

BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine

Lecture 6: Clustering

Friday, 19th September 2025

Lecturer : Yu LI (李煜)

Scriber : VERMINSHU, Britney Alexandra (1155206525)

Agenda(s) :

1	Recap from last lecture
2	Clustering
3	Similarity and dissimilarity measurement
4	Hierarchical clustering

1. Recap

a. Sequence Mapping

- Mapping: slide each read along the genome and calculate the difference, or use dynamic programming to calculate the alignment score for each read.

b. Data Cleaning and Exploration

- Steps of data cleaning: denoise, remove outliers, handle missing data, remove duplicates, encode categorical data, normalize data.
- Summary statistics: mean, median, range, variance, percentiles, mode.
- Visualization: histograms and box plots.

2. Clustering

- Cluster items for better organization that enables faster searching, e.g., online shopping website.
- Cluster people: for different group of people with different treatments, e.g., hospital patients (children, elders), or people with their different needs (not necessarily by age or gender), e.g., customers.
- Cluster genes to identify co-expressed genes involved in the same pathway to fulfil some specific functions, or to identify differentially expressed genes related to diseases.
- Cluster samples/cells to identify new disease sub-types, e.g., cancer stages (where early treatment will result in a higher rate of recovery), or to identify new cell types.
- Clustering analysis is finding groups of objects such that the objects within a group are similar or related to each other (intra-cluster) and different from the objects in other groups (inter-cluster).

- To understand data: insight into data distribution, pre-processing step for other algorithms.
 - To summarize data: reduce the size of large data sets, which then helps preserve patients' privacy.
 - Clustering with a computer: design a program to assist with grouping large amount of data sets, i.e., input : data $\rightarrow$ clustering methods $\rightarrow$ output : clustering indicator (cluster ID).
 - Elements for clustering: data (to be clustered), similarity measurement, and clustering algorithm (the executive procedure).
- *) Aligning two sequences is not clustering itself, but a step towards clustering.

3. Similarity and dissimilarity measurement

- Similarity: measures how alike two data objects are, in the range $[0,1]$.
- Dissimilarity/distance: measures how different two data objects are, with a minimum of 0.

a. Cosine Similarity

- Calculate the cosine of the angle between two vectors.
- 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 $\cdot$ indicates vector dot product and $|d|$ is the length of the vector d .

b. Correlation (Pearson Correlation Coefficient)

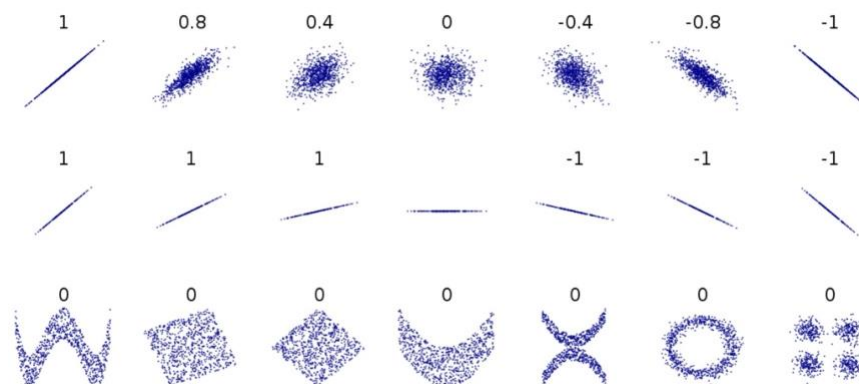
- Measures the linear relationship between two data objects.
- Defined as:

$$\rho_{X,Y} = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

where: $\text{cov}(X,Y)$ is the covariance of X and Y ;

$\sigma_X \sigma_Y$ is the product of standard deviation of X and Y .

- $\rho_{X,Y} = 1$ indicates a perfect positive linear relationship between X and Y .
- $\rho_{X,Y} = -1$ indicates a perfect negative linear relationship between X and Y .
- $\rho_{X,Y} = 0$ indicates no linear relationship between X and Y .



c. Euclidean Distance

- Calculate the length (distance) of a line between two points.
- Defined as:

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

where: m is the number of dimensions (attributes);

p_k and q_k are the k -th attributes (components) or data objects p and q respectively.

- Example:
 - Distance matrix: to measure the distance between a pair of data points ($n \times n$ matrix).
 - $p_1 = (0,2)$; $p_2 = (2,0)$

$$Ed(p_1, p_2) = \sqrt{(0 - 2)^2 + (2 - 0)^2} = \sqrt{8} = 2.828$$

*) If attributes (dimensions) differ in scale, it is necessary to normalize the data.

d. Minkowski Distance

- Is a generalization of Euclidean distance.
- Defined as:

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

where: r is a parameter;

m is the number of dimensions (attributes);

p_k and q_k are the k -th attributes (components) or data objects p and q respectively.

- Special cases:
 - Manhattan distance ($r = 1$)
 - Also known as the L_1 norm, taxicab or city block distance.
 - Is the sum of the absolute differences of the coordinates.
 - Defined as:

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

- Example: $p_1 = (0,2)$; $p_2 = (5,1)$
 $L_1 = |0 - 5| + |2 - 1| = 6$

- Euclidean distance ($r = 2$)

- Supremum distance ($r \rightarrow \infty$)

- Also known as L_{max} norm, L_{∞} norm.
- Is the maximum difference between any component of the vectors.
- Calculate the maximum absolute differences between any coordinates of the two points.

- Defined as:

$$\text{dist}(p, q) = \max_{k=1}^m (|p_k - q_k|)$$

- Example: $p_1 = (0, 2)$; $p_2 = (5, 1)$

$$L_\infty = \max(|0 - 5|, |2 - 1|) = \max(5, 1) = 5$$

- Visual representations of different cases of Minkowski distance:
 - Manhattan distance (red)
 - Euclidean distance (blue)
 - Supremum distance (green)



4. Hierarchical measurement

- This agenda will be covered in Lecture 7.

Disclaimer: all figures are adapted from Prof. Li BMEG3105 Lecture 6 Notes