

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Clustering I: k-Means

Monday 9 November 2026 · 50 minutes

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

By the End of Today You Can

IIT Madras Zanzibar

1. Tell supervised from unsupervised learning
2. Run k-means by hand: assign, move, repeat
3. Choose k with the elbow and the silhouette
4. Profile and name clusters so someone can act
5. Explain k-means' weaknesses and their fixes

Today's Plan (45 minutes + 5 minutes questions)

IIT Madras Zanzibar

minutes	part	what you get
0 – 8	Learning without labels	clusters, distance, scaling
8 – 22	k-means by hand	8 farms, round by round; code from scratch
22 – 30	Choosing k	elbow and silhouette
30 – 38	10,000 real farms	profile, name, honest check
38 – 45	Weaknesses, exercise	
45 – 50	Questions	

Learning Without Labels

IIT Madras Zanzibar

Supervised (Weeks 6 – 7)

every training farm has an answer
learn to predict it
check against the answer

classification

Unsupervised (today)

no answers at all
find natural groups
judge by tight and separate

clustering

Definition: cluster = a group of records more similar to each other than to records in other groups

Like sorting unlabelled laundry: you naturally make piles of whites, darks and towels.

Similar Means Close: Distance

$$d(A, B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots}$$

step	what	example
1	subtract each attribute	A = (1, 1), B = (4, 5): 3 and 4
2	square and add	9 + 16 = 25
3	**square root**	**√25 = 5**

Small distance = similar farms. k-means puts close farms together.

Distance Needs Scaling First

IIT Madras Zanzibar

attribute	what it tells us	preparation
farm size	scale of the operation	log, then z-score (sizes are skewed)
average yield	productivity vs same crop	z-score within crop, averaged per farm
fertilizer used	input practice	Yes = 1, No = 0, then z-score
irrigated	water management	Yes = 1, No = 0, then z-score

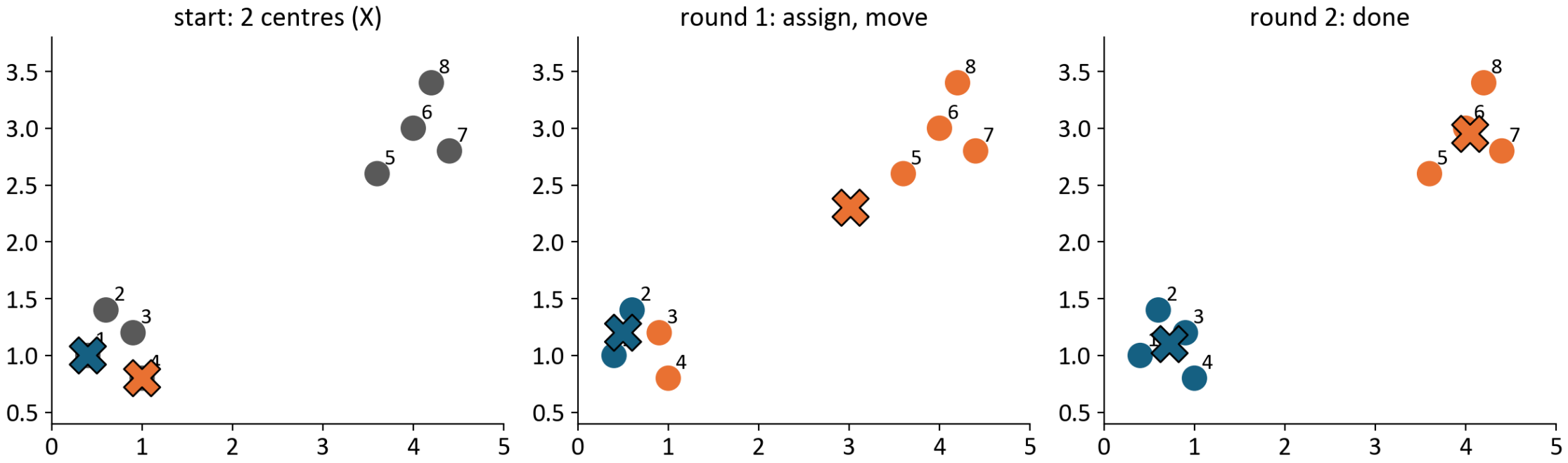
Without z-scoring, the attribute with the biggest numbers decides every distance (Week 3).

The k-Means Algorithm in Five Steps

1	**Choose k**	how many clusters you want
2	**Start**	pick k starting centres (e.g. random farms)
3	**Assign**	each farm joins its NEAREST centre
4	**Update**	move each centre to the MEAN of its farms
5	**Repeat**	until no farm changes cluster

It always stops, but may stop at a poor answer: run several starts (n_init = 10).

k-Means in Pictures: 8 Farms, k = 2



Start badly (both centres on the left) → round 1: assign, move → round 2: farms 3, 4 switch → stop.

Round 1 in Numbers: Assign

IIT Madras Zanzibar

farm	position	distance to c_1	distance to c_2	joins
1	(0.4, 1.0)	0.00	0.63	cluster 1
2	(0.6, 1.4)	0.45	0.72	cluster 1
3	(0.9, 1.2)	0.54	0.41	cluster 2
4	(1.0, 0.8)	0.63	0.00	cluster 2
5 – 8	(3.6 – 4.4, 2.6 – 3.4)	3.6 – 4.5	3.2 – 4.1	cluster 2

Start: c_1 = farm 1 (0.4, 1.0), c_2 = farm 4 (1.0, 0.8). After assigning: {1, 2} and {3, 4, 5, 6, 7, 8}.

Round 1 Update, Round 2, Stop

IIT Madras Zanzibar

step	what happens	result
update c_1	mean of farms 1, 2	(0.50, 1.20)
update c_2	mean of farms 3 – 8	(3.02, 2.30)
round 2 assign	farms 3 and 4 switch to c_1	{1, 2, 3, 4} and {5, 6, 7, 8}
round 2 update	new means	c_1 (0.73, 1.10), c_2 (4.05, 2.95)
round 3	**nobody changes cluster**	**STOP**

k-Means From Scratch: 9 Lines of Python

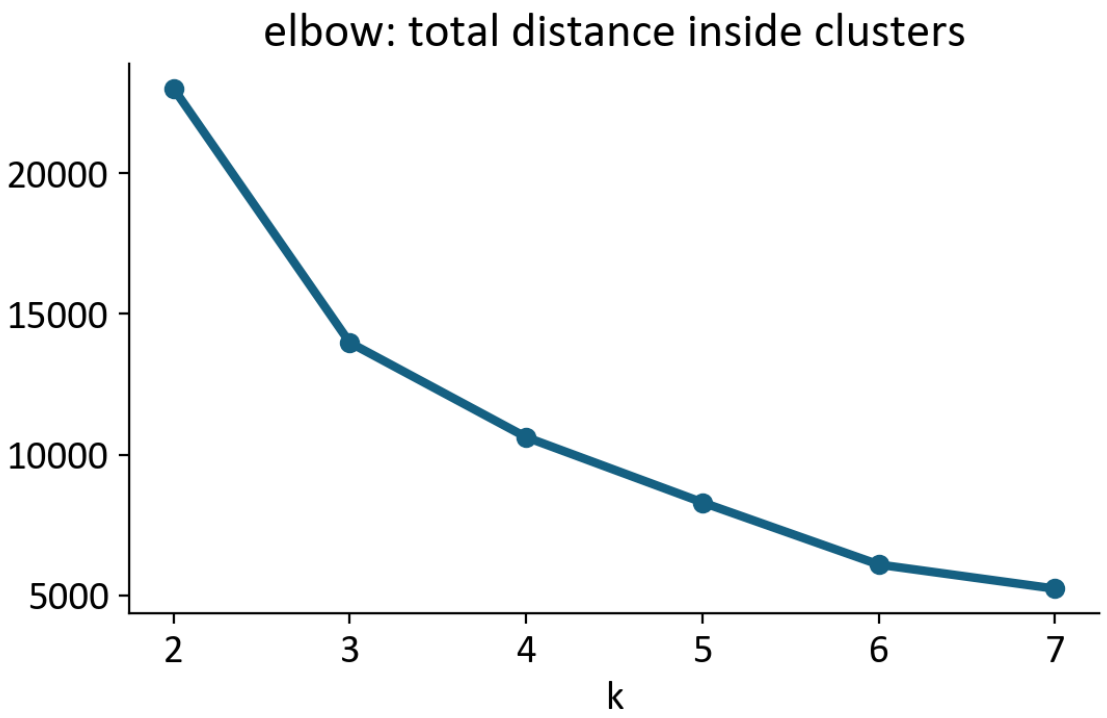
IIT Madras Zanzibar

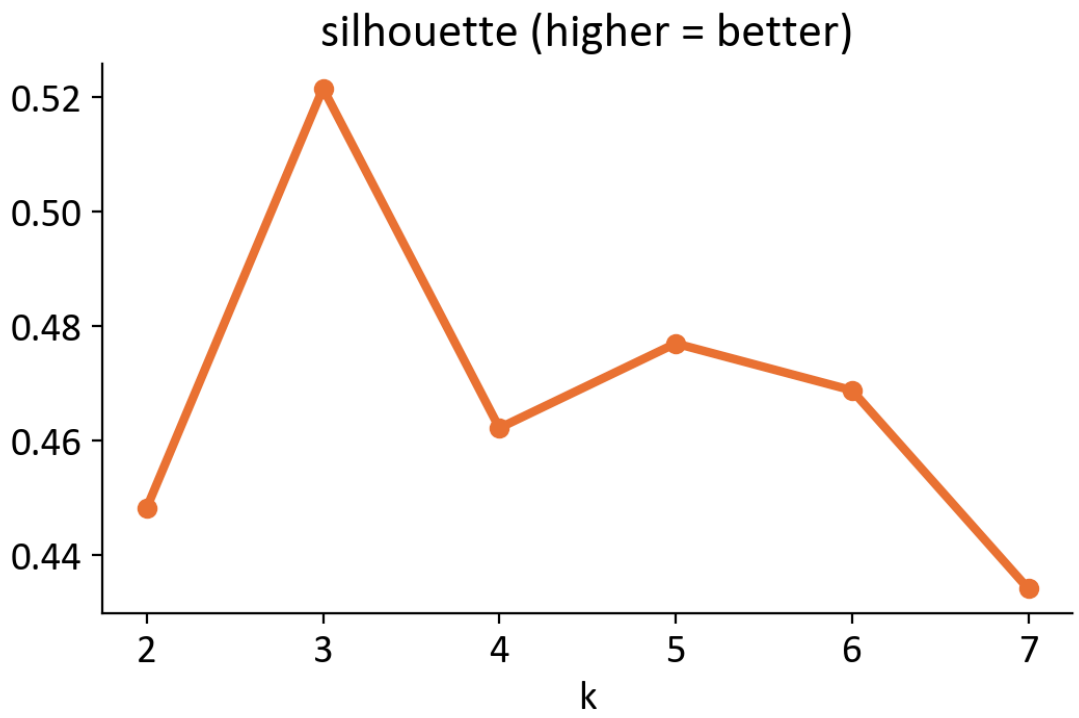
```
1  import numpy as np                          # P = 8 farms as an (8, 2) array
2  c = P[[0, 3]]                               # start: farm 1 and farm 4
3  for round_ in range(10):
4      lab = np.argmin(((P[:, None] - c[None]) ** 2).sum(2), axis=1)   # assign
5      new = np.array([P[lab == k].mean(axis=0) for k in range(2)])      # update
6      if np.allclose(new, c):                 # nothing moved?
7          break                               # converged
8      c = new
9  print(lab + 1, c.round(2))   # [1 1 1 1 2 2 2 2]  [[0.72 1.1] [4.05 2.95]]
```

line	what it does
2	start centres
4	ASSIGN: nearest
5	UPDATE: means
6 – 7	stop if no move
9	same as by hand

Two lines do the real work: assign (line 4) and update (line 5).

Choosing k: Elbow and Silhouette





Elbow bends at 3 – 4; silhouette peaks at k = 3 (0.52), k = 4 close (0.46). Numbers guide, usefulness decides.

The Silhouette in Plain Words

IIT Madras Zanzibar

$$s = (b - a) \div \max(a, b)$$

1. **a** — average distance to farms in MY cluster (want small)
2. **b** — average distance to the NEAREST other cluster (want big)

near 1

deep inside its cluster

near 0

on the border

negative

probably in the wrong cluster

k-Means on 10,000 Farms in Python

IIT Madras Zanzibar

```
1  from sklearn.cluster import KMeans; from sklearn.preprocessing import StandardScaler
2  cols = ["log_size", "avg_yield_z", "fert", "irr"]
3  Z = StandardScaler().fit_transform(farms[cols])            # z-score
4  km = KMeans(n_clusters=4, n_init=10, random_state=0).fit(Z)  # 10 starts
5  farms["cluster"] = km.labels_
6  print(silhouette_score(Z, km.labels_, sample_size=5000))    # ≈ 0.46
7  print(farms.groupby("cluster")[["farm_size_ha", "fert", "irr", "avg_yield_z"]].mean())
```

line	what it does
3	z-score
4	k = 4, 10 starts
5	cluster per farm
7	PROFILE them

Line 7 matters most: describe each cluster in the original units so people can act.

Profile the Clusters, Then Name Them

IIT Madras Zanzibar

name	farms	median size	fertilizer	irrigated	yield vs crop (z)
Irrigated achievers	1,373	1.78 ha	63%	100%	+0.62
Mid-size, no inputs	3,004	1.17 ha	0%	0%	−0.16
Small, no inputs	2,362	0.45 ha	2%	2%	−0.54
Fertilized, rain-fed	3,261	1.16 ha	100%	0%	+0.28

Act: subsidise inputs for 'small, no inputs'; fertilizer advice for 'mid-size, no inputs'; learn from the 'irrigated achievers'.

An Honest Check: Clusters Reflect Your Choices

IIT Madras Zanzibar

cluster	commercial	cooperative	smallholder	subsistence
Irrigated achievers	510	393	462	8
Mid-size, no inputs	62	603	2,242	97
Small, no inputs	0	37	1,017	1,308
Fertilized, rain-fed	354	962	1,848	97

Agreement with farm type: adjusted Rand index = 0.13 (1 = same, 0 = chance).
Clusters reflect the attributes YOU chose: a lens, not the truth.

Weaknesses of k-Means, and the Fixes

IIT Madras Zanzibar

k must be chosen

fix: elbow, silhouette, common sense

Depends on the start

fix: several starts; k-means++

Outliers drag means

fix: z-score; k-medoids

Round clusters only

fix: DBSCAN (Wednesday)

Good news: each round is linear in the number of farms, so k-means scales well.

Clustering in the Real World

IIT Madras Zanzibar

Mobile operators

user segments → bundles

Extension services

farm types → targeted advice

Image compression

a photo reduced to k colours

News

articles grouped by topic

Public health

disease hotspots

Market research

survey respondents by attitude

Find groups first, then describe them in plain words so someone can act.

Your Turn (5 minutes, pairs)

IIT Madras Zanzibar

Farm sizes (ha): 1, 2, 3, 8, 9, 10. k = 2. Start centres: $c_1 = 1$, $c_2 = 2$.

1. Assign each farm to its nearest centre
2. Compute the new centres
3. Repeat until nothing changes: final clusters and centres
4. Add a 100 ha farm and rerun from the final centres. What happens, and why?

Solution, Step by Step

IIT Madras Zanzibar

step	working	result
round 1 assign	{1} vs {2, 3, 8, 9, 10}	$c_1 = 1$, $c_2 = 2$
update	$c_2 = (2 + 3 + 8 + 9 + 10) \div 5 = 32 \div 5$	6.4
round 2	{1, 2, 3} vs {8, 9, 10}; means 2 and 9	$c_1 = 2$, $c_2 = 9$
round 3	nobody moves	STOP
add 100 ha	$c_2 = (8 + 9 + 10 + 100) \div 4 = 31.75 \rightarrow$ 8, 9, 10 move to c_1	{1,2,3,8,9,10} vs {100}

Summary

IIT Madras Zanzibar

1	**Unsupervised**	no labels: tight inside, separate between
2	**k-means**	assign to nearest centre, move to mean, repeat
3	**Choosing k**	elbow + silhouette (peak 3: 0.52), then usefulness
4	**Interpret honestly**	profile, name; clusters reflect your attributes (ARI 0.13)

Next: hierarchical clustering: clusters inside clusters, and evaluating clusters.

Questions (5 minutes)

IIT Madras Zanzibar

Q: Does k-means always give the same answer?

A: No: it depends on the start. Use several starts (n_init = 10).

Q: Can it use categories like crop?

A: Not directly; use 0/1 columns or k-modes.

Q: Is the best-yield cluster 'the best farms'?

A: Only on the attributes chosen; not profit or sustainability.

Q: How big can the data be?

A: Millions fit; Wednesday: methods for data beyond memory.