Name: Chiong Yan Lok
SID: 1155193602

Scribing: Lecture 7 – Clustering

1. Clustering

Why clustering?

1. Cluster items
   - ➢ Better organization: helps in arranging information or items systematically.
   - ➢ Faster searching: enables quicker retrieval of information
2. Cluster people
   - ➢ Patients: different treatment for different groups
     - Children, elderly
   - ➢ Customers: different groups with different needs
     - Not necessarily grouping the people by age or gender
     - Optimize the product based on the need of the targeting group
3. Cluster genes
   - ➢ Identify co-expressed genes
     - Involved in the same pathway
   - ➢ Identify differentially expressed genes
     - Related to diseases
4. Cluster samples/cells
   - ➢ Identify new disease sub-types
   - ➢ Identify new cell types

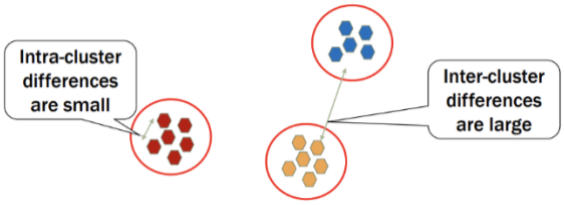
What is clustering analysis?

Definition:

- ➢ Finding groups of objects such that the objects in a group are similar (or related) to one another and different from those in other groups.

Key Concepts:

- ➢ Intra-cluster differences are small
- ➢ Inter-cluster differences are large



Application of clustering analysis:

- As a stand-alone tool to get insight into data distribution
- As a pre-processing step for other algorithms
- Examples:
  - group related documents for browsing
  - group genes and proteins that have similar functionality
  - group stocks with similar price fluctuations
  - discover new groups (cell types)

Summarization:

- Reduce the size of large data sets
- Preserve privacy (e.g., in medical data)

What are needed to do clustering?

1. Data to be clustered
2. Similarity measurement
3. Clustering algorithm (the executive procedure)

2. Similarity and dissimilarity measurement

Similarity and dissimilarity

Similarity

- Numerical measure of how alike two data objects are
- Higher when objects are more alike
- Often falls in the range [0,1]

Dissimilarity (distance)

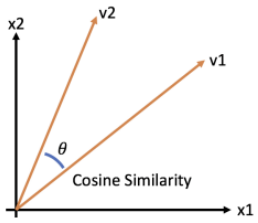
- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

Cosine similarity

- If d_1 and d_2 are two vectors, then

$$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$$

- Where $\cdot$ indicate vector dot product and $|d|$ is the length of the vector d
- A cosine similarity of 1 indicates that the vectors are identical in direction, while a cosine similarity of 0 indicates that they are orthogonal (completely dissimilar).



➢ Example

d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

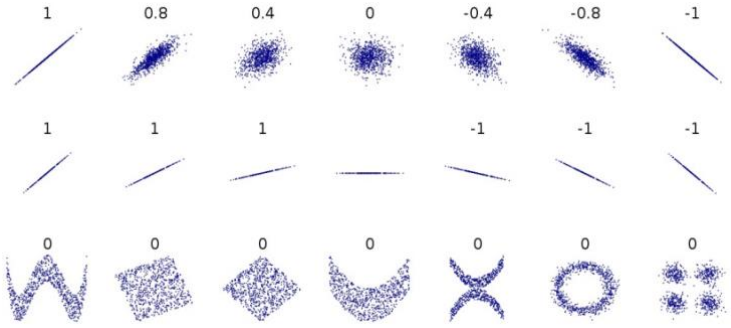
$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481
$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

cos(d_1, d_2) = 0.3150

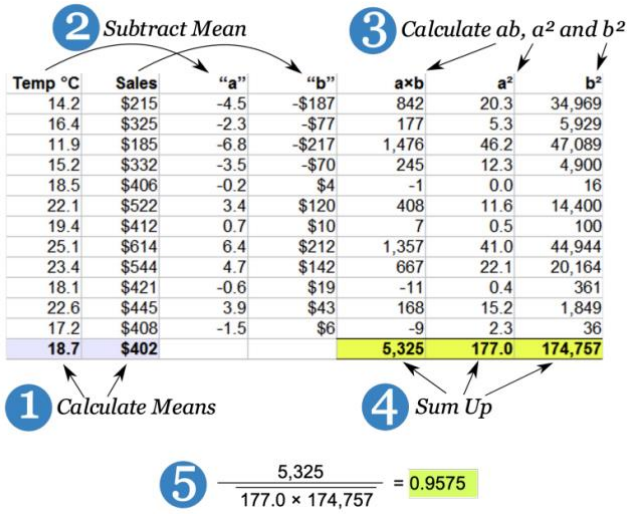
Correlation

➢ Correlation measures the linear relationship between objects

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$



➢ Example

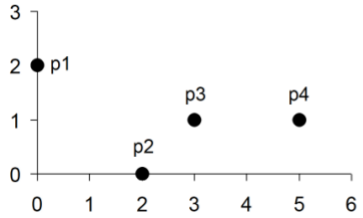


Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
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**

Euclidean distance

$$Ed(\boldsymbol{p},\boldsymbol{q})=\sqrt{\sum_{k=1}^{m}(p_k-q_k)^2}$$

- Where m is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.
- Normalization is necessary, if scales of different dimension differ
- Example



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	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

Minkowski distance

- Minkowski Distance is a generalization of Euclidean Distance

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

- Where r is a parameter, m is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.
- Different cases of Minkowski Distance:
  - r = 1. City block (Manhattan, Taxicab, L_1) distance
    - e.g. Hamming distance, which is the number of bits that are different between two binary vectors
  - r = 2. Euclidean distance

- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance
  - This is the maximum difference between any component of the vectors

➢ Example

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	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

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

3. Hierarchical clustering

Hierarchical clustering

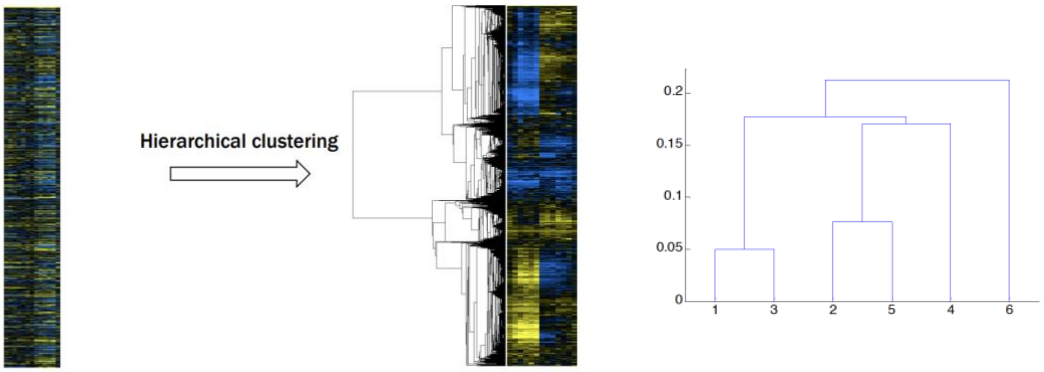
➢ Produces a set of nested clusters organized as a hierarchical tree

➢ Can be visualized as a dendrogram

➢ A tree like diagram that records the sequences of merges

➢ They may correspond to meaningful taxonomies

➢ Gene clusters, phylogeny reconstruction, animal kingdom…

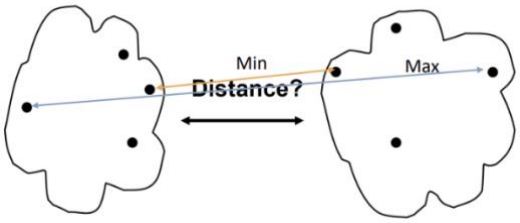


Steps of hierarchical clustering

1. Compute the Similarity or Distance matrix
2. Let each data point be a cluster
3. Merge the two closest clusters
4. Update the similarity or distance matrix
5. Repeat step 3 and step 4 until only a single cluster remains

Methods to update the distance matrix after merging?

- Min
- Max
- Group average
- Distance between centroids



A running example

Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

1. Use correlation (linear correlation) to compute the data matrix

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

   Then, we will get this

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770	1				
At1g30720	0.9733	1			
At4g27450	-1	-0.9733	1		
At2g34930	0.9493	0.9909	-0.9493	1	
At2g05540	0.5774	0.562	-0.5774	0.4528	1

2. Let each gene be a cluster and remove the 1 in the matrix
3. Merge the two closest clusters, At1g30730 and At2g34930, as the correlation coefficient between them is the largest (0.9909)

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733			
At2g34930	0.9493	0.9909	-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528	

4. Update with minimum distance (largest correlation)

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733 ->-0.9493			
At2g34930	0.9493 ->0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528 ->0.562	

5. Merge At2g34930, At1g30720 and At4g35770

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9493			
At2g34930	0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.562	

6. Update with minimum distance (largest correlation)

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-1 ->-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.562 ->0.5774	-0.5774	0.562 ->0.5774	

7. Merge At2g34930, At1g30720, At4g35770, and At2g05540

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.5774	-0.5774	0.5774	

8. Update with minimum distance (largest correlation)

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.5774	-0.5774			
At2g34930			-0.5774		
At2g05540			-0.5774		