BMEG3105: Data analytics for personalized genomics and precision medicine - Fall 2023

Lecture 7 - Clustering

Lecturer: Professor Yu LI Scriber: Chiu Chi Chung (SID: 1155174790)

1. Introduction

Lecture 6 mainly focused on clustering, including:

a. reasons and examples on using clustering;
b. definition and methods of clustering;
c. hierarchical clustering.

2. Reasons and examples

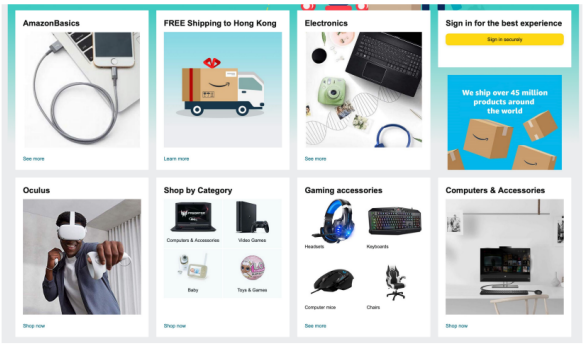
Clustering is usually used for understanding data as useful information and summarization. Sometimes, it is a stand-alone tool to get insights into data distribution, or a pre-processing step for other algorithms. Some examples are

1. grouping related documents for browsing;
2. grouping genes and proteins that have similar functionality;
3. or grouping stocks with similar price fluctuations.

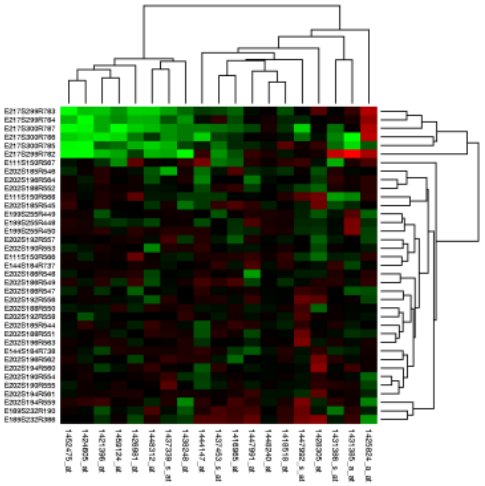
For summarization, it could be used for reducing the size of large data sets or preserving privacy, especially for medical data.

Here are some more examples that can be clustered:

1. items: for better organization and faster searching.



2. people: for understanding different needs of different groups of people, like patients or customers. Age, gender, and income are some common attributes used in this area.
3. genes: for identifying co-expressed and differentially expressed genes that are related to diseases.

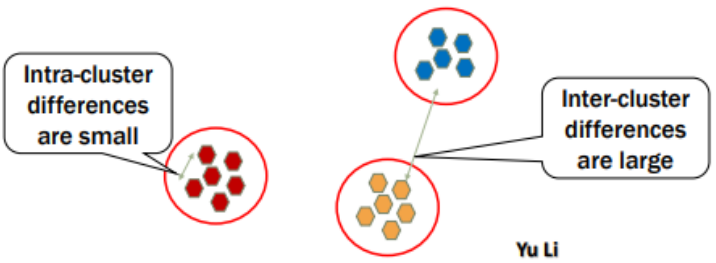


4. samples/cells: for identifying new disease subtypes, new cell types, etc..

3. Definition and methods

Clustering is finding groups of objects such that they have small intra-cluster differences and big inter-cluster differences. It involves three components to be defined, which are:

1. similarity and dissimilarity,
2. measurements of differences,
3. algorithms.



Difference could be measured in two ways, similarity and dissimilarity. **Similarity** is usually the numerical measure of how alike two data objects are. The higher it is, the more similar objects are, but it usually bounded between 0 and 1. Meanwhile, **dissimilarity** is the numerical measure of how different are two data objects. Opposingly, the lower it is, the more similar objects are. It usually doesn't have an upper limit, but its minimum is often 0.

In the following, different measurement methods will be introduced, while one of the algorithms will be introduced in the next part.

Cosine similarity:

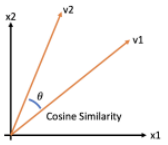
It is a mathematical calculation that is usually used with vector data.

❖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

❖Example:



d_1 = 3 2 0 5 0 0 0 2 0 0

d_2 = 1 0 0 0 0 0 0 1 0 2

$d_1 \cdot d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

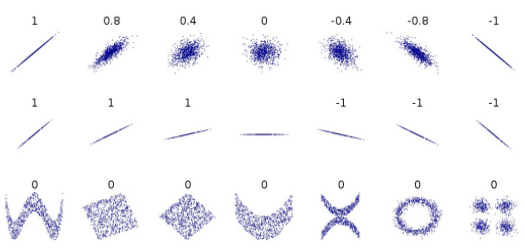
$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

cos(d_1, d_2) = 0.3150

Correlation:

Correlation measures the linear relationship between objects, and we usually combine all combinations of correlation of a group of data into a correlation matrix/ heatmap.

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



Euclidean distance:

It measures the linear distance between two points in a coordinate system, with the following equation. However, it requires normalization before calculation.

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

➢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 $\boldsymbol{q}$.

Minkowski distance:

Minkowski Distance is a generalization of Euclidean Distance, with an extra parameter, r. When r is 2, it becomes the Euclidean Distance, but when r goes to infinity, the answer will be bound by the maximum absolute difference among pairs of elements in two vectors.

❖$r = 1$. 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

❖$r = 2$. Euclidean distance

❖$r \to \infty$. "supremum" ($L_{\max}$ norm, L_∞ norm) distance.
➢This is the maximum difference between any component of the vectors

It will be the Max of dist(P,Q) a set

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

It denotes the result

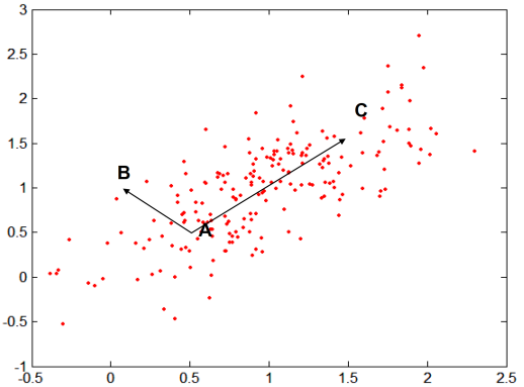
Mahalanobis distance:

The Mahalanobis distance is the distance of the test point from the center of mass divided by the width of the ellipsoid in the direction of the test point.

❖Mahalanobis distance

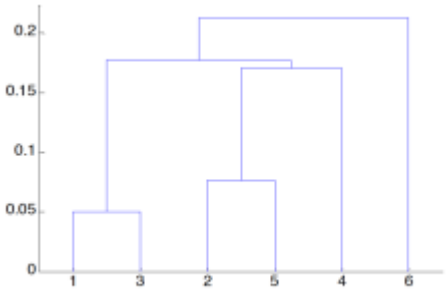
$$mahalanobis(\boldsymbol{p},\boldsymbol{q})=(\boldsymbol{p}-\boldsymbol{q})^T\Sigma^{-1}(\boldsymbol{p}-\boldsymbol{q})$$

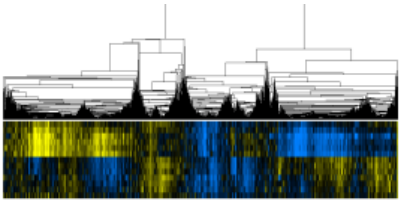
❖Where $\sum$ is the covariance matrix



4. Hierarchical clustering

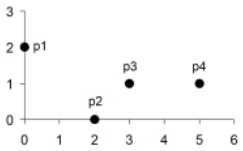
Hierarchical clustering produces a set of nested clusters organized as a hierarchical tree, which can also be visualized as a dendrogram. It has been used in many areas, like gene clusters, phylogeny reconstruction, animal kingdom, etc..





It can be break down into following steps:

1. compute the Similarity or Distance matrix of a set of data;
2. let each data point be a cluster;



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

3. merge the two closest clusters;
4. update the similarity or distance matrix
5. repeat steps 3 and 4 until only one cluster left

The most complicated step would be step 4, which requires us to recalculate the similarity or dissimilarity. We first have to define a means to compute the similarity between a cluster and a point/ a cluster and a cluster. Then, we only need to replace the similarity score of those which are related to the newly formed cluster. The remaining score will remain and be passed on, so we do not need the original data set throughout the process.