t*	***s+t***
	Sum(col.)	***q+s***	***r+t***	***p***

contingency table

Proximity Measures for Binary Attributes

Example. Suppose that we have the sample data of following table, so the distance between each pair of the three patients—Jack, Mary, and Jim—is



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

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

- d(Jack, Jim) =
- d(Jack, Mary) =
- d(Jim, Mary) =

Object j

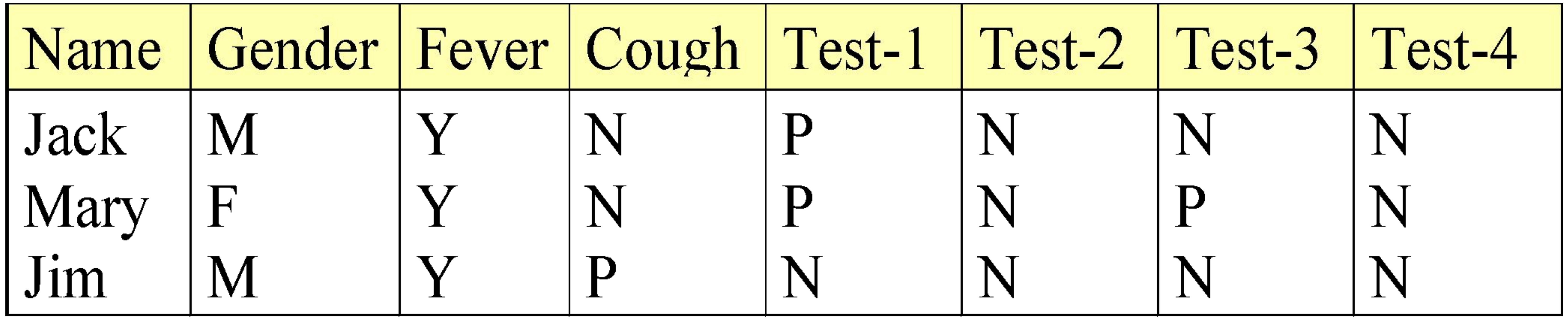
Object i

	1	0	Sum (row)
1	*q*	*r*	***q+r***
0	*s*	*t*	***s+t***
Sum(col.)	***q+s***	***r+t***	***p***

contingency table

Proximity Measures for Binary Attributes

Example. Suppose that we have the sample data of following table, so the distance between each pair of the three patients—Jack, Mary, and Jim—is



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

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

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$$

Object j

		1	0	Sum (row)
Object i	1	*q*	*r*	***q+r***
	0	*s*	*t*	***s+t***
	Sum(col.)	***q+s***	***r+t***	***p***

contingency table

Dissimilarity of numeric data: Minkowski distance

Distance measures are commonly used for computing the dissimilarity of objects described by numeric attributes.

- **Euclidean distance:** The most popular distance measure
  - Let i = (xi1, xi2, . . . , xip) and j = (xj1, xj2, . . . , xjp) be two objects described by p numeric attributes.
  - The Euclidean distance between objects i and j is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- **Manhattan (or city block) distance:**

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

Dissimilarity of numeric data: Minkowski distance

Both the Euclidean and the Manhattan distance satisfy the following mathematical properties:

- **Nonnegativity:** d(i, j) ≥ 0: Distance is a nonnegative number.
- **Identity of indiscernibles:** d(i, i) = 0: The distance of an object to itself is 0.
- **Symmetry:** d(i, j) = d(j, i): Distance is a symmetric function.
- **Triangle inequality:** d(i, j) ≤ d(i, k)+d(k, j): Going directly from object i to object j in space is no more than making a detour over any other object k.

A measure that satisfies these conditions is known as **metric**.

Dissimilarity of numeric data: Minkowski distance

- **Minkowski distance:** is a generalization of the Euclidean and Manhattan distances. It is defined as

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

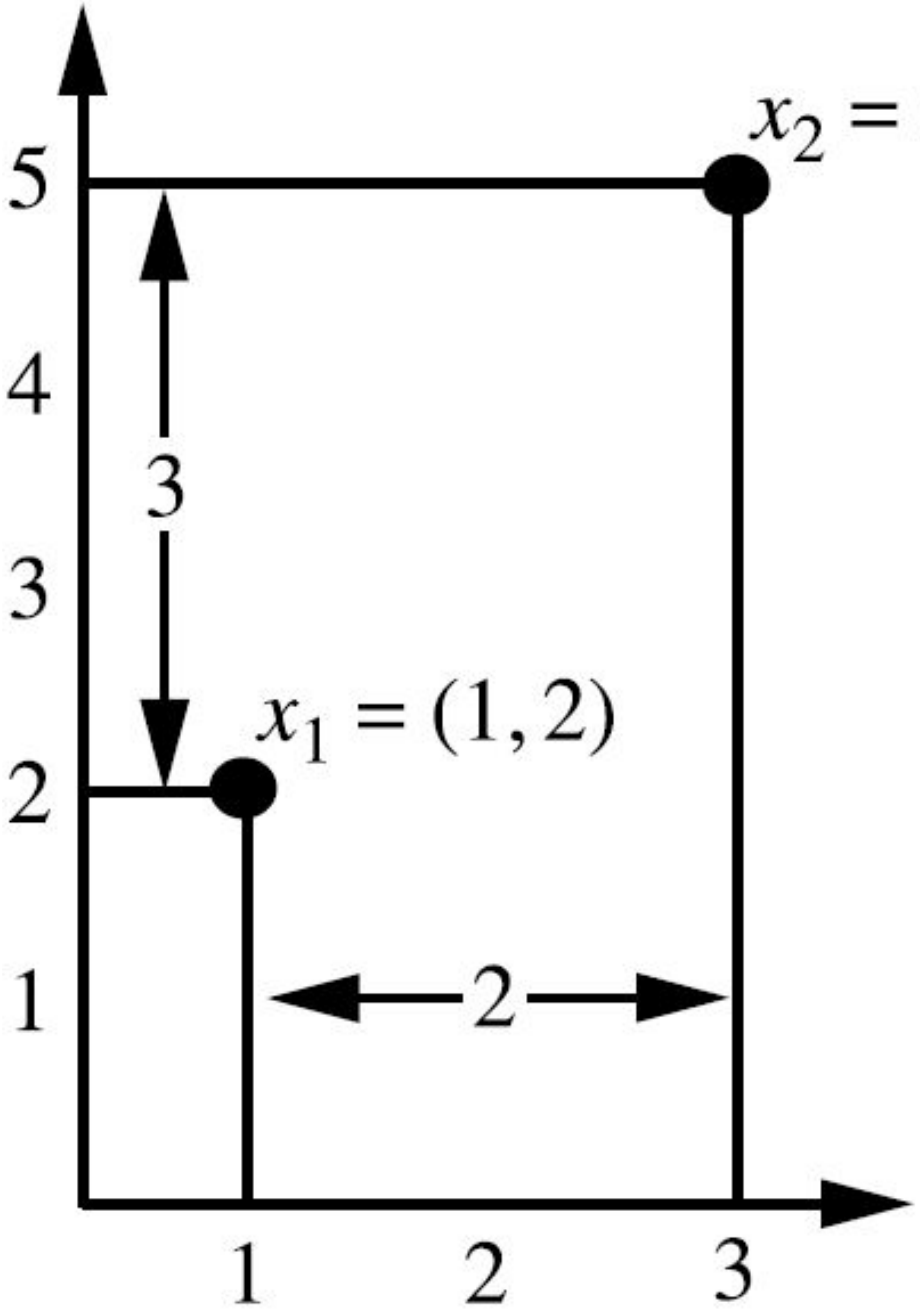
  - where h is a real number such that h ≥ 1
    - Manhattan distance when h = 1 (L1 norm)
    - Euclidean distance when h = 2 (L2 norm)

- **Supremum distance (Lmax, L∞ norm, and the Chebyshev distance):** a generalization of the Minkowski distance for h→∞

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

Dissimilarity of numeric data: Minkowski distance

Example. Let x1 = (1, 2) and x2 = (3, 5) represent two objects as shown



- **Euclidean distance:**

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- **Manhattan distance:**

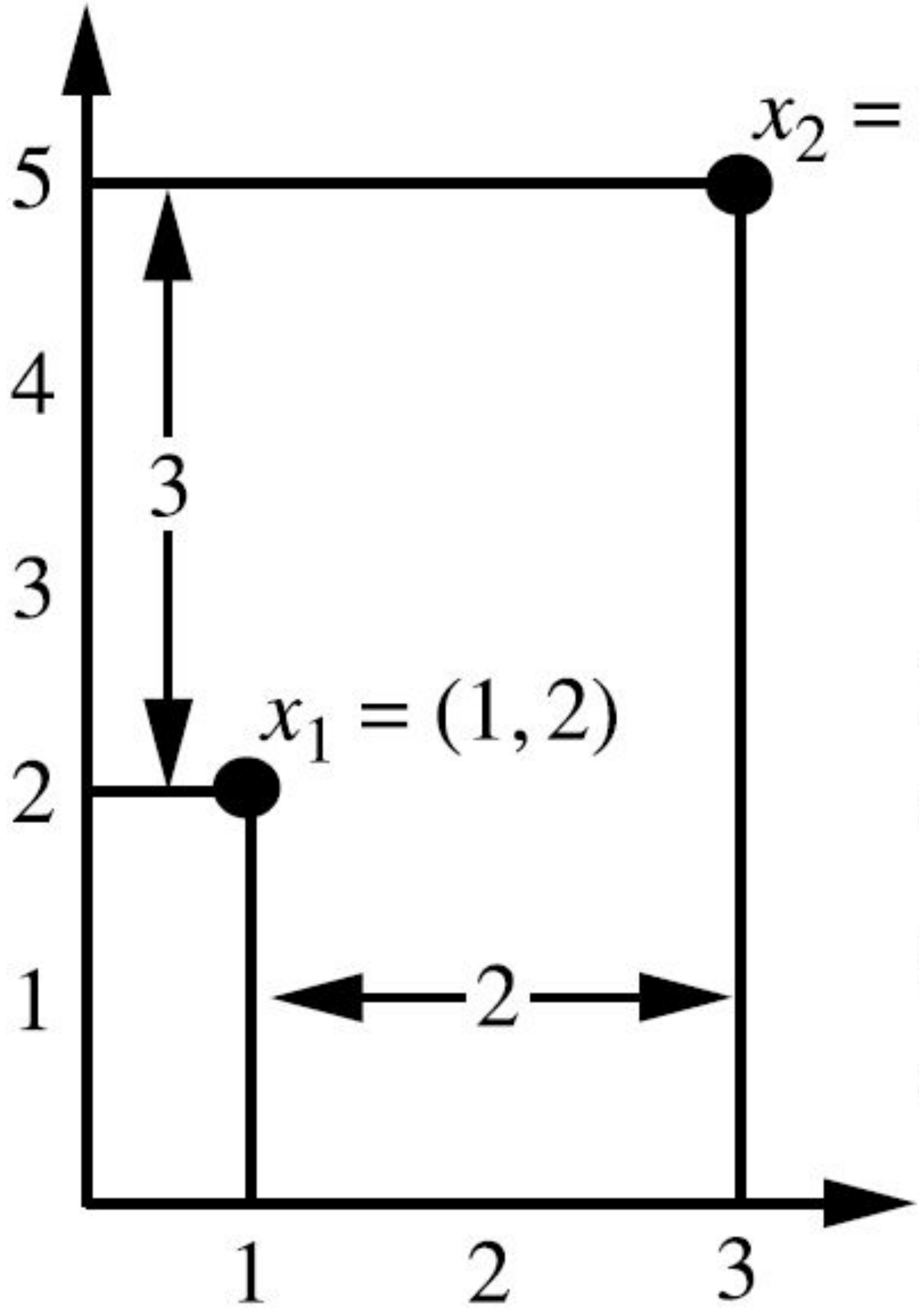
$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- **Supremum distance:**

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

Dissimilarity of numeric data: Minkowski distance

Example. Let x1 = (1, 2) and x2 = (3, 5) represent two objects as shown



Euclidean distance
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance
$= 2 + 3 = 5$

Supremum distance
$= 5 - 2 = 3$

Proximity measures for ordinal attributes

The values of an ordinal attribute have a meaningful order or ranking about them. e.g. *drink_size* = {*small, medium, large*}

Suppose that f is an ordinal attribute and has M_f ordered states. Let 1, . . . , M_f represent ranking of these ordered states. The dissimilarity of f can be calculated by:

1.Normalize the rank *r_{if}* of the object i and attribute f by

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

2.**Compute the dissimilarity using distance methods**

Proximity measures for ordinal attributes

Example. Suppose that we have the sample data shown as follows. use the Euclidean distance, the dissimilarity matrix is?

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

Object Identifier	Test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

Proximity measures for ordinal attributes

Example. Suppose that we have the sample data shown as follows. use the Euclidean distance, the dissimilarity matrix is?

$Z_{1f} =$

$Z_{2f} =$

$Z_{3f} =$

$Z_{4f} =$

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

Object Identifier	Test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

Proximity measures for ordinal attributes

Example. Suppose that we have the sample data shown as follows. use the Euclidean distance, the dissimilarity matrix is?

$M_f = 3$, [fair, good, excellent] = [1,2,3]

$Z_{1f} = 1$

$Z_{2f} = 0$

$Z_{3f} = 0.5$

$Z_{4f} = 1$

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

Object Identifier	Test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

Proximity measures for ordinal attributes

Example. Suppose that we have the sample data shown as follows. use the Euclidean distance, the dissimilarity matrix is?

$M_f = 3$, [fair, good, excellent] = [1,2,3]

$z_{1f} = 1$

$z_{2f} = 0$

$z_{3f} = 0.5$

$z_{4f} = 1$

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

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Object Identifier	Test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

Dissimilarity for attributes of mixed types

A database may contain all attribute types

- Nominal, symmetric binary, asymmetric binary, numeric, ordinal

Suppose that the data set contains p attributes of mixed types. The dissimilarity d(i, j) between objects i and j is defined as

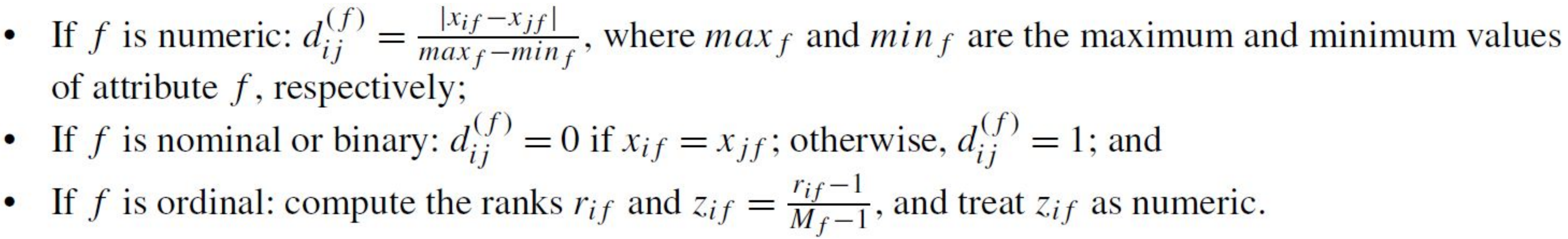
$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

where the indicator $\delta_{ij}^{(f)}$ = 0 if

1. x_{if} or x_{jf} is missing (i.e., there is no measurement of attribute f for object i or object j),
2. $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary;
3. otherwise, $\delta_{ij}^{(f)} =1$.

Dissimilarity for attributes of mixed types

The contribution of attribute f to the dissimilarity between i and j (i.e., $d_{ij}^{(f)}$) is computed dependent on its type:



- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_f - min_f}$, where max_f and min_f are the maximum and minimum values of attribute f, respectively;
- If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$; and
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as numeric.

Dissimilarity for attributes of mixed types

Example. compute a dissimilarity matrix for the objects in following table

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_f - min_f}$, where max_f and min_f are the maximum and minimum values of attribute f, respectively;
- If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$; and
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as numeric.

Dissimilarity for attributes of mixed types

Example. compute a dissimilarity matrix for the objects in following table

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3(numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$d_{ij}^{(1)} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity for attributes of mixed types

Example. compute a dissimilarity matrix for the objects in following table

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3(numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$d_{ij}^{(1)} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad d_{ij}^{(2)} = \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Dissimilarity for attributes of mixed types

Example. compute a dissimilarity matrix for the objects in following table

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3(numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$d_{ij}^{(1)} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad d_{ij}^{(2)} = \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix} \quad d_{ij}^{(3)} = \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Dissimilarity for attributes of mixed types

Example. compute a dissimilarity matrix for the objects in following table

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3(numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$d_{ij}^{(1)} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad d_{ij}^{(2)} = \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix} \quad d_{ij}^{(3)} = \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

$$\delta_{ij}^{(f)} = 1, \quad d(3,1) = \frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65$$

Dissimilarity for attributes of mixed types

Example. compute a dissimilarity matrix for the objects in following table

Object Identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3(numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$d(i, j) = \begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

Cosine Similarity

Cosine similarity: measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.

- Often used to measure document similarity in text analysis.
- A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document. Thus each document is an object represented by what is called a **term-frequency vector**.

Let x and y be two term-frequency vectors for comparison. Using the cosine measure as a similarity function, we have

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$$

where $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$.

||*y*|| is the Euclidean norm of vector *y*

Cosine Similarity

Example. Suppose that x and y are the first two term-frequency vectors in the following table, That is, x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) and y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1).

How similar are x and y?

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||}$$

Document vector or term-frequency vector.

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss	Season
1	5	0	3	0	2	0	0	2	0	0
2	3	0	2	0	1	1	0	1	0	1
3	0	7	0	2	1	0	0	3	0	0
4	0	1	0	0	1	2	2	0	3	0

Cosine Similarity

Example. Suppose that x and y are the first two term-frequency vectors in the following table, That is, x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) and y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1).

How similar are x and y?

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$$

$$\begin{aligned} \boldsymbol{x} \cdot \boldsymbol{y} = {} & 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 \\ & + 0 \times 0 + 0 \times 1 = 25 \end{aligned}$$

$$||\boldsymbol{x}|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$||\boldsymbol{y}|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

Cosine Similarity

Example. Suppose that x and y are the first two term-frequency vectors in the following table, That is, x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) and y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1).

How similar are x and y?

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$$

$$\begin{aligned}\boldsymbol{x} \cdot \boldsymbol{y} &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 \\ &\quad + 0 \times 0 + 0 \times 1 = 25\end{aligned}$$

$$||\boldsymbol{x}|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$||\boldsymbol{y}|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

sim(x, y) = 0.94

Kullback-Leibler divergence

Kullback-Leibler divergence(the KL divergence): a measure that has been popularly used in the data mining literature to measure the difference between two probability distributions over the same variable x.

- closely related to relative entropy, information divergence, and information for discrimination
- is a nonsymmetric measure of the difference between two probability distributions p(x) and q(x)
- the KL divergence of q(x) from p(x), denoted $D_{KL}(p(x)||q(x))$, is a measure of the information loss when q(x) is used to approximate p(x).

Kullback-Leibler divergence

Let p(x) and q(x) be two probability distributions of a discrete random variable x. That is, both p(x) and q(x) sum up to 1, and p(x) > 0 and q(x)>0 for any x in X. $D_{KL}(p(x)||q(x))$ is defined as

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Typically p(x) represents the "true" distribution of data. The measure q(x) typically represents a theory, model, description, or approximation of p(x).

- it is not a distance measure, because it is not a metric measure.
- It is not symmetric: the KL from p(x) to q(x) is generally not the same as the KL from q(x) to p(x).
- $D_{KL}(p(x)||q(x))$ is a nonnegative measure. $D_{KL}(p(x)||q(x)) \geq 0$ and $D_{KL}(p(x)||q(x)) = 0$ if and only if p(x) = q(x)

Kullback-Leibler divergence

Example. Suppose there are two sample distributions P and Q as follows: P : (a : 3/5, b : 1/5, c : 1/5) and Q: (a : 5/9, b : 3/9, d : 1/9). Compute the KL divergence $D_{KL}(P \,||Q)$

Kullback-Leibler divergence

Example. Suppose there are two sample distributions P and Q as follows: P : (a : 3/5, b : 1/5, c : 1/5) and Q: (a : 5/9, b : 3/9, d : 1/9). Compute the KL divergence $D_{KL}(P\,||Q)$

No sample d in P, and no sample c in Q?

Avoiding the Zero-Probability Problem

Kullback-Leibler divergence: smoothing

Example. Suppose there are two sample distributions P and Q as follows: P : (a : 3/5, b : 1/5, c : 1/5) and Q: (a : 5/9, b : 3/9, d : 1/9). Compute the KL divergence $D_{KL}(P \| Q)$

- Introduce a small constant e = 0.001,
- **smoothing:** the missing symbols can be added to each distribution accordingly, with the small probability e.
  - P' : (a : 3/5 − e/3, b : 1/5 − e/3, c : 1/5 − e/3, d : e)
  - Q' : (a : 5/9 − e/3, b : 3/9 − e/3, c : e, d : 1/9 − e/3)

D_{KL}(P',Q') can be calculated.

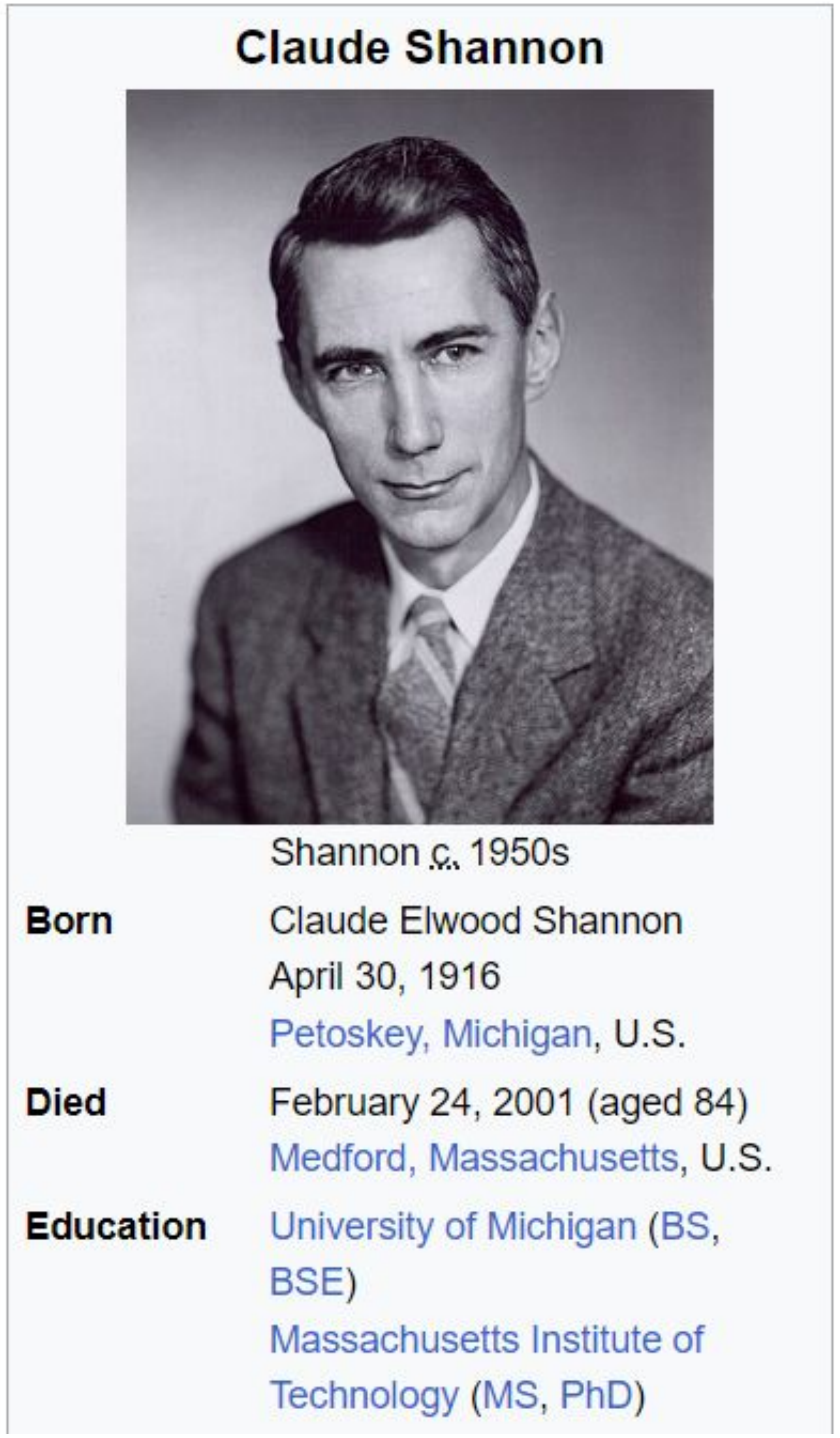
Kullback-Leibler divergence

Example. Suppose there are two sample distributions P and Q as follows: P : (a : 3/5, b : 1/5, c : 1/5) and Q: (a : 5/9, b : 3/9, d : 1/9). Compute the KL divergence $D_{KL}(P||Q)$

No sample d in P, and no sample c in Q?

Avoiding the Zero-Probability Problem

Information Theory



Claude Shannon

Shannon c. 1950s

Born	Claude Elwood Shannon April 30, 1916 Petoskey, Michigan, U.S.
Died	February 24, 2001 (aged 84) Medford, Massachusetts, U.S.
Education	University of Michigan (BS, BSE) Massachusetts Institute of Technology (MS, PhD)

Claude Elwood Shannon was an American mathematician, electrical engineer, computer scientist and cryptographer known as the "**father of information theory**" and as the "**father of the Information Age**". … and was one of the **founding fathers of artificial intelligence**.

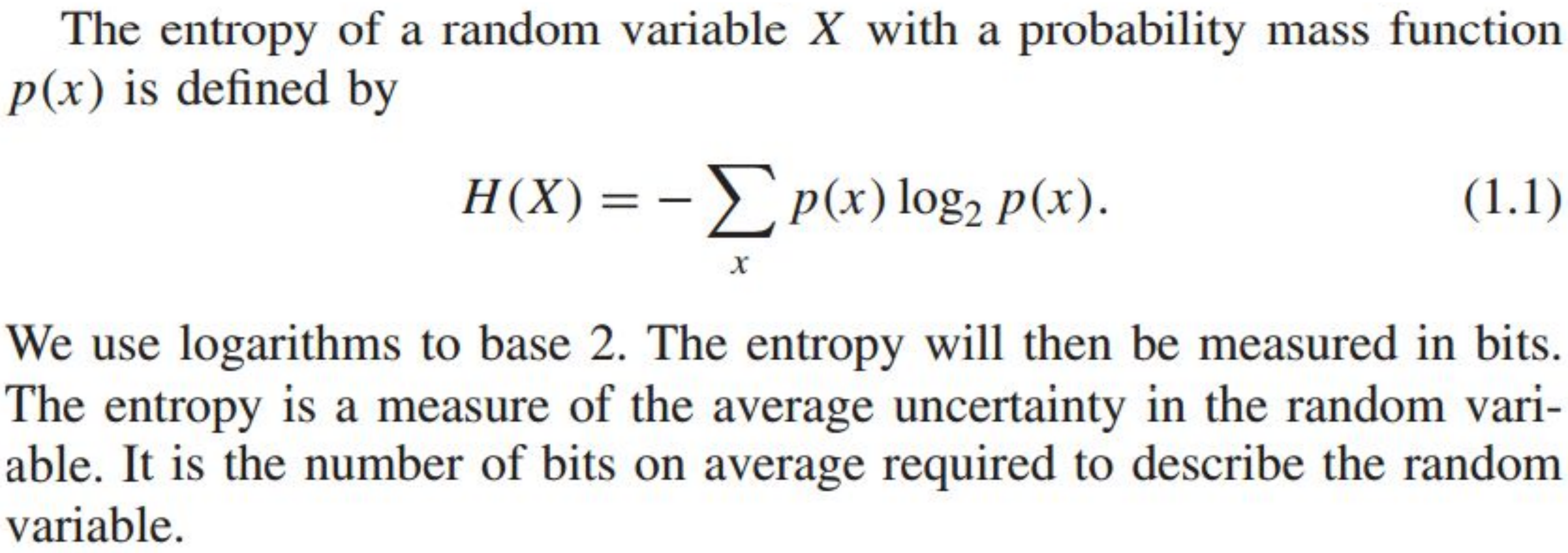
The Mathematical Theory of Communication (1948)

Information Theory & Entropy

Goal: is to reliable and efficiently transmit a message from a sender to a recipient.
In digital age, message are composed of bits. Bit = 0 or 1, when we communicate a message, we want as much useful information as possible to get through.

What is Entropy?

- Entropy measures the **uncertainty** in a probability distribution.
- It quantifies the **average amount of information** you gain from observing an outcome.



The entropy of a random variable X with a probability mass function $p(x)$ is defined by

$$H(X) = -\sum_{x} p(x) \log_2 p(x). \qquad (1.1)$$

We use logarithms to base 2. The entropy will then be measured in bits. The entropy is a measure of the average uncertainty in the random variable. It is the number of bits on average required to describe the random variable.

Entropy

Example: Consider rolling a **fair eight-sided die** where each face (1-8) is equally likely. Each outcome has a probability of 1/8. so the entropy is ___ ?

$$H(X) = -\sum_{x} p(x) \log_2 p(x).$$

Entropy

Example: Consider rolling a **fair eight-sided die** where each face (1-8) is equally likely. Each outcome has a probability of 1/8. so the entropy is ___ ?

$$H(X) = -8 \times \left(\frac{1}{8}\log_2\left(\frac{1}{8}\right)\right)$$

$$H(X) = -\sum_x p(x)\log_2 p(x).$$

$$H(X) = -\log_2\left(\frac{1}{8}\right)$$

$$H(X) = -\log_2\left(2^{-3}\right)$$

$$H(X) = 3\,\text{bits}$$

This means, on average, you get 3 bits of information per roll.

Entropy in Biased Dice

Example: Now consider a **biased eight-sided die** where:

- P(1)=P(2)=0.35
- P(3)=P(4)=0.1
- P(5)=P(6)=0.04
- P(7)=P(8)=0.01

$$H(X) = -\sum_{x} p(x)\log_2 p(x).$$

so the entropy is ___ ?

Entropy in Biased Dice

Example: Now consider a **biased eight-sided die** where:

- P(1)=P(2)=0.35
- P(3)=P(4)=0.1
- P(5)=P(6)=0.04
- P(7)=P(8)=0.01

$$H(X) = -\sum_{x} p(x) \log_2 p(x).$$

so the entropy is ___ ?

$$H(X) = -(0.35\log_2(0.35) + 0.35\log_2(0.35) + 0.1\log_2(0.1) + 0.1\log_2(0.1) + 0.04\log_2(0.04) + 0.04\log_2(0.04) + 0.01\log_2(0.01) + 0.01\log_2(0.01))$$

- $0.35\log_2(0.35) \approx -0.530$
- $0.1\log_2(0.1) \approx -0.332$
- $0.04\log_2(0.04) \approx -0.185$
- $0.01\log_2(0.01) \approx -0.066$

$$H(X) = -(2 \times -0.530 + 2 \times -0.332 + 2 \times -0.185 + 2 \times -0.066)$$

$$H(X) = 1.06 + 0.664 + 0.37 + 0.132 = 2.226 \text{ bits}$$

This means, on average, you get 2.226 bits of information per roll.

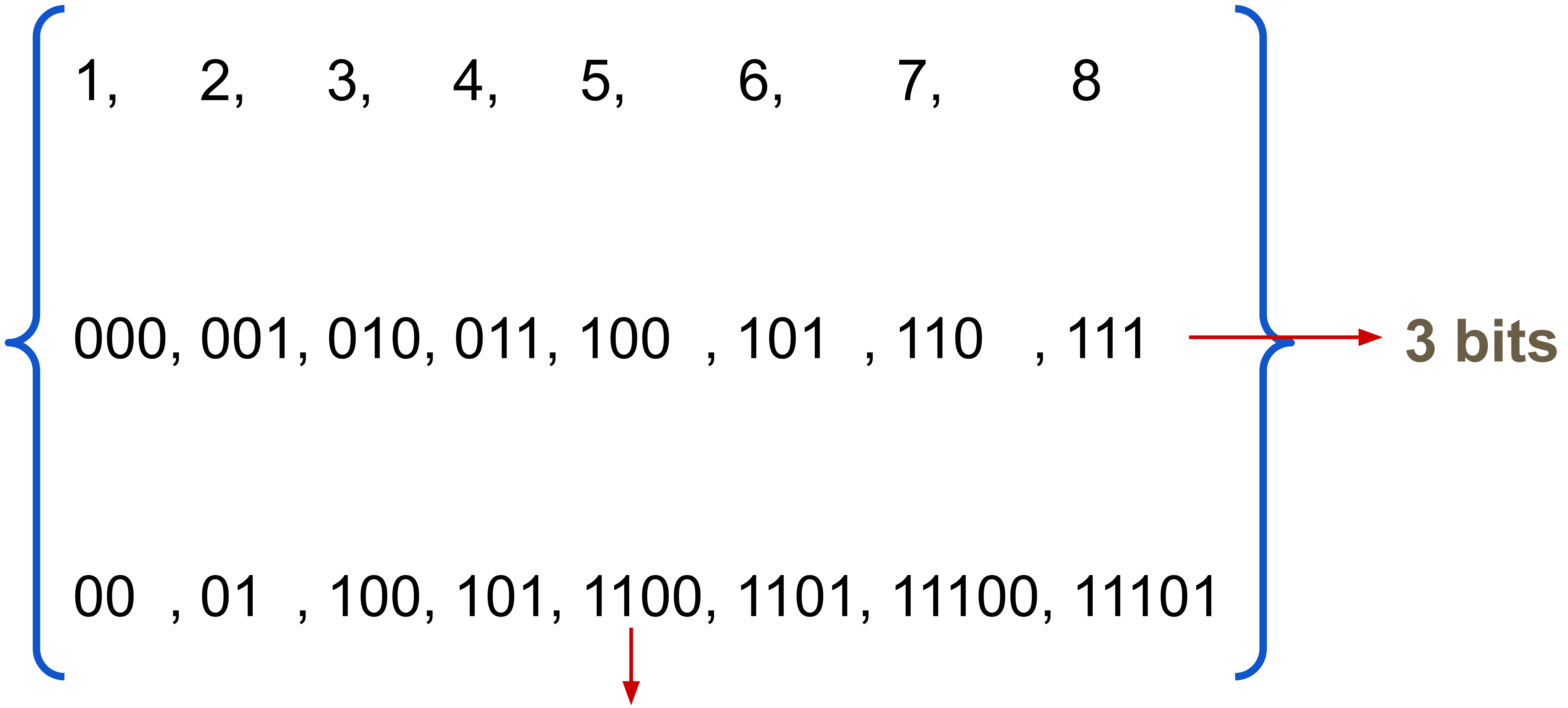
Entropy in Biased Dice

Example: Now consider a **biased eight-sided die** where:

$$H(X) = -\sum_{x} p(x)\log_2 p(x).$$

- P(1)=P(2)=0.35
- P(3)=P(4)=0.1
- P(5)=P(6)=0.04
- P(7)=P(8)=0.01

so the entropy is ____ ?



1, 2, 3, 4, 5, 6, 7, 8

000, 001, 010, 011, 100 , 101 , 110 , 111

3 bits

00 , 01 , 100, 101, 1100, 1101, 11100, 11101

0.35*2*2 + 0.1*3*2 + 0.04*4*2 + 0.01*5*2 = **2.42 bits**

This means, on average, you get **2.226** bits of information per roll.

Entropy in Biased Dice

Example: Now consider a **biased eight-sided die** where:

$$H(X) = -\sum_{x} p(x)\log_2 p(x).$$

- P(1)=P(2)=~~0.35~~ 0.01
- P(3)=P(4)=0.1
- P(5)=P(6)=0.04
- P(7)=P(8)=~~0.01~~ 0.35

H(x) = **2.226**

1, 2, 3, 4, 5, 6, 7, 8

000, 001, 010, 011, 100 , 101 , 110 , 111 → **3 bits**

00 , 01 , 100, 101, 1100, 1101, 11100, 11101

~~0.35~~0.01*2*2 + 0.1*3*2 + 0.04*4*2 + ~~0.01~~0.35*5*2 = **4.46 bits**

Cross-Entropy and Message Encoding

What is Cross-Entropy?

- **Cross-Entropy** measures the average number of bits required to transmit outcomes from a distribution p (true) when using a code based on distribution q(predicted).

1,	2,	3,	4,	5,	6,	7,	8
00 ,	01 ,	100,	101,	1100,	1101,	11100,	11101

$$H(p,q) = -\sum_i p(x_i)\log q(x_i)$$

Where:

- $p(x_i)$ is the probability of the true distribution for event x_i,
- $q(x_i)$ is the probability of the predicted distribution for event x_i,

$$q = \left\{\frac{1}{2^2} = 0.25, \frac{1}{2^2} = 0.25, \frac{1}{2^3} = 0.125, \frac{1}{2^3} = 0.125, \frac{1}{2^4} = 0.0625, \frac{1}{2^4} = 0.0625, \frac{1}{2^5} = 0.03125, \frac{1}{2^5} = 0.03125\right\}$$

$$p = \{0.35, 0.35, 0.1, 0.1, 0.04, 0.04, 0.01, 0.01\}$$

Cross-Entropy and Message Encoding

What is Cross-Entropy?

- **Cross-Entropy** measures the average number of bits required to transmit outcomes from a distribution p (true) when using a code based on distribution q(predicted).

$$H(p,q) = -\sum_i p(x_i)\log q(x_i) = \underline{H(p) + D_{KL}(p||q)}$$

- It also measures the difference between the true probability distribution p and the predicted distribution q, quantifying how well the predicted distribution approximates the true one.
- If p=q, the cross-entropy will be equal to the entropy of p. If p and q differ, the cross-entropy will be greater than the entropy of p, the difference between p and q is the **KL Divergence**.

Summary

- Similarity and distance measures
  - Proximity Measures for
    - Nominal Attributes
    - Binary Attributes
    - Numeric Attributes
    - Ordinal attributes
    - Mixed types
  - Cosine Similarity
  - Kullback-Leibler divergence
  - Entropy & Cross Entropy