BMEG 3105 Fall 2024

Data analytics for personalized genomics and precision medicine

Lecture 7: Clustering

Lecturer: Yu LI (李煜) from CSE

Scriber: KANNAPPAN, Kannammai (1155163190)

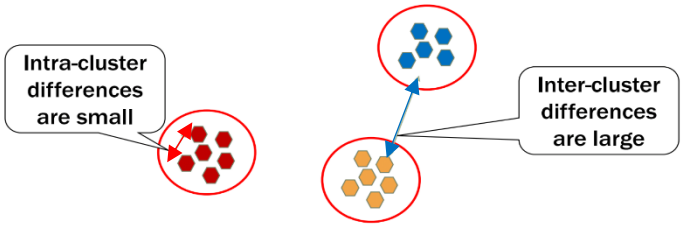
Friday, 27th September 2024

1. Clustering Introduction:

"Finding groups of objects such that the objects in a group will be similar (related) to one another and different from (or unrelated to) the objects in other groups."

Clustering is a process to group objects with similar properties/characteristics which can help to:

- Understand data: insight into data distribution, pre-processing step for many algorithms
- Summarize data: reduces large data sets into several groups, preserve patient privacy



Why is it important to cluster?

General Applications:	Biological Application:

- Better organisation → Faster searching.
- Example: shopping online by category.
 	
People:
- Clustered by age, gender, etc.:
  - For analysing treatment based on cluster of people.
- Cluster by interest:
  - To target products based on need of group.
	Genes:
	- Can identify co-expressed genes:
	  - Suggesting similar function/pathway
	- Can identify differentially expressed genes:
	  - Gene correlation to disease cause.
	
	Cells/Samples:
	- Can identify new disease type:
	  - Develop personalised treatments for this group.
	- To identify new cell types.

2. Clustering with a computer:

Order of process:

1. Feed the computer with data
2. Label each piece of data
3. Pass through clustering algorithm.
4. Output is data grouped with a clustering ID representing each group.
5. Check the efficiency of clustering

Like sequence comparison we need 3 things to cluster data:

- Data
- Similarity Measurement (Criteria for grouping objects)
- Clustering algorithm (Operational method)

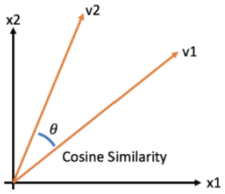
3. Similarity and Dissimilarity Measurement

Quantify how alike 2 objects are
- Higher = more alike
- Range: [0,1]

Quantify how different 2 objects are
- Lower = more alike
- Range: [0, upper limit depends on data]

A. Cosine Similarity:

Measures the similarity between two vectors, d_1 and d_2:
- By calculating the angle between them and using the cosine of the angle between the two vectors.
- As since cos (0) = 1 then both vectors align therefore similar and can cluster together.



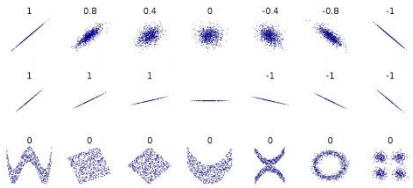
Formula:

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

· is vector dot product
|d| is the length of vector

B. Correlation:

Measuring linear relationship between data points:
- By measuring how likely to variables change together e.g. if x increases and y also increases/decreases then higher correlation.
- If data similar, then correlation is close to 1 (positively correlated) or -1 (negative correlated) then can cluster together.
- No similarity then correlation = 0.



Formulas:

Covariance

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

Expectation of X and Y deviation from mean

Standard deviation

$$corr(X,Y) = \frac{(X - \mu_X)(Y - \mu_Y)}{(X - \mu_X)^2 (Y - \mu_Y)^2}$$

C. Minkowski Distance:

Measuring distance between two points (p, q) in a m dimensional space.
- Generalised formula of the Euclidean distance (dimension = 2)
- This tool can be used to measure the dissimilarity of the two points. Therefore smaller distance between data points means they can be clustered together.

Formula:

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

r = parameter
m= number of dimensions (attributes)
p_k/q_k = Value of p/q in the k-th dimension (attribute)

Types of Minkowski distance when r is fixed:

- City block/Manhattan/Taxicab L_1 norm distance
  - r=1
  - Calculates the sum of the distance in each dimension
  - $dist\ (p,q) = \sum_{k=1}^{m} |p_k - q_k|$
  - Example:
    - $p_1 = (0,2)$, $p_2 = (2,0)$
    - Distance (p_1, p_2) = |0-2| +|2-0| = 4
- Euclidean distance
  - r=2
  - Length of a straight line between two data points within m dimensions
  - $Ed\ (p,q) = \sqrt{\sum_{k=1}^{m}(p_k - q_k)^2}$
  - Normalisation is necessary is attributes (dimensions) differ in levels of measurement.
  - Example:
    - $p_1 = (0,2)$, $p_2 = (2,0)$
    - Distance (p_1, p_2) = $\sqrt{(0-2)^2 + (2-0)^2} = \sqrt{8} = 2.828$
- Supremum distance/L_{max}, L_∞ distance
  - r → ∞
  - Maximum difference in any component of the vector.
  - Calculate all difference of the vector in each dimension the supremum distance is the maximum value.
  - $dist\ (p,q) = \max|p_k - q_k|$
  - Example:
    - $p_1 = (0,2)$, $p_2 = (2,0)$
    - Distance (p_1, p_2) = max (|0-2|, |2-0|) = 2

Visualisation of different Minkowski distances:

Red = Manhattan distance
Blue = Euclidean Distance
Green = Supremum distance

All Figures are from Prof. Li BMEG 3105 Lecture 7 Notes