BMEG3105 Fall 2025

Lecture 8 – Clustering

Lecturer: Yu Li(李煜) from CSE SID: 1155212306

Clustering

Friday, 26 September 2025

1. Outline of lecture 8
   1.1. Hierarchical clustering
   1.2. Mahalanobis distance (Advanced topic)

2. Recap of lecture 6
   2.1. Clustering analysis

Clustering is used to find object that have very similar feature and divide them into groups and put different object in different groups. Therefore, clustering can help us to organise data easier and allow faster searching of data by reducing a large set of data into more simple and optimized data set.

   2.2. Minkowski distance and Euclidean distance

Euclidean distance is used to calculate the similarity of different data and plot them inside a graph. Equation is below:

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

The smaller the distance between the two points and the straight line, the more similar the data are.
Minkowski distance is the generalization of Euclidean distance. The equation is below:

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

For Example:
R = 1 is City block distance
R = 2 is Euclidean distance
R = Infinity is “Supremum” distance

3. Hierarchical clustering

This method of clustering produces a set of nested cluster organised as a hierarchical tree and it can be visualised as a dendrogram

Steps of clustering:

I. Compute the similarity or Distance matrix
II. Let each data point to be a cluster
III. Merge the two closest clusters
IV. Update the matrix until only a single cluster remain

Method to update the matrix:

- Max
- Min
- Groups average
- Distance between centroid

Running Example and correlation:

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

Correlation equation:

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

Step1: Let each gene to be a cluster, and compute the matrix

	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

Step2: Merge At1g30730 and At2g34930 and update the matrix

	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	

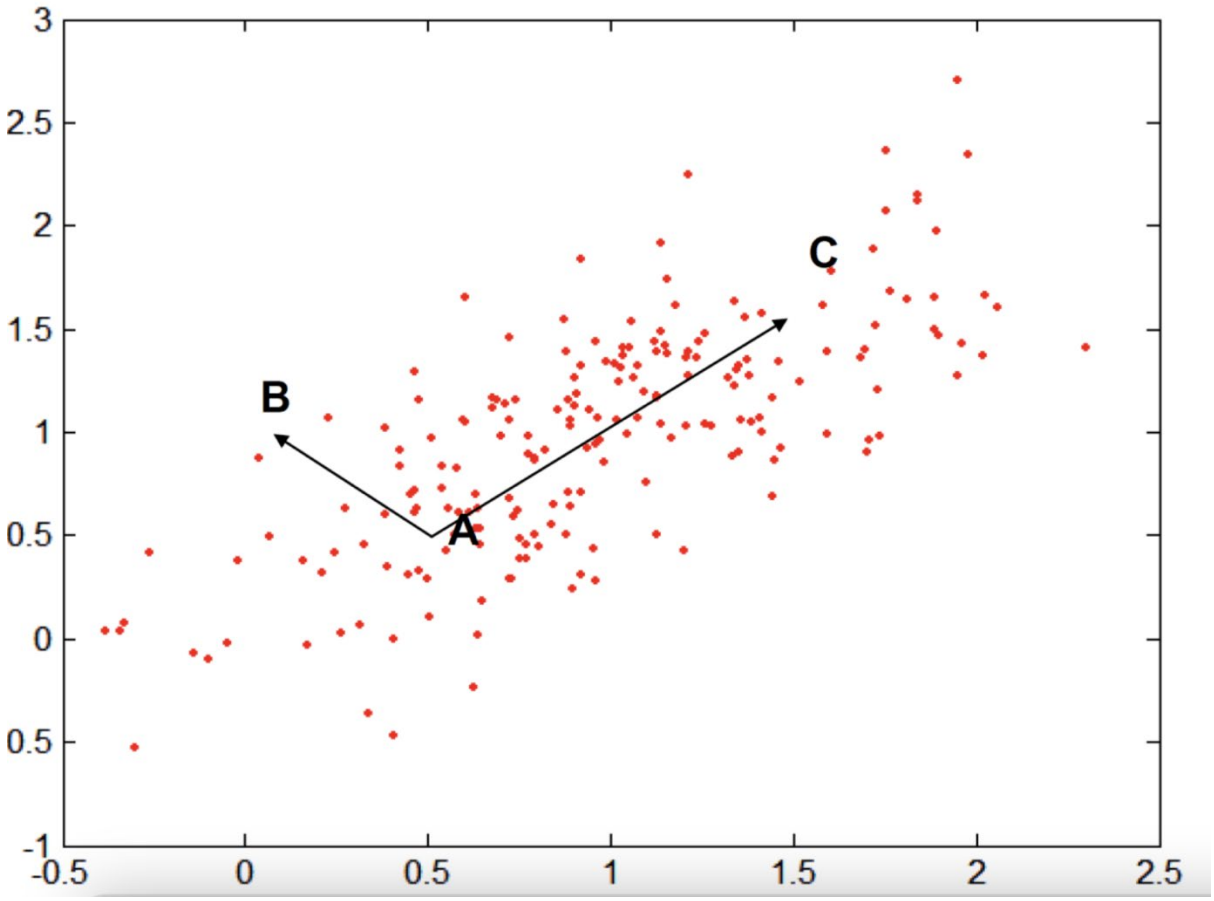
Step3: Merge the above two with At4g35770 and update the matrix

	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	

4. Mahalanobis distance

This method calculate the distance with respect to the data distribution. Below is the equation:

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



From the graph, we can see that A located at (0.5,0.5), B located at (0,1) and C located at (1.5,1.5)

Mahal(A,B) = 5
Mahal(B,C) = 4

5. Resource and Uncovered Topic
    - Introduction to data mining: Chapter 2.4 & Chapter 8
    - K-means clustering
    - Density-based clustering
    - How to determine the number of cluster
    - How good is your clustering (lecture8-9