BMEG3105 - Fall 2024

Lecture 7: Clustering

Lecture Date: 27 Sept. Deadline: 04 Oct. 11:59 p.m.

Lecturer: Prof. LI Yu *Scribe: LIU Linqi*

1 Recap from Last Lecture

1.1 Sequence Mapping

- Method: Slide each read along the genome, calculate the difference.
- Each time, we may use dynamic programming to calculate the difference.

1.2 Data Exploration and Cleaning

- Data cleaning: Denoise, remove outliers, handle missing data, remove duplicates, and normalize data.
- Data exploration: Summary statistics, including mean, median, range, variance, percentiles.
- Visualization: Histograms and box plots.

1.3 Percentiles

- Given an ordinal or continuous attribute x and a number p between 0 and 100, the p-th percentile is a value of x such that $p\%$ of the observed values of x are less than x_p.
- Sort N values of attribute x in decreasing order. The $N \times (1 - p/100)$-th value corresponds to the p-th percentile.
- When $p = 50$, x_{50} is close to the median value.

2 Introduction to Clustering

2.1 Why Clustering?

- Cluster items: Better organization, faster searching
- Cluster people: Different needs for diffirent groups
- Cluster in biology:

    - Cluster genes to identify co-expressed or differentially expressed genes.
    - Cluster samples or cells to identify new disease sub-types or cell types.

2.2 What is Clustering?

Definiation Clustering is about finding groups of objects that are similar to each other within the group (intra-cluster) and different from other groups (inter-cluster).

2.3 How to Do Clustering?

$$\text{Data} \xrightarrow{\text{Clustering methods}} \begin{array}{c}\text{Output:}\\ \text{clustering indicator}\end{array}$$

3 Similarity and dissimilarity measurement

- Similarity: Measures how alike two data objects are, often in the range [0,1].
- Dissimilarity (Distance): Measures how different two objects are, with a minimum of 0.

3.1 Cosine Similarity

If $\mathbf{d_1}$ and $\mathbf{d_2}$ are two vectors, then the cosine similarity between them is defined as:

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

where $\cdot$ denotes the dot product of the vectors, and $|\mathbf{d}|$ represents the magnitude of vector $\mathbf{d}$.

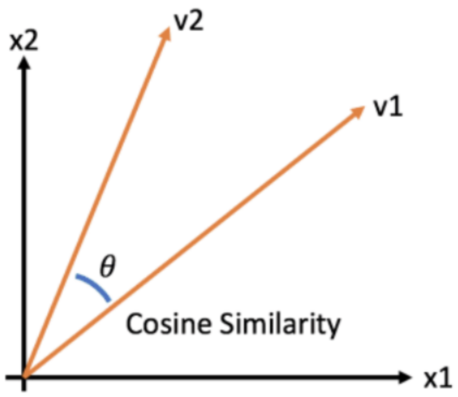

Figure 1: Example of cosine similarity in 2D space. [1]

3.2 Correlation

The correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}$$

where:

- $\text{Cov}(X,Y)$ is the covariance between X and Y,
- σ_X and σ_Y are the standard deviations of X and Y respectively.

Correlation measures the linear relationship between objects.

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

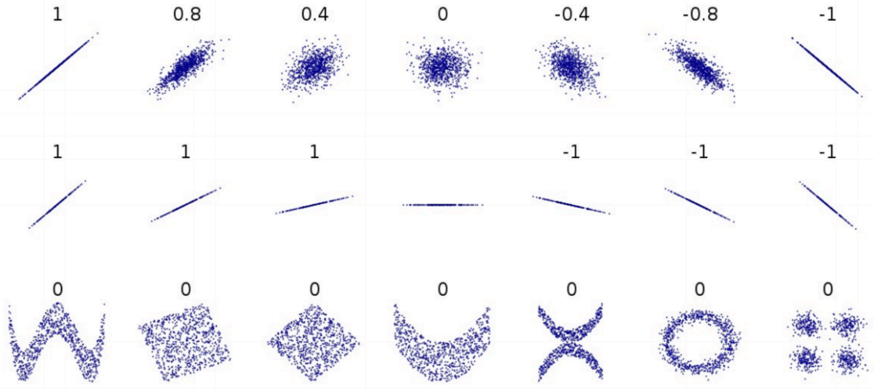

Figure 2: Different correlation coefficients and their corresponding scatter plot shapes. [1]

3.3 Euclidean Distance

Euclidean distance measures the straight-line distance between two points in Euclidean space. It is defined as:

$$Ed(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

where:

- n is the number of dimensions (attributes).
- X_i and Y_i are, respectively, the i-th attributes (components) or data objects X and Y.

3.4 Minkowski Distance

Minkowski Distance is a generalization of Euclidean Distance. It is defined as:

$$d(X,Y) = \left(\sum_{i=1}^{n} |X_i - Y_i|^p \right)^{\frac{1}{p}}$$

where:

- p is a parameter that determines the type of distance.
- n is the number of dimensions (attributes).
- X_i and Y_i are, respectively, the i-th attributes (components) or data objects X and Y.

Manhattan Distance When $p = 1$, this represents the City block (Manhattan, taxicab, L^1 norm) distance:

$$d(X,Y) = \sum_{i=1}^{n} |X_i - Y_i|$$

Euclidean Distance When $p = 2$, it represents the Euclidean distance.

Supremum Distance As $p \to \infty$, it becomes the supremum distance (L^∞ norm), which is defined as:

$$d(X,Y) = \max_i (|X_i - Y_i|)$$

This represents the maximum difference between any component of the vectors.

4 Hierarchical Clustering

This topic would be covered in the next lecture.

References

[1] Li, Yu (2024). *BMEG3105: Data analytics for personalized genomics and precision medicine Clustering*.