BMEG3105 Data Analytics for Personalized Genomics and Precision Medicine
Lecture 8: Clustering

26 September, 2025

Lecturer: Yu LI
Scriber: Lan Hoi Lee(1155191196)

1 Recap

1.1 Sequence Mapping

Approach: Shift each read across the genome and compute the degree of difference. In each iteration, dynamic programming might be applied to calculate this difference.

1.2 Data Exploration and Cleaning

- Data cleaning: Involves reducing noise, eliminating outliers, addressing missing data, removing duplicate entries, and normalizing the dataset.
- Data exploration: Includes generating summary statistics—such as the mean, median, range, variance, and percentiles.
- Visualization: Common methods include histograms and box plots.

1.3 Percentiles

For an ordinal or continuous attribute x and a value p (which falls between 0 and 100), the p-th percentile is a specific value of x (marked as xp). Its defining feature is that p% of the observed values of x are smaller than xp.

First, arrange the N values of attribute x in descending order. The value located at the position of $N \times (1 - p/100)$ is the p-th percentile. When p is 50, x_{50} (the 50th percentile) is approximate to the median.

2 Introduction to Clustering

2.1 Why Clustering?

- Clustering items: Helps achieve better organization and faster searching.
- Clustering people: Enables identifying different needs for distinct groups.

- Clustering in biology:

1. Clustering genes to recognize co-expressed genes or differentially expressed genes.
2. Clustering samples or cells to discover new disease sub-types or cell types.

2.2 What is Clustering?

Clustering refers to the process of identifying groups (clusters) of objects, where the objects within the same group (intra-cluster) share high similarity with one another, while the objects from different groups (inter-cluster) exhibit clear differences.

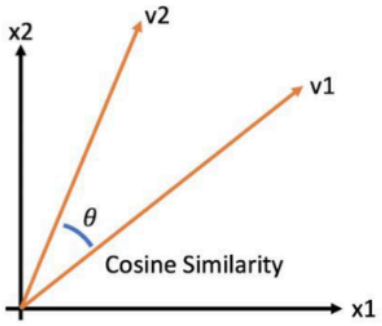
3 Similarity and Dissimilarity

- Similarity: Quantifies the degree to which two data objects resemble each other. It typically falls within the range of [0,1], where higher values indicate greater similarity.
- Dissimilarity (also called Distance): Quantifies the degree to which two data objects differ from each other. Its minimum value is 0, meaning the two objects are identical.

3.1 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 · indicate vector dot product and $|d|$ is the length of the vector d



3.2 Correlation

The correlation coefficient is defined by the following formula:

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

Where:

- Cov(X, Y) represents the covariance between variables X and Y.
- σ_x and σ_γ denote the standard deviations of variables X and Y, respectively.

3.3 Euclidean distance

Euclidean distance quantifies the straight-line distance between two points in a Euclidean space (e.g., 2D, 3D, or higher-dimensional space).

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

- n: Represents the number of dimensions (i.e., the number of attributes) of the data objects X and Y.
- Xi and Yi: Respectively stand for the i-th attribute (or component) of the data objects X and Y (where i ranges from 1 to n).

3.4 Minkowski distance

Minkowski Distance serves as a generalized form of Euclidean Distance. Its formal definition is as follows: For two data objects X and Y in an n-dimensional space, the Minkowski Distance d(X, Y)between them is:

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

Where the parameters are defined as:

- p: A parameter that specifies the type of distance (different values of p correspond to different specific distance metrics derived from Minkowski Distance).
- n: The number of dimensions (i.e., the number of attributes) of the data objects X and Y.

- Xi and Yi: Respectively represent the i-th attribute (or component) of the data objects X and Y (with i ranging from 1 to n).