BMEG3105 Data analytics for personalized genomics and precision medicine

Topic: Lecture7 Clustering

Scriber: 1155192506 Xu Ching Laam

Lecture Outcome:

1. Reasons of clustering
2. Description of clustering
3. Similarity and dissimilarity of measurement
4. Hierarchical clustering

1. Reasons of Clustering

- **Clustering items**
  - Better organization
  - Faster searching

- **Clustering people**
  - Optimize the product based on the need of the targeting group
  - i.e. Different treatment/products for different groups of users
    e.g. age, gender, needs

- **Clustering genes**
  - Identify co-expressed genes (with same pathway)
  - Identify differentially expressed genes (related to diseases)

- **Clustering samples/cells**
  - Identify new disease sub-types
  - Identify new cell types

2. Description of Clustering

Finding groups of objects such that the objects in group

-will be similar to one another (small intra-cluster distances)

-and different from the objects in other groups (large inter-cluster distance)

Usage of clustering:

- **Understanding**
  - As a stand-alone tool to get insight into data distribution
  - As a pre-processing step for another algorithm

- **Summarization**
  - Reduce size of large data sets
  - Preserve privacy

How to do clustering:

1. Collect data to be clustered
2. Similarity measurement
3. Clustering algorithm (the executive procedure)

3. Similarity and Dissimilarity of Measurement

- **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 two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies (may not have upper limit)

How to measure the similarity:

- **Cosine similarity (i.e. find the angle between two vectors)**

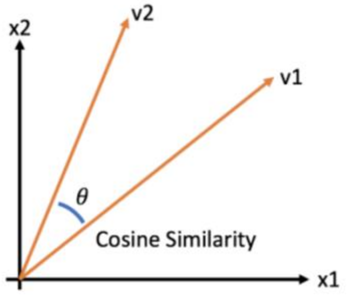
  2 vectors: d_1 and d_2

  Indicates the dot product and |d| (length of vector) and sub into the formula below

  Then, the angle between two vectors can be found (which indicates the similarity)

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

  Example:

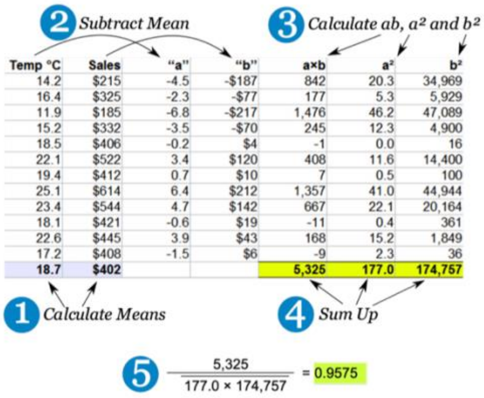


- **Correlation (i.e. find the linear relationship between objects)**

  General formula:

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

  Example:



2 *Subtract Mean*

3 *Calculate ab, a² and b²*

Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

1 *Calculate Means*

4 *Sum Up*

5 $\frac{5{,}325}{\overline{177.0 \times 174{,}757}} = 0.9575$

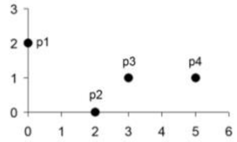
- **Euclidean distance (i.e. find the straight-line distance between two test points)**

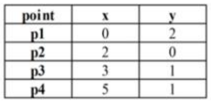
$$Ed(\boldsymbol{p},\boldsymbol{q})=\sqrt{\sum_{k=1}^{m}(p_k-q_k)^2}$$

- m is the number of dimensions
- p_k = k-th attributes or data objects p
- q_k = k-th attributes or data objects q

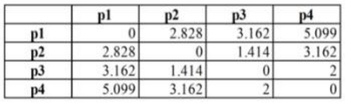
*Normalization is needed if scales of different dimensions differ

Example:





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

- **Minkowski distance (i.e. generalization of Euclidean distance)**

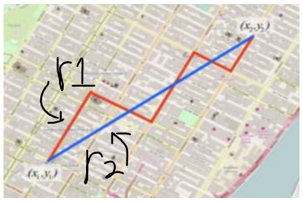
$$dist(\boldsymbol{p},\boldsymbol{q})=\left(\sum_{k=1}^{m}|p_k-q_k|^r\right)^{\frac{1}{r}}$$

- m is the number of dimensions
- p_k = k-th attributes or data objects p
- q_k = k-th attributes or data objects
- r = a parameter:

  r1 = City block (Manhattan, taxicab, L_1 norm) distance

  r2 = Euclidean distance

  r3 ---> infinity (supremum distance) (maximum distance difference between and component of vectors)



- **Mahalanobis distance (i.e. distance considering data distribution)**

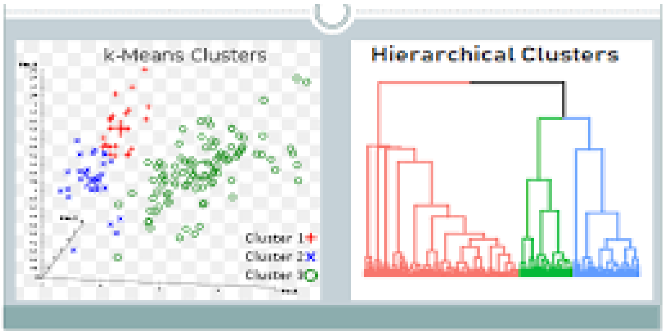
$$(\boldsymbol{p},\boldsymbol{q})=(\boldsymbol{p}-\boldsymbol{q})^T\Sigma^{-1}(\boldsymbol{p}-\boldsymbol{q})$$

4. Hierarchical Clustering

What is hierarchical clustering?

- Produce a set of nested clusters organized as a hierarchical tree
- Can be organized as a dendrogram
  i.e. a tree diagram that records the sequences of merges
- May correspond to meaningful taxonomies
  e.g. gene clusters, phylogeny reconstruction, animal kingdom...

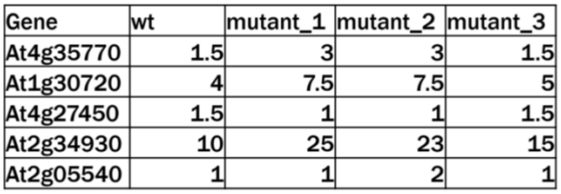
Example of Hierarchical clusters:



Steps of hierarchical clustering:

1. **Compute the Similarity or Distance matrix**
2. **Let each data point be a cluster**
3. **Merge the two closest clusters**
4. **Update the similarity or distance matrix (first time)**
    - Methods of updating the distance matrix after merging:
      - Minimum
      - Maximum
      - Group average
      - Distance between centroids
5. **Continue the previous two steps...**
6. **Until only a single cluster remains**

Example: Given the data matrix below (after normalization)



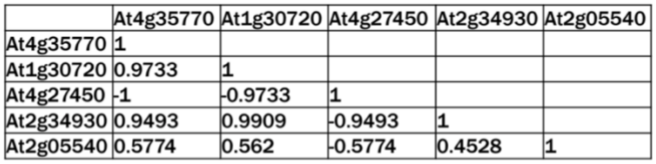
Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

Visualization after normalization

1. Compute distance matrix with linear correlation

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

2. Each gene be a cluster.

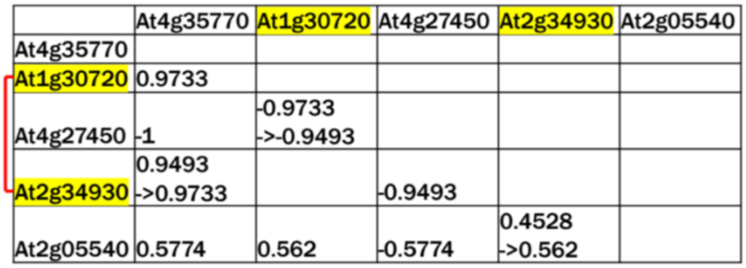


	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770	1				
At1g30720	0.9733	1			
At4g27450	-1	-0.9733	1		
At2g34930	0.9493	0.9909	-0.9493	1	
At2g05540	0.5774	0.562	-0.5774	0.4528	1

3. Find two closest clusters and merge them. (and Remove 1)

i.e. Merge At2g34930 and At1g30720

Then update the data with minimum distance

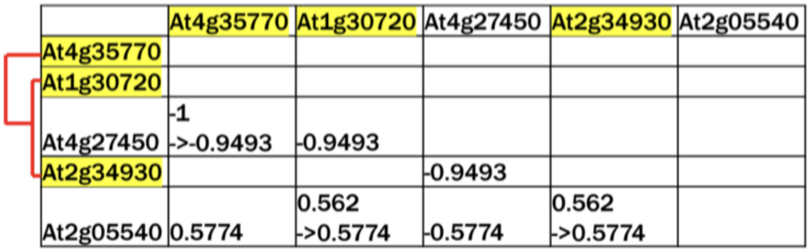


	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720	0.9733				
At4g27450	-1	-0.9733 ->-0.9493			
At2g34930	0.9493 ->0.9733		-0.9493		
At2g05540	0.5774	0.562	-0.5774	0.4528 ->0.562	

4. Find two closest clusters and merge them.

i.e. Merge At2g34930 , At1g30720 and At4g35770

Then update the data with minimum distance (largest correlation)

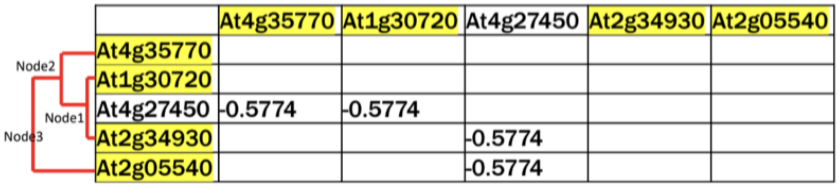


	At4g35770	At1g30720	At4g27450	At2g34930	At2g05540
At4g35770					
At1g30720					
At4g27450	-1 ->-0.9493	-0.9493			
At2g34930			-0.9493		
At2g05540	0.5774	0.562 ->0.5774	-0.5774	0.562 ->0.5774	

5. Find two closest clusters and merge them.

i.e. Merge At2g34930, At1g30720, At4g35770 and At4g35770

Then update the data with minimum distance (largest correlation)


Node2
Node1
Node3

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

6. Finish

This is the end of scribing of clustering.