Scribe: Koo Hoi Ying Kody

BMEG3105: Data analytics for personalized genomics and precision medicine

Lecture 7: Clustering

Friday, 27 September 2024

Why clustering?

- Cluster items:
  - Better organization
  - Faster searching (identify the wanted item easily by going into a category)
- Cluster people:
  - Different treatment for different groups
    - e.g. symptoms, age
  - Different groups with different groups
    - Optimize the product based on needs of the targeting group
- Cluster genes:
  - Identify co-expressed genes
    - Involved in the same pathway (biological process or similar function)
  - Identify differentially expressed genes
    - Related to diseases
- Cluster samples/ cells:
  - Identify new diseases sub-types
  - Identify new cell types

What is clustering analysis?

Finding a groups of objects such that

1. Intra-cluster differences are small (objects within a group is similar / related)
2. Inter-cluster differences are large (objects in different groups are different / unrelated)

Clustering analysis

- Give insights in data distribution
- Pre-processing step for other algorithms
- Discover new groups

⇨ Clustering analysis can reduce the size of large data sets

Clustering methods

➔ The magic tool to give clustering indicator (ID)

What we need to do clustering?

- The data
- Measurements to determine the grouping ways (Similarity measurement)
- Clustering algorithm (the executive procedure)

(Example) Which is not a clustering analysis?

A. Grouping genes based on gene expression
B. Aligning 2 sequences
C. Dividing the customers based on their age
D. Forming interest society

Ans: B (Aligning sequences is actually a step of clustering)

Similarity and dissimilarity

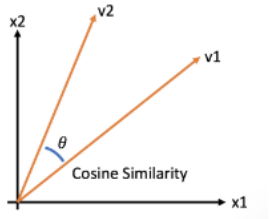
Similarity:

- Numerical measure of how alike two data objects are
- More alike, Higher similarity
- Often falls in the range [0,1]

Dissimilarity (distance):

- Numerical measure of how different two data objects are
- More alike, Lower dissimilarity
- Min is often 0
- Upper limits varies

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**

Correlation

- Measure the linear relationship between objects

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

Euclidean distance

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

➢ **Where m is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects $\boldsymbol{p}$ and $\boldsymbol{q}$.**

- Normalization is necessary if scales of different dimension differ
- Data matrix & Distance matrix
  - Date matrix: show the points of the data
  - Distance matrix: Each row and column are representing distance between two points (Euclidean distance)

Minkowski distance

- Generalization of Euclidean Distance

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

❖ **Where r is a parameter, m is the number of dimensions (attributes) and p_k and q_k are, respectively, the k-th attributes (components) or data objects $\boldsymbol{p}$ and $\boldsymbol{q}$.**