19/07/2025

1155184178 ZHOU Wan Ieng

Clustering

Why ?

-- Better organization, easy to find items and fast searching

Eg. Patient are put into different groups and assigned to speciality hospital –for example elderly stroke centre [aim to optimize the need for that specific group]

Clustering in biology

Eg. Cluster in gene

- Co-expressed : same cluster
- Diff- expressed: diff cluster

Cluster in cell

- Identify new disease subtypes : eg. diff cancer stages / different treatment = more effective treatment

❖Summarization

➢Reduce the size of large data sets

➢Preserve privacy (blur individual information , instead representing them as 1 big cluster and this prevent data leak)

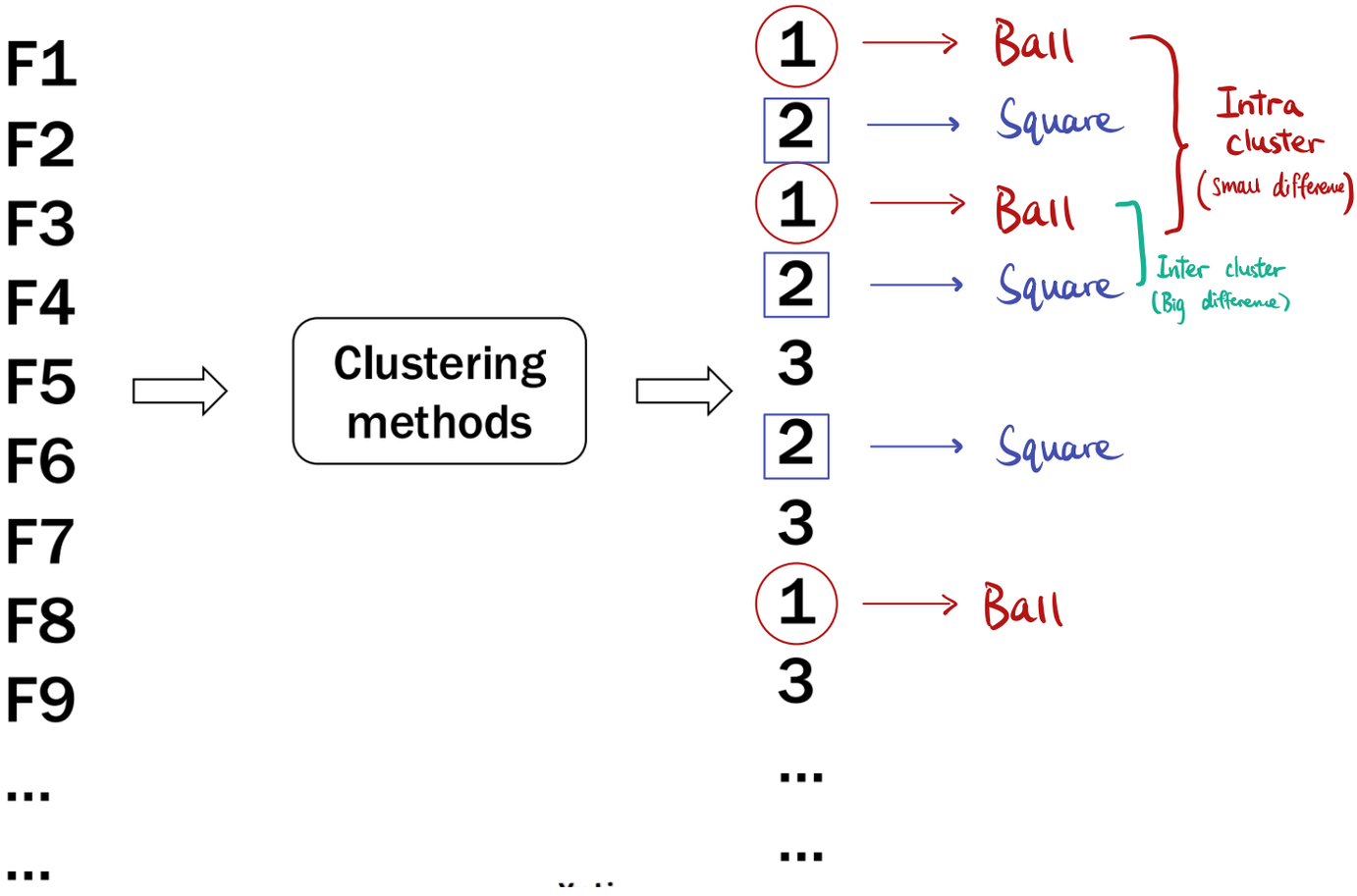
What?

Related to objects in the same group = 1 cluster [Intra-cluster differences are small]

Different from objects in the same group = belongs to other cluster [Inter-cluster differences are large]

How ?

Each pixel represent a value



** this clustering method Is treated as a blackbox for now

Correctness check

Parameters in clustering

➢**Data to be clustered : check data type and normalize**

➢**Similarity measurement :** to define the level of difference between clusters

➢**Clustering algorithm (the executive procedure) : can be reproduce**

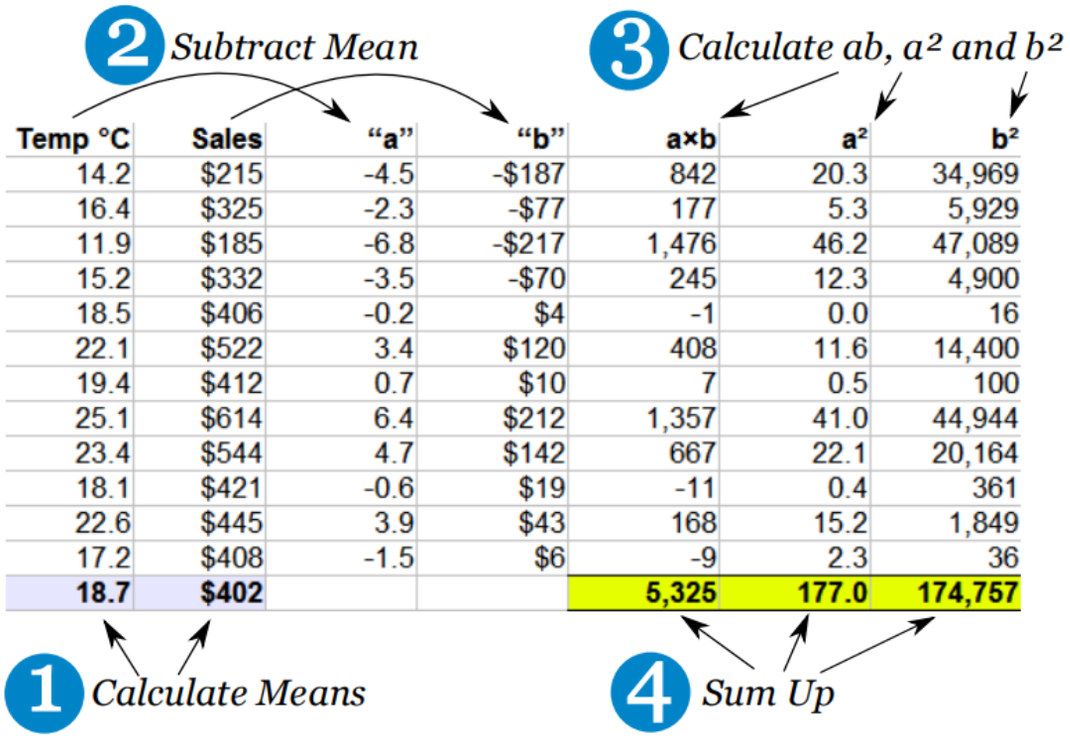
Define

- **Similarity : how alike range [0,1] 1 –higher = same**
- **Dissimilarity: how different 0=unalike**

Similarity measurement method :

Qualitative e.g. level of gene expression
↓
Quantitative e.g. distance

1. **Cosine similarity:**
2. **Pearson correlation coefficients : linear relationship / -1~1**



Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

5 $\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$

** temp and ice cream sale has a high positive correlation = + 0.9575

3. Euclidean distance : Normalization before calc. , scale must be the same

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Multiple pairs, nxn matrix (n = number of attributes)

4. Minkowski distance : r can be more than 2, Generalized formula

$$dist(\boldsymbol{p},\boldsymbol{q}) = \left(\sum_{k=1}^{m} |p_k - q_k|^r\right)^{\frac{1}{r}}$$

❖r = 1. Block distance eg. lrl, city is divided into blocks and avenues / difference along x axis and y axis , no matter which route taken will always be the same [total number of differences]

❖r = 2 Euclidean distance eg. In city, transport route is unlikely to be straight line [overall similarity]

❖$r \to \infty$. Supreme distance : take limit, max along all dimensions [maximum difference between any component of the vectors]

> What is the Minkowski distance between the below two vectors, when r is equal to infinite?
>
> X=[1,2,3,4,5], Y=[5,4,3,2,1]
>
> A. 1
>
> B. 2
>
> C. 3
>
> D. 4
>
> E. 5
>
> **You have submitted: 4/D**

5-1=4