

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Similarity and Distance Measures

Lecture 5 · Thursday 24 September 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–5 min) Recap; why proximity is the core question
2. (5–15 min) Data matrix, dissimilarity matrix, what makes a distance
3. (15–35 min) Numeric: Manhattan, Euclidean, Chebyshev — and why units matter
4. (35–48 min) Binary (SMC, Jaccard) and nominal attributes
5. (48–62 min) Ordinal, mixed types (Gower) and cosine similarity
6. (62–68 min) Distance at scale: n^2 pairs, curse of dimensionality
7. (68–75 min) Exercise and summary

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Build	a dissimilarity matrix from a data matrix
Compute	Manhattan, Euclidean and Chebyshev distances by hand
Compute	simple matching and Jaccard for binary data, and nominal mismatch
Combine	mixed attribute types into one distance (Gower)
Choose	the right measure for the data, and explain why scale makes it hard

Similarity and Distance Measures

→ **Why 'How Close?' Is the Core Question**

Data Matrix and Dissimilarity Matrix

Numeric Attributes: The Minkowski Family

Binary and Nominal Attributes

Ordinal, Mixed Types and Cosine

Distance at Scale

In-Class Exercise

Recap: Attributes Together → Records Together

IIT Madras Zanzibar

Correlation (yesterday)

compares two **columns**

price ↔ revenue across 500 farms: r = 0.92

Distance (today)

compares two **rows**

farm 1 vs farm 2 across their attributes

Same table — read across instead of down

Recall: d(A, B) = 20 and d(A, C) = 700 for the shop customers — visits hardly counted. Today we see why and fix it.

Where Distance Is Used

IIT Madras Zanzibar

1 Clustering (Wk 8–9)

put close farms in the same group

2 k-NN classification

label a new loan like its 5 nearest past loans

3 Recommenders

suggest what similar customers bought

4 Outlier detection

a payment far from all others

5 Search

find documents or images like this one

6 De-duplication

two records so close they are the same farm

Common thread: every one of these breaks if 'close' is measured badly.

Similarity and Distance Measures

IIT Madras Zanzibar

Why 'How Close?' Is the Core Question

→ **Data Matrix and Dissimilarity Matrix**

Numeric Attributes: The Minkowski Family

Binary and Nominal Attributes

Ordinal, Mixed Types and Cosine

Distance at Scale

In-Class Exercise

Data Matrix: n Objects × p Attributes

IIT Madras Zanzibar

farm	size (ha)	rainfall (mm)	yield (t)
F1	0.25	1780.8	6.11
F2	4.18	1765.6	2.71
F3	0.68	1214.4	3.05
F4	1.40	1249.9	1.68
F5	3.60	1398.6	1.51
F6	0.20	1355.9	2.22

1. **n = 6 objects**
   rows: one farm each

2. **p = 3 attributes**
   columns: size, rainfall, yield

3. **$x_i f$**
   value of attribute f for object i: x_1,size = 0.25

Two-mode: rows and columns are different kinds of thing (farms vs attributes).

Dissimilarity Matrix: n × n Distances

IIT Madras Zanzibar

Dissimilarity matrix (Euclidean, z-scores)

	F1	F2	F3	F4	F5	F6
F1	0.00	3.31	2.90	3.39	3.80	2.79
F2	3.31	0.00	3.32	2.92	1.70	3.24
F3	2.90	3.32	0.00	0.92	2.32	0.81
F4	3.39	2.92	0.92	0.00	1.64	0.98
F5	3.80	1.70	2.32	1.64	0.00	2.39
F6	2.79	3.24	0.81	0.98	2.39	0.00

1. **Diagonal = 0**
   a farm vs itself
2. **Symmetric**
   d(i, j) = d(j, i): store one triangle
3. **Most alike**
   F3–F6: 0.81
4. **Least alike**
   F1–F5: 3.80
5. **Size**
   n(n − 1)/2 pairs: 6 farms → 15

One-mode: rows and columns are both farms. Clustering algorithms take this matrix as input.

What Makes a Proper Distance?

IIT Madras Zanzibar

1 **Non-negative**

d(i, j) ≥ 0

distances are never negative

2 **Identity**

d(i, j) = 0 only if i = j

zero means identical

3 **Symmetry**

d(i, j) = d(j, i)

Zanzibar → Dar = Dar → Zanzibar

4 **Triangle inequality**

d(i, j) ≤ d(i, k) + d(k, j)

direct is never longer than a detour

Why it matters: the triangle inequality lets fast algorithms skip pairs that cannot possibly be close.

Similarity and Distance Measures

IIT Madras Zanzibar

Why 'How Close?' Is the Core Question

Data Matrix and Dissimilarity Matrix

→ **Numeric Attributes: The Minkowski Family**

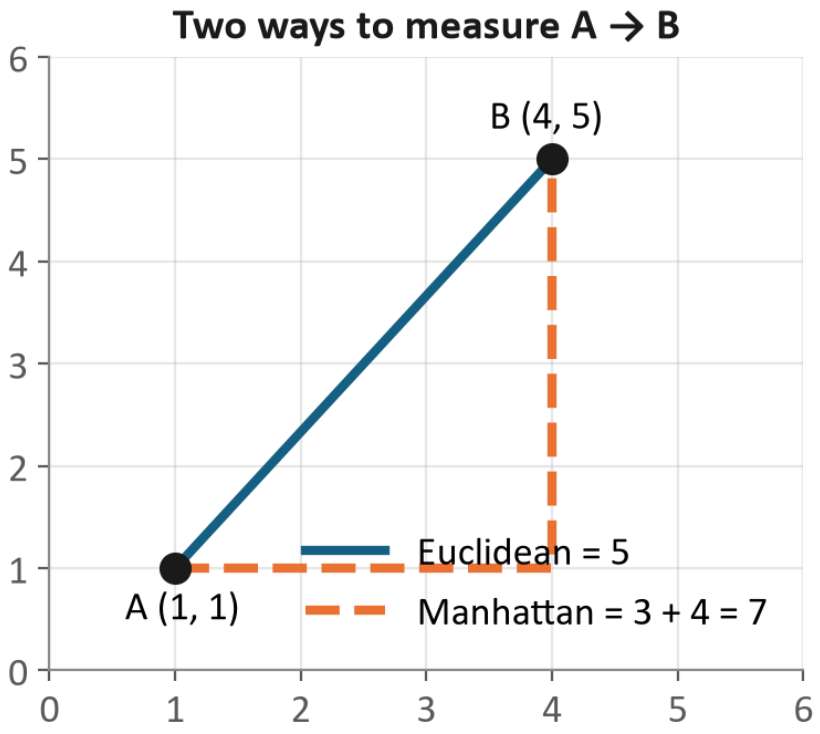
Binary and Nominal Attributes

Ordinal, Mixed Types and Cosine

Distance at Scale

In-Class Exercise

Two Ways from A to B



1. **Differences**
   $|4 - 1| = 3$ and $|5 - 1| = 4$

2. **Euclidean (straight)**
   $\sqrt{(3^2 + 4^2)} = \sqrt{25} = 5$

3. **Manhattan (grid)**
   $3 + 4 = 7$

4. **Chebyshev (largest)**
   $\max(3, 4) = 4$

5. **Always**
   Manhattan ≥ Euclidean ≥ Chebyshev

Analogy: Euclidean = the crow's flight; Manhattan = walking Stone Town's alleys.

One Formula, One Dial: Minkowski Distance

IIT Madras Zanzibar

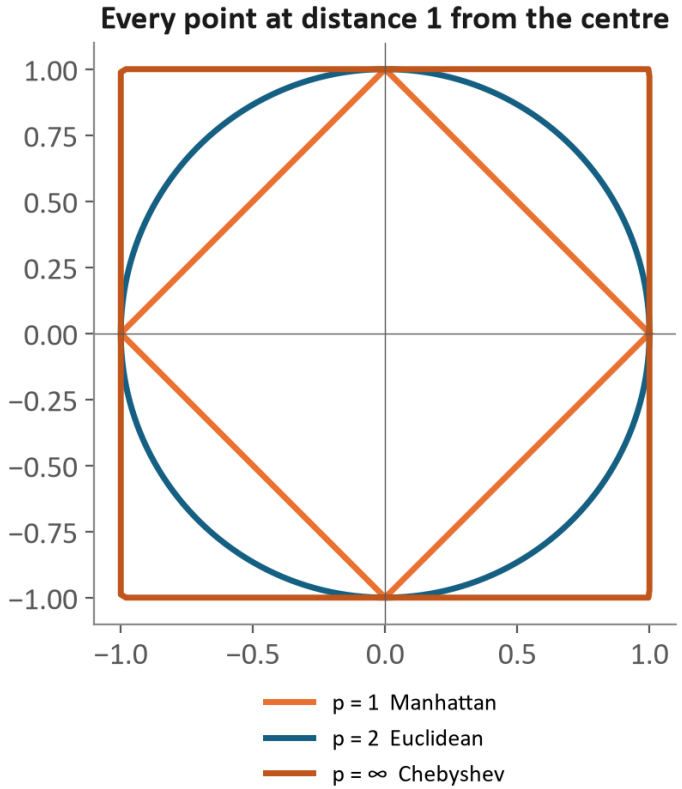
$$d(i, j) = (\Sigma f \, |x_i f - x_j f|^h)^{\wedge}(1/h)$$

h = 1 Manhattan (L_1) · h = 2 Euclidean (L_2) · h → ∞ Chebyshev ($L\infty$)

1. **h = 1**
   add the absolute differences

2. **h = 2**
   square, add, square-root

3. **h → ∞**
   only the largest difference counts



Larger h: more weight on the single biggest difference.

Units Decide Who Wins: Farms F1 and F2

IIT Madras Zanzibar

	size	rainfall	yield	Euclidean
F1 raw	0.25 ha	1780.8 mm	6.11 t	
F2 raw	4.18 ha	1765.6 mm	2.71 t	
raw difference²	$3.93^2 = 15.4$	$15.2^2 = \mathbf{231.0}$	$3.40^2 = 11.6$	$\sqrt{258.0} = 16.06$
z-score difference²	$2.71^2 = \mathbf{7.34}$	$0.06^2 = 0.00$	$1.90^2 = 3.61$	$\sqrt{10.95} = 3.31$

1 **Raw**
rainfall (a tiny 15 mm gap) dominates because mm are big numbers

2 **Standardized**
z = (x − mean) ÷ std: size, the real difference, dominates

Rule: standardize numeric attributes before computing distance (Week 3: normalization).

Similarity and Distance Measures

IIT Madras Zanzibar

Why 'How Close?' Is the Core Question

Data Matrix and Dissimilarity Matrix

Numeric Attributes: The Minkowski Family

→ **Binary and Nominal Attributes**

Ordinal, Mixed Types and Cosine

Distance at Scale

In-Class Exercise

Binary Attributes: Count Matches

IIT Madras Zanzibar

item	seed	urea	hoe	pesticide	gloves	sacks
A	1	1	0	1	0	0
B	1	0	0	1	1	0
case	q	r	t	q	s	t

	B = 1	B = 0
A = 1	q = 2	r = 1
A = 0	s = 1	t = 2

1 **q**
both 1 — shared purchase

2 **r, s**
only one of them — mismatch

3 **t**
both 0 — neither bought it

Total: p = q + r + s + t = 2 + 1 + 1 + 2 = 6 attributes.

Symmetric (SMC) or Asymmetric (Jaccard)?

IIT Madras Zanzibar

$$\text{SMC} = (q + t) \div (q + r + s + t)$$

symmetric binary: 0 and 1 equally important

$$\text{Jaccard} = q \div (q + r + s)$$

asymmetric binary: ignore the 0–0 matches

1 **A vs B**
SMC = (2 + 2) ÷ 6 = 0.67 → distance 0.33

2 **A vs B**
Jaccard = 2 ÷ 4 = 0.50 → distance 0.50

Why ignore 0–0? two shoppers who both skipped 9,990 of 10,000 products would look 99.9% alike by SMC. Shared purchases (and shared symptoms) are the signal; shared absences are not.

Nominal Attributes: Count Mismatches

IIT Madras Zanzibar

$$d(i, j) = (p - m) \div p$$

p = number of nominal attributes; m = number that match

farm	region	crop	fertilizer
F3	South Unguja	Coconut	Yes
F6	North Unguja	Coconut	Yes
match?	no	yes	yes

1 **m = 2, p = 3**
d = (3 − 2) ÷ 3 = 0.33

2 **Alternative**
one-hot: a 0/1 column per category, then Jaccard

Similarity and Distance Measures

IIT Madras Zanzibar

Why 'How Close?' Is the Core Question

Data Matrix and Dissimilarity Matrix

Numeric Attributes: The Minkowski Family

Binary and Nominal Attributes

→ **Ordinal, Mixed Types and Cosine**

Distance at Scale

In-Class Exercise

Ordinal Attributes: Ranks onto [0, 1]

IIT Madras Zanzibar

$$z = (r - 1) \div (M - 1)$$

r = rank of the value; M = number of ordered levels

Poor	Fair	Good	Excellent
rank 1 → z = 0.00	rank 2 → z = 0.33	rank 3 → z = 0.67	rank 4 → z = 1.00

1 **Fair vs Excellent**
|0.33 − 1.00| = 0.67

Same range: every ordinal attribute now runs 0 to 1, so none dominates.

Mixed Types: Gower's Distance

$$d(i, j) = \Sigma f\ \delta_{ij}^{(f)}\ d_{ij}^{(f)} \div \Sigma f\ \delta_{ij}^{(f)}$$

$d^{(f)}$ in [0, 1] by type; $\delta = 0$ skips an attribute (missing value, or 0–0 on an asymmetric binary)

attribute	type	F3	F6	rule	$d^{(f)}$
region	nominal	South Unguja	North Unguja	mismatch → 1	1
crop	nominal	Coconut	Coconut	match → 0	0
fertilizer	binary (sym.)	Yes	Yes	match → 0	0
farm size	numeric	0.68	0.20	|0.48| ÷ 8.94	0.054

1 **Average**
(1 + 0 + 0 + 0.054) ÷ 4 = 0.26

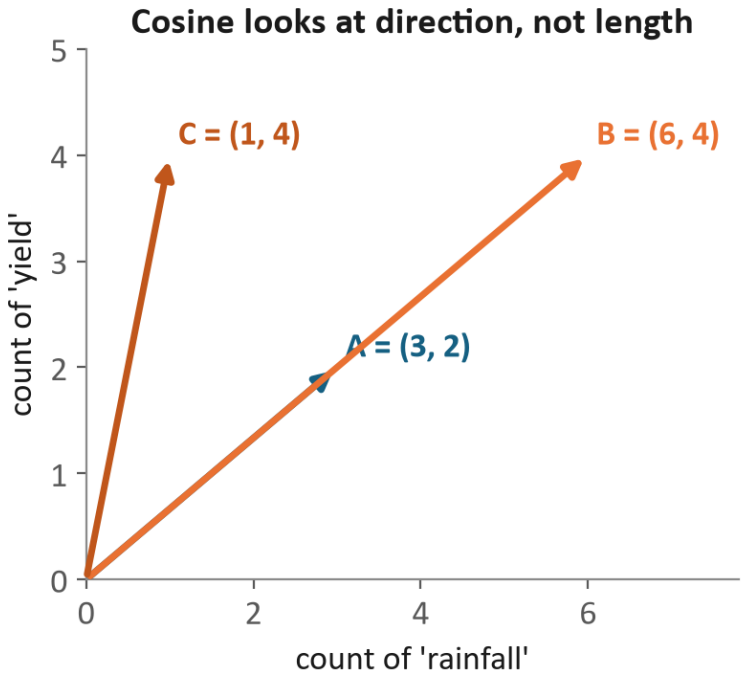
Cosine Similarity: Direction, Not Length

$$\cos(x, y) = (x \cdot y) \div (\|x\| \, \|y\|)$$

$x \cdot y = \Sigma\, x_k y_k$ (dot product); $\|x\| = \sqrt{\Sigma\, x_k^2}$ (length)

1 **A = (3, 2), B = (6, 4)**
26 ÷ (√13 × √52) = 26 ÷ 26 = 1.00 — same direction

2 **A = (3, 2), C = (1, 4)**
11 ÷ (√13 × √17) = 11 ÷ 14.87 = 0.74



Use for: word counts in documents and SMS, user–item ratings. A report twice as long on the same topic still scores 1.0.

Choosing the Right Measure

IIT Madras Zanzibar

Data	Measure	Example
Numeric	standardize, then Euclidean / Manhattan	farm size, rainfall, yield
Symmetric binary	simple matching (SMC)	fertilizer used yes/no
Asymmetric binary / sets	Jaccard	purchases, symptoms
Nominal	mismatch fraction $(p - m) \div p$	region, crop
Ordinal	rank → $z = (r - 1)/(M - 1)$, then numeric	crop health
Mixed	Gower: average of per-attribute distances	the whole farm table
Text / sparse counts	cosine similarity	SMS, reports, ratings

Similarity and Distance Measures

IIT Madras Zanzibar

Why 'How Close?' Is the Core Question

Data Matrix and Dissimilarity Matrix

Numeric Attributes: The Minkowski Family

Binary and Nominal Attributes

Ordinal, Mixed Types and Cosine

→ **Distance at Scale**

In-Class Exercise

Scale Problem 1: Too Many Pairs

IIT Madras Zanzibar

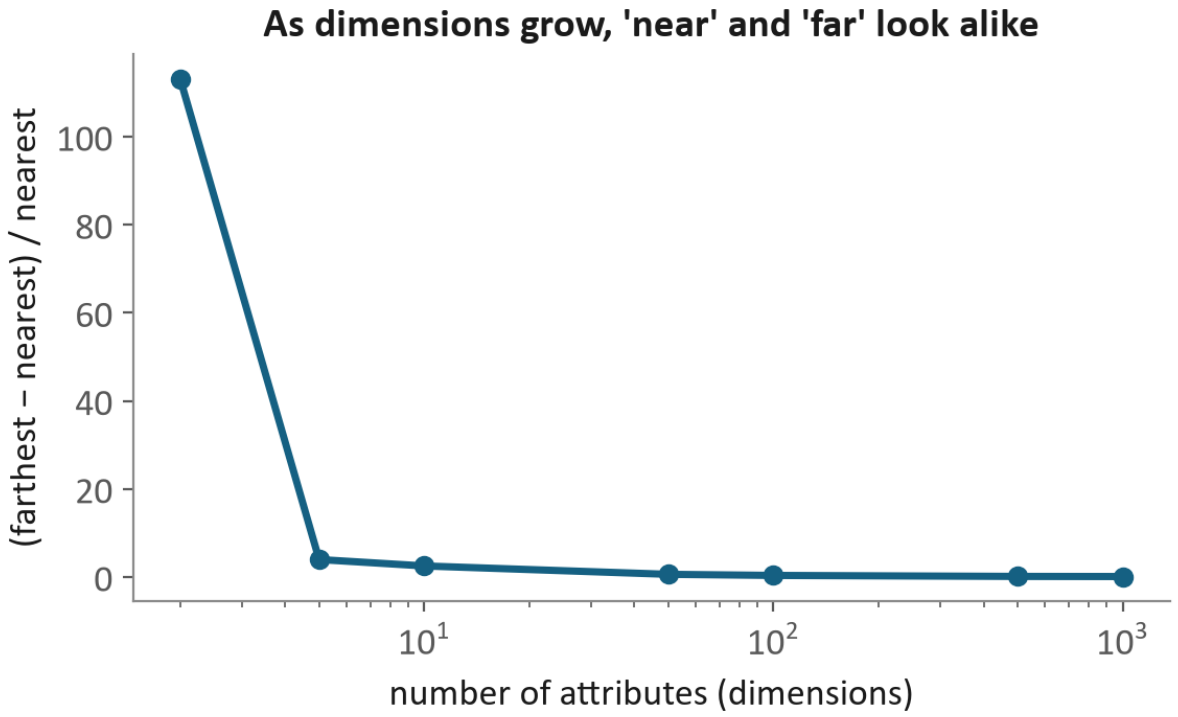
objects n	pairs n(n − 1)/2	at 10^6 pairs per second
500 farms	124,750	0.1 seconds
10,000 farms	≈ 50 million	50 seconds
1 million customers	≈ 500 billion	≈ 6 days

1. **Never all pairs**
   at scale, compare only likely candidates

2. **LSH**
   hash similar items into the same bucket (MMDS Ch. 3)

Analogy: to find a friend in a stadium, check your own section first, not all 60,000 seats.

Scale Problem 2: The Curse of Dimensionality



dimensions	(far − near) ÷ near
2	113.19
10	2.52
100	0.37
1000	0.12

Fix: reduce dimensions first (PCA, Week 3) or select the attributes that matter — otherwise 'nearest' means little.

Similarity and Distance Measures

IIT Madras Zanzibar

Why 'How Close?' Is the Core Question

Data Matrix and Dissimilarity Matrix

Numeric Attributes: The Minkowski Family

Binary and Nominal Attributes

Ordinal, Mixed Types and Cosine

Distance at Scale

→ **In-Class Exercise**

In-Class Exercise (6 min, pairs)

IIT Madras Zanzibar

IN-CLASS TASK

Two standardized farms: P = (0.5, −1.0, 2.0), Q = (1.5, 1.0, 0.0). Two patients' symptoms (fever, cough, rash, headache, vomiting): U = (1, 1, 0, 1, 0), V = (1, 0, 0, 1, 1).

1. Compute the Manhattan, Euclidean and Chebyshev distances between P and Q.
2. For U and V, find q, r, s, t; compute SMC and Jaccard.
3. Which of SMC and Jaccard suits symptoms, and why?
4. Two SMS word-count vectors: (2, 0, 1) and (4, 0, 2). Cosine similarity?

Solution

IIT Madras Zanzibar

1. **P vs Q**
   differences 1, 2, 2 → Manhattan 5 · Euclidean √(1 + 4 + 4) = 3 · Chebyshev 2

2. **U vs V**
   q = 2 (fever, headache), r = 1 (cough), s = 1 (vomiting), t = 1 (rash) → SMC = 3/5 = 0.6 · Jaccard = 2/4 = 0.5

3. **Which?**
   Jaccard: symptoms are asymmetric — sharing 'no rash' says little

4. **Cosine**
   (8 + 0 + 2) ÷ (√5 × √20) = 10 ÷ 10 = 1.0 — same direction, one is twice as long

Summary & Exam Checkpoints

IIT Madras Zanzibar

Two matrices

data matrix n × p; dissimilarity matrix n × n, symmetric, zero diagonal

Numeric

Minkowski: h = 1 Manhattan, h = 2 Euclidean, h → ∞ Chebyshev; standardize first

Categorical

SMC (symmetric), Jaccard (asymmetric), mismatch (nominal), ranks (ordinal), Gower (mixed)

At scale

n^2 pairs → LSH; many dimensions → reduce first; cosine for text

What's Next

IIT Madras Zanzibar

Tut **Tutorial 2 (today 4:45 pm)**
correlation, distance and similarity on the 500 farms in pandas

Fri **Lecture 6 (25 Sep): Data Quality & Cleaning**
missing values, noise, outliers, integration, χ^2

Wk 3 **Normalization, PCA, sampling**
the fixes for today's units and dimensionality problems

Practise: scipy.spatial.distance.cdist and sklearn.metrics.pairwise on the farm file.

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed., Section 2.3 (2.2.3–2.2.4 in some printings). Morgan Kaufmann.
- Leskovec, J., Rajaraman, A. & Ullman, J. (2020). Mining of Massive Datasets, 3rd ed., Ch. 3 (similarity, LSH). mmds.org.
- Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics 27(4).
- Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. New Phytologist 11(2).
- Beyer, K. et al. (1999). When Is 'Nearest Neighbor' Meaningful? ICDT.