# Data

IT 326: Data Mining

First semester 2021

Chapter 2, "Data Mining: Concepts and Techniques" (3rd ed.)

# Outline



- Data Objects and Attribute Types
- Measuring Data Similarity and Dissimilarity
- Summary

# Types of Data Sets



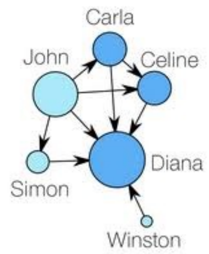
- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data
  - Video data
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data تسلسل جيني
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures تراكيب جزيئية

| Tid | Items bought |
| --- | --- |
| 10 | Tea, Nuts, Water |
| 20 | Tea, Coffee, Water |
| 30 | Tea, Water, Eggs |
| 40 | Nuts, Eggs, Milk |
| 50 | Nuts, Coffee, Water, Eggs, Milk |

كل ترانزكشن فيه Set of Item → Transaction data

| | team | coach | pla y | ball | score | game | wi n | 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 |

Document data




Graph

# Data Objects



- Data sets are made up of data objects.
- A data object represents an entity.
  - Examples:
    - sales database: customers, store items, sales
    - medical database: patients, treatments
    - university database: students, professors, courses
  - Also called samples , examples, instances, data points, objects, tuples.
- Data objects are described by attributes.
  - Attributes also called dimensions, features, variables.
- Database: rows → data objects; columns → attributes.

Attributes

Objects

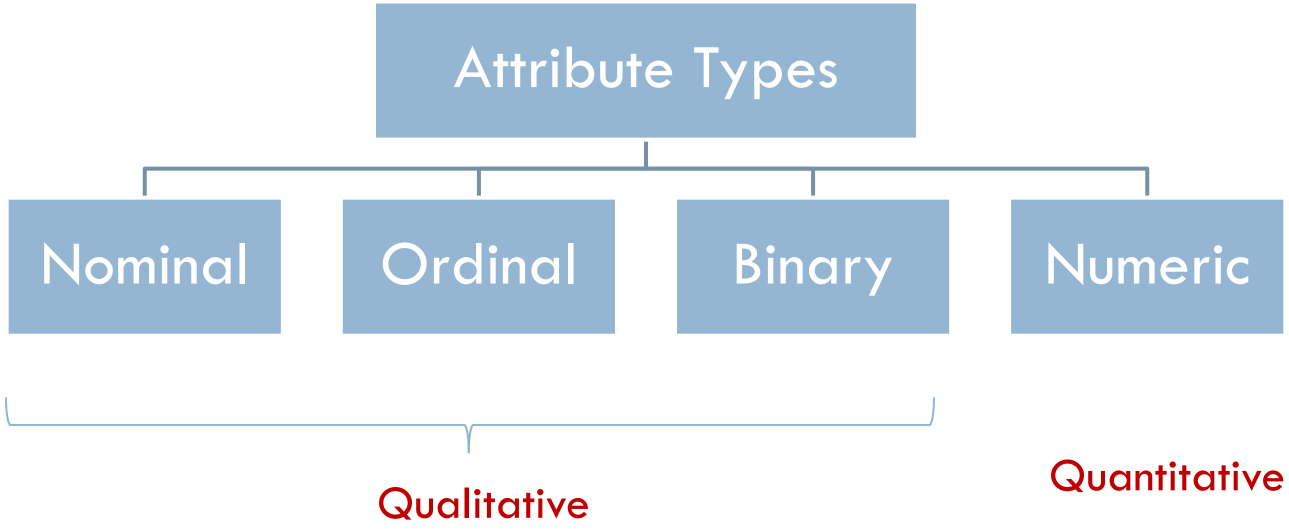
| Student | ID | Name | GPA | Age |
|---|---|---|---|---|
| Student 1 | 8000 | Sam | 3.45 | 19 |
| Student 2 | 5001 | Jill | 2.65 | 21 |

**Relational records**

# Attributes



- Attribute: a data field, representing a characteristic or feature of a data object.
  - E.g., customer_ID, name, address
- The type of an attribute is determined by the set of possible values the attribute can have.



# Attribute Types (Qualitative)



- Nominal: categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes

- Ordinal:
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

# Attribute Types (Qualitative)



- Binary:
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

# Attribute Types (Quantitative)



- Numeric: a measurable quantity, represented in integer or real values.
  - Numeric attributes can be further categorized into ( interval or ratio)

| Interval | Ratio |
| --- | --- |
| ○ Measured on a scale of equal-sized units<br>○ Values have order<br>○ No true zero-point<br>○ e.g., temperature in C°or F°, calendar dates | ○ Inherent zero-point<br>○ we can speak of a value as being a multiple (or ratio) of another value<br>○ e.g., counts( years of experiences, number of words), monetary quantities |

# Proximity

# Similarity and Dissimilarity



- Proximity refers to a similarity or dissimilarity
- 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

# Data Matrix and Dissimilarity Matrix



- **Data matrix:** stores the n data objects in the form of a relational table.

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

$n$-by-$p$ matrix ($n$ objects $\times p$ attributes)

- **Dissimilarity matrix**: This structure stores a collection of proximities that are available for all pairs of $n$ objects.
  - A triangular matrix

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

# Proximity Measure for Nominal Attributes



- Can take 2 or more states, e.g., red, yellow, blue, green (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

**Example:**

| Obj | Color |
|---|---|
| 1 | Red |
| 2 | Yel |
| 3 | Red |
| 4 | Green |

→

| Obj | col-red | col-yel | col_blue | col_green |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 |

# Example: Nominal attributes



A Sample Data Table Containing Attributes of Mixed Type

| 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 |

normal data

S=1

S=0

d=0 = إذا متشابهين

d=1 ≠ مختلفين //

**Dissimilarity matrix**

"test-1"

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

$$\begin{array}{c} \\ \neq \\ \neq \\ \\ \end{array} \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

# Proximity Measure for Binary Attributes

s asy



- A contingency table for binary attributes: ⇒

| | | Object $j$ | | |
|---|---|---|---|---|
| | | 1 | 0 | sum |
| | 1 | $q$ | $r$ | $q+r$ |
| Object $i$ | 0 | $s$ | $t$ | $s+t$ |
| | sum | $q+s$ | $r+t$ | $p$ |

- Distance measure for symmetric binary variables:

يحسب الاختلاف

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

عدد ١ و ٠ ترتيبين →

* r 10
s 01
t

- Distance measure for asymmetric binary variables:

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

(1) نفس المعادلة بس شلنا t →

* ⇒ يحسب التشابه

- Jaccard coefficient: similarity measure for asymmetric binary variables.

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

# Example: Binary Attributes



| | | Jim | |
|---|---|---|---|
| | | 1 | 0 |
| Jack | 1 | q | r |
| | 0 | s | t |

$$\frac{r+s}{q+r+s}$$

Relational Table Where Patients Are Described by Binary Attributes

| *name* | *gender* | *fever* | *cough* | *test-1* | *test-2* | *test-3* | *test-4* |
|---|---|---|---|---|---|---|---|
| Jack | M | Y 1 | N 0 | P 1 | N 0 | N 0 | N 0 |
| Jim | M | Y 1 | Y 1 | N 0 | N 0 | N 0 | N 0 |
| Mary | F | Y 1 | N 0 | P 1 | N 0 | P 1 | N 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

have 2 possible value

Y N P N
asy

M yes = 1
F No = 0
P = 1
Sy N = 0

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

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

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

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

1 Coding 2 distance (contengcy table) 3 تعويض

gender ⇒ Symmetric
طردية

# Distance on Numeric Data: Minkowski Distance

ان 2 object لأنه ارقام فيمكن calculation



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

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

Where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are be two objects described by $p$ numeric attributes, and $h$ is the order (the distance so defined is also called $L - h$ norm)

- Properties:
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positive definiteness) 1
  - d(i, j) = d(j, i) (Symmetry) * 2
  - d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric

# Distance on Numeric Data: Special cases



- h = 1: Manhattan (city block, L1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

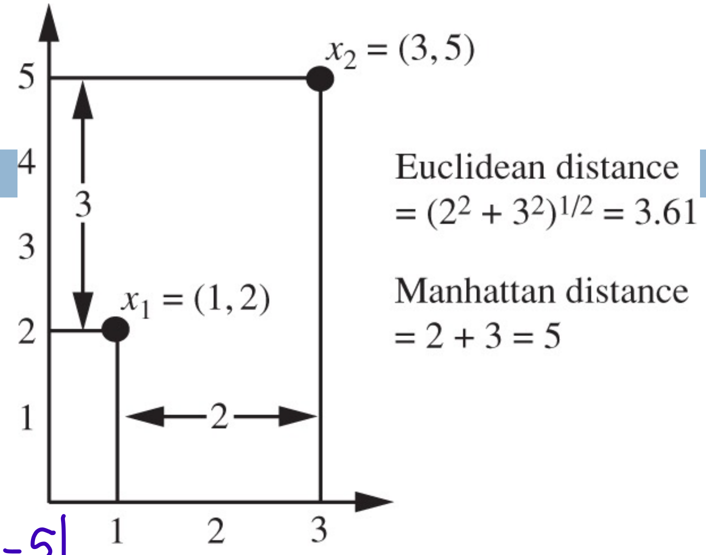
$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

- h = 2: (L2 norm) Euclidean distance

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

# Example: Numeric Distance

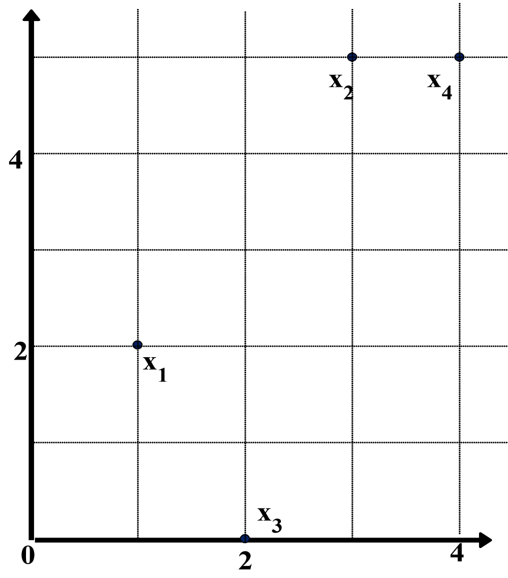





## Data Matrix

object →

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




## Dissimilarity Matrices:

① Similirty betwen distance ①

$|x_2 - x_1|$

**Manhattan ($L_1$)** $= |1-3| + |2-5| = 2+3=5$

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

② **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 |

$E = \sqrt{(1-3)^2 + (2-5)^2}$

$\Rightarrow = \sqrt{4+9} = 3.61$

# Proximity Measure for Ordinal Attributes ⇒

الاختلاف بنعين دين الترتيب

is order is important



- Order is important, e.g., rank
- Can be treated like interval-scaled:
  1. Replace $x_{if}$ by their rank $r_{if} \in \{1, \ldots, M_f\}$ 1
  2. Map the range of each variable onto [0, 1] by replacing $i^{th}$ object in the $f^{th}$ variable by: 2

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

$M_f = 3$

$M_f$: number of possible values for variable $f$.

  3. Compute the dissimilarity using methods for numeric variables.

| x | r | |
|---|---|---|
| S | 1 | $\frac{1-1}{3-1} = 0$ |
| m | 2 | $\frac{2-1}{3-1} = 0.5$ |
| L | 3 | $\frac{3-1}{3-1} = 1$ |

$m = 3$

# Example: Ordinal Attributes



A Sample Data Table Containing Attributes of Mixed Type

| 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 |

**Dissimilarity matrix**
"test-2"

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

| Obj ID | Test-2 | rank | normalize |
|---|---|---|---|
| 1 | Excellent | 3 | (3-1)/(3-1) = 1 |
| 2 | Fair | 1 | 0 |
| 3 | Good | 2 | 0.5 |
| 4 | Excellent | 3 | 1 |

$M_f = 3$ 2
Fair 1 0
good 2 0.5
Excellent 3 1

d(1,2)
m =
|1-0| = 1
d(2,3)
|0-0.5| = 0.5
d(3,1) = |1-0| = 1

# Dissimilarity for Attributes of Mixed Types



- A database may contain all attribute types:
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects:

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

  - $f$ is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$.
  - $f$ is numeric: use the normalized distance $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$
  - $f$ is ordinal:
    - Compute ranks $r_{if}$ and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - Treat $z_{if}$ as numeric

$\boldsymbol{\delta_{ij}}$ **= 0** IF either

(1) $x_{if}$ or $x_{jf}$ is missing OR

(2) $x_{if} = x_{jf} = 0$ and attribute $f$ is asymmetric binary;

Otherwise, $\boldsymbol{\delta_{ij}}$ **= 1**.

# Example: Mixed Attributes



- Calculate d(3,1).

| Obj ID | Test-1 (nominal) | Test-2 (ordinal) | Test-3 (numeric) |
|---|---|---|---|
| 1 | Code A | Excellent → 1 | 45 |
| 2 | Code B | Fair | 22 ← min |
| 3 | Code C | Good → 0.5 | 64 ← max |
| 4 | Code A | Excellent | 28 |

$$d_{3,1}^{Test-1} = 1 \qquad d_{3,1}^{Test-2} = 0.5 \qquad d_{3,1}^{Test-3} = \frac{|64-45|}{64-22} = 0.45$$

$$d(3,1) = \frac{1(1) + 1(0.5) + 1(0.45)}{1+1+1} = 0.65$$

# Mixed Attributes



**Nominal matrix (only for test#1)**

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

**Ordinal matrix (only for test#2)**

$$\begin{vmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{vmatrix}$$

**Numeric matrix (only for test#3)**

$$\begin{vmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{vmatrix}$$

**Mixed variables matrix**

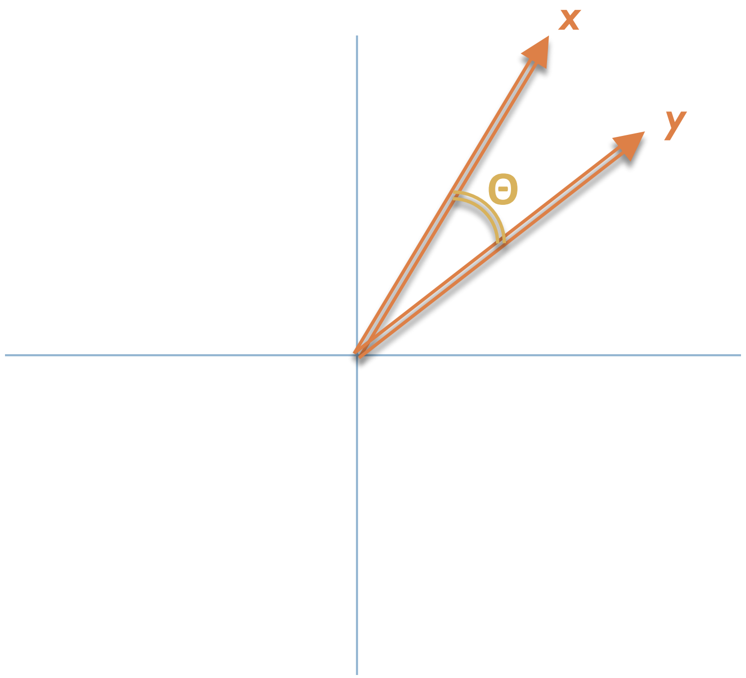
$$\begin{vmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{vmatrix}$$

$$d(2,1) = \frac{(1*1)+(1*1)+(23/42*1)}{1+1+1} = \frac{1+1+0.55}{3} = 0.85$$

$$d(4,3) = \frac{(1*1)+(0.5*1)+(36/42*1)}{1+1+1} = \frac{1+0.5+0.86}{3} = 0.79$$

# Cosine Similarity

- Cosine similarity measures the similarity between two vectors by the cosine of the angle between them.




| Θ | 0 | 22.5 | 45 | 67.5 | 90 |
|---|---|---|---|---|---|
| cos(Θ)≈ | 1 | 0.92 | 0.71 | 0.38 | 0 |

# Cosine Similarity



- Cosine measure: If x and y are two vectors (e.g., term-frequency vectors), then

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

where • indicates vector dot product, and $||x||$ is the length of vector $x$.

$$x \bullet y = \sum_i x_i * y_i \qquad \|x\| = \sqrt{\sum_i x_i^2}$$

# Example: Cosine Similarity



- 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.

Document Vector or Term-Frequency Vector

| 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 |

- Term-frequency vectors are very long and sparse (contains many zeros).
- Traditional distance measures are not suitable;
  - Many zero-matches between two documents does not mean that they are similar.
- Must ignore zero-matches → cosine similarity.

# Example: Cosine Similarity



- Ex: Find the similarity between documents 1 and 2.
- d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
- d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$$d1 \bullet d2 = (5\times 3) + (0\times 0) + (3\times 2) + (0\times 0) + (2\times 1) + (0\times 1) + (0\times 0) + (2\times 1) + (0\times 0) + (0\times 1)$$
$$= 25$$

$$\|d1\| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$\|d2\| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

$$\cos(d1, d2) = \frac{25}{6.48 \times 4.12} = 0.94$$

# Exercise



1


2

3


4

# Summary



- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Measure data similarity and dissimilarity