BMEG 3105 Fall 2022

Data analytics for personalized genomics and precision medicine

Course introduction

Lecturer: Yu LI (李煜) from CSE

Liyu95.com, liyu@cse.cuhk.edu.hk

Monday, 5 September2022

Outline of lecture

I. Recap from last lecture

A. Data cleaning
B. Data exploration
C. Clustering
D. 0 VS missing value

II. Taday's agenda

A. Similarity and dissimilarity Measurement
   a) Identity
   b) Measurement methods
      1. Cosine similarity If and are two vectors, then
      2. Correlation
      3. Euclidean distance
      4. Minkowski distance
      5. Mahalanobis distance (Advanced)

B. Hierarchical clustering
   a) Identity
   b) Steps

III. Self-study material

A. Data mining
B. programming

I. Recap from last lecture

A. Data cleaning

a) Denoise data (if applicable)
b) Remove outliers
c) Handling missing data
d) Remove duplicates
e) Categorical data encoding
f) Data normalization

B. Data exploration

a) Summary statistics:
    - Location: mean and median;
    - Spread: range, variance, percentiles;
    - Frequency: mode
b) Visualization (distribution and trend)
    - Histogram
    - Box plots

C. Clustering

a) Datato be clustered
b) Similarity measurement
c) Clustering algorithm (the executive procedure)

D. 0 VS Missing value

- 0 is informative
- Missing value means we do not have the information

II. Today's agenda

A. Similarity and dissimilarity measurement

a) Identity

Similarity

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

Dissimilarity (distance)

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

b) Measurement Methods

1. Cosine similarity
   If d1 and d2 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

2. Correlation
   Correlation measures the linear relationship between objects
   $$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

3. 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 p and q.
   (Normalization is necessary, if scales of different dimension differ)

4. Minkowski distance
   Minkowski Distance is a generalization of Euclidean Distance
   $$dist(\boldsymbol{p},\boldsymbol{q}) = \left(\sum_{k=1}^{m} |p_k - q_k|^r\right)^{\frac{1}{r}}$$
   Where r is a parameter, m is the number of dimensions (attributes), and p_k and q_k are, respectively, the k-th attributes (components) or data objects p and q.

There are 3 cases:

- Case 1: r =1. City block (Manhattan, taxicab, L_1 norm) distance. A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors.

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

- Case 2: r =2. Euclidean distance
- Case3: r →∞. "supremum" (L_{max} norm, L_∞ norm) distance. This is the maximum difference between any component of the vectors

$$dist(p,q) = (\sum_{k=1}^{m} a_k^{\infty})^{\frac{1}{\infty}}$$

Which means the maximum absolute value among all the distances for each pair.

5. Mahalanobis distance (Advanced)
Calculating distance considering the data distribution

❖ **Mahalanobis distance**

$$mahalanobis(\boldsymbol{p},\boldsymbol{q}) = (\boldsymbol{p}-\boldsymbol{q})^T \Sigma^{-1} (\boldsymbol{p}-\boldsymbol{q})$$

❖ **Where $\sum$ is the covariance matrix**

How to calculate the inverse of the covariance matrix?

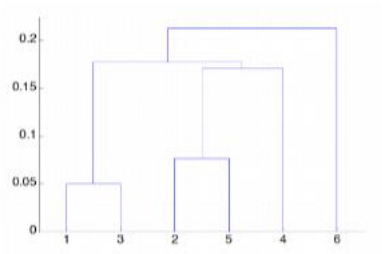
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\underbrace{ad-bc}_{\text{determinant}}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

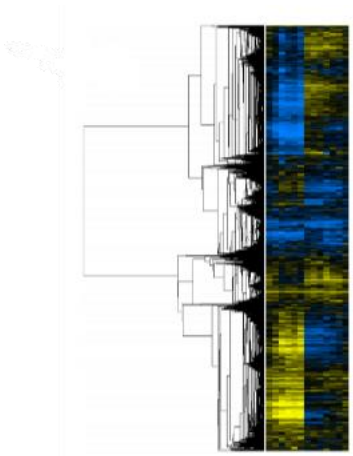
$$\begin{bmatrix} 4 & 7 \\ 2 & 6 \end{bmatrix}^{-1} = \frac{1}{4\times6-7\times2} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 6 & -7 \\ -2 & 4 \end{bmatrix} = \begin{bmatrix} 0.6 & -0.7 \\ -0.2 & 0.4 \end{bmatrix}$$

B. Hierarchical clustering

a) Identity

- Producesa set of nested clusters organized as a hierarchical tree
- Can be visualized as adendrogram. A tree like diagram that records the sequences of merges
- They may correspond tomeaningful taxonomies. For example, Gene clusters, phylogeny reconstruction, animal kingdom...





b) Steps

Step1: Compute the Similarity or Distance matrix. (Use the methods mentioned above)

Step2: Let each data point be a cluster.

Step3: Merge the two closest clusters. (e.g. The clusters with the largest correlation)

Step4: Update the similarity or distance matrix. (e.g. Replaced with a larger correlation)

Step5: Repeat Step3&4 until only a single cluster remains.

Example:

1) Computer the similarity by correlation.

Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

2) Now we get the distance matrix. Find the largest correlation, which means the shortest distance. Here we get gene2&4. Then combine gene2&4.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733			
At2g34930	0.9493	0.9909	-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528	

3) After merging gene2&4, we need to update the distance matrix. We update all values related to gene2&4 by a larger correlation, which means a shorter distance as well.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733 ->-0.9493			
At2g34930	0.9493 ->0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528 ->0.562	

4) After updating the matrix, find the largest correlation again. Here we find "0.9733" so we choose gene 3 to combine with gene2&4.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9493			
At2g34930	0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.562	

5) After Merging gene 1, we update the values with the larger correlation again.

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-1 ->-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.562 ->0.5774	-0.5774	0.562 ->0.5774	

6) Repeat steps, until only one single cluster (gene 3) remains.

Node2
Node1
Node3

	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-0.5774	-0.5774			
At2g34930			-0.5774		
At2g05540			-0.5774		

III. Self-learning Materials

A. Data mining: (Introduction to data mining: Chapter 2.4& Chapter 8)

- K-means clustering
- Density-based clustering
- How to determine the number of clusters
- How good is your clustering (lecture 8-9)

B. Programming: (Scikit-learn)

https://scikit-learn.org/stable/ClusteringYu LiLecture6-39

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py