

Clustering

Yu LI

Friday, 26 September 2025

Clustering Analysis

[using similarity or dissimilarity to measure the differences]

- Intra-cluster differences: small; Inter-cluster differences: large
- Common measures:
 - o Cosine similarity
 - o Correlation
 - o Euclidean distance
 - o Manhattan distance
 - o Mahalanobis distance

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:

1. Compute the Similarity or Distance Matrix
2. Let each data point be a cluster
3. Repeat the following steps until only a single cluster remains:
 - o Merge the two closest clusters.
 - o Update the Similarity or Distance Matrix: 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

Linear correlation:

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



Visualization after normalization

1. each gene be a cluster

	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	

2. merge At2g34930 and At1g30720

3. 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	

	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	

4. merge At2g34930, At1g30720 and At4g35770

5. update with minimum distance (largest correlation)

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

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

6. merge At2g34930, At1g30720, At4g35770, and At2g05540

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
Node2					
Node1					
Node3					
At4g35770					
At1g30720					
At4g27450	-0.5774	-0.5774			
At2g34930			-0.5774		
At2g05540			-0.5774		

About programming

Scikit-learn: <https://scikit-learn.org/stable/>

Hierarchical clustering in Python: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram

Potential project-2

Data preprocessing for the gene expression matrix

- Data collecting and merging (if needed)
- Exploration
- Visualization
- Data cleaning
- Get distance matrix
- Perform clustering

Dimension reduction (Lec-11)

Advanced Topic

Mahalanobis distance

[Calculating distance considering the data distribution]

$$mahalanobis(p, q) = (p - q)^T \Sigma^{-1} (p - q)$$

where Σ is the covariance matrix

inverse of the covariance matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

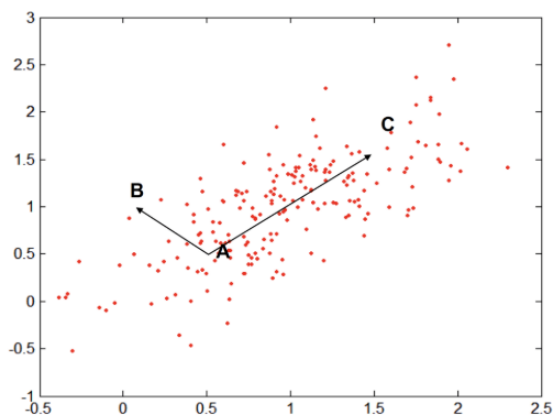
↑
determinant

$$\begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}^{-1} = \frac{1}{4 \times 6 - 7 \times 2} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix}$$

$$= \frac{1}{10} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{bmatrix}$$

example 1:



Covariance matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

*C looks geometrically farther from A than B does, but Mahalanobis distance from A to C (4) is smaller than to B (5).

example 2:

Quiz	Standard deviation	
1	10	The difference between Student A and Student B is 1 point.
2	1	The difference between Student C and Student D is 1 point.

Should a 1-point difference in Quiz 1 be considered the same as a 1-point difference in Quiz 2?
To compare differences fairly, see how many standard deviations apart the two points are.

Quiz 1: 1pt/10std=0.1 standard deviations

Quiz 2: 1pt/1std=1 standard deviation

Conclusion: the 1pt difference in quiz 2 is 10 times more significant than in quiz 1.

➔ The Mahalanobis distance- this intuition formalized

Mahalanobis distance & Normalization

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

↓ Min-max normalization

Person	Height	Weight
P1	0.625	0.875
P2	0	0
P3	0.25	0.375
P4	1	1

the two features (height and weight) are in different scales (meters and kilograms) which can distort distance calculations.

→ utilize min-max normalization for pre-processing

Resources and uncovered topics

- introduction to data mining: chapter 2.4 & 8
- k-means clustering
- density-based clustering
- how to determine the number of clusters
- how good is your clustering (lec8-9)