Lec 8 : Clustering & Classification

Clustering

what are the components needed ?

1) data
2) similarity measurement
3) clustering algorithm

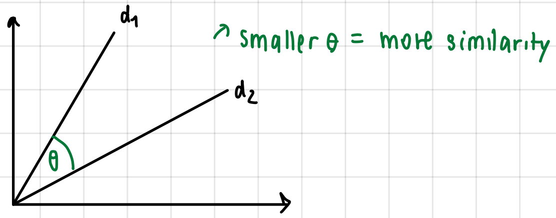
How to measure similarity

① Cosine similarity

d_1 = vector 1 d_2 = vector 2

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

↗ dot product



cosine

② correlation

linear relationship

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

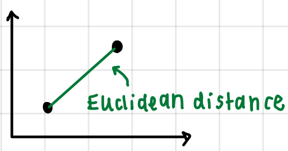
Correlation and covariance both indicate the relationship between two variables

covariance can only indicate the direction (proportional / inproportional), but the correlation can indicate both direction and strength (can tell the dependency degree of relationship)

③ Euclidean distance

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

k represents dimensions



Euclidean

④ Minkowski distance

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

generalization of Euclidean (but with gimmick ✧)

- If r = 1

  the summation of all the paths (in other names: Manhattan)

  or it can be number of bits between two binary vectors.

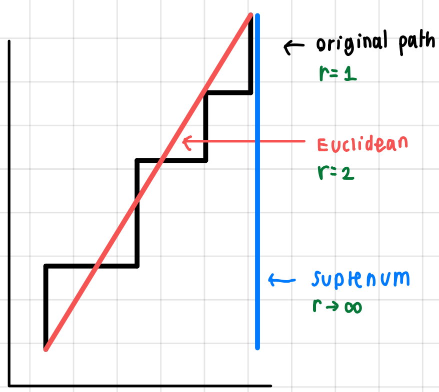
- If r = 2

  Euclidean distance ! (Easy)

- If r = ∞

  Suprenum (L_∞ norm) distance

$$\lim_{r \to \infty} \sum_{k=1}^{m} (|p_k - r_k|^r)^{\frac{1}{r}} = \max_k [p_k - r_k]$$



Minkowski

Hierarchal Clustering

↳ just a nested cluster

What are the steps ?

1) compute similarity distance
2) merge
3) update
4) repeat 1)-3) until the data is all clustered

How ?

Gene	wt	mutant_1	mutant_2	mutant_3	μ
At4g35770	1.5	3	3	1.5	2.25
At1g30720	4	7.5	7.5	5	6
At4g27450	1.5	1	1	1.5	1.25
At2g34930	10	25	23	15	28.25
At2g05540	1	1	2	1	1.25

① choose similarity distance

- Euclidean
- Minkowski
- cosine
- correlation

$$corr(X,Y) = \frac{E[(X-\mu_X)(Y-y_X)]}{\sigma_X \sigma_Y}$$

② calculate similarity distance

Find correlation between gene types At1g30720 and At4g35770

$\mu_X = 2.25$

$\mu_Y = 6$

$\sigma_X = 0.75$

$\sigma_Y = 1.541$

$$cov(X,Y) = \frac{(1.5-2.25)(4-6)+(3-2.25)(7.5-6)+(3-2.25)(7.5-6)+(1.5-2.25)(5-6)}{4}$$

$$= 1.125$$

$$corr(X,Y) = \frac{1.125}{0.75 \cdot 1.541} = 0.9734$$

	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 choose the best pair and merge

→ choose 0.9909 of At1g30720 , At2g34930

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

④ choose scoring matrix

- choose max (largest correlation)

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

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

At4g27450

Mahalanobis Distance

- one way for similarity distance
- consider data distribution (variance / SD)

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

Σ^{-1}: inverse of covariance matrix

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

$$\Sigma^{-1} = \frac{1}{(0.3 \cdot 0.3) - (0.2 \cdot 0.2)} \begin{bmatrix} 0.3 & -0.2 \\ -0.2 & 0.3 \end{bmatrix}$$

$$= \frac{1}{0.05} \begin{bmatrix} 0.3 & -0.2 \\ -0.2 & 0.3 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix}$$

A = [0.5, 0.5]
B = [0, 1]
C = [1.5, 1.5]

$$\text{mahal}(AB) = \begin{bmatrix} 0.5 & -0.5 \end{bmatrix} \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & -5 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$$

$$= 2.5 + 2.5 = 5 \ \#$$

$$\text{mahal}(AC) = \begin{bmatrix} -1 & -1 \end{bmatrix} \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

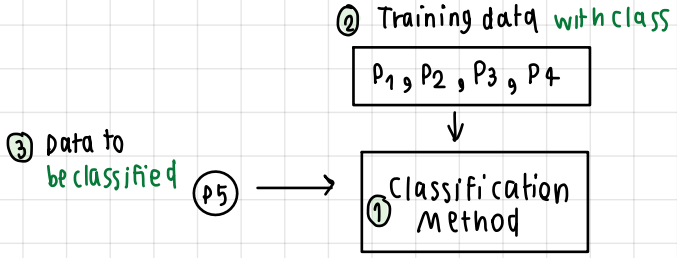
$$= \begin{bmatrix} -2 & -2 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$= 2 + 2 = 4 \ \#$$

Classification

- method to assign class of previously unseen records based on the attributes & training set

Components needed



① KNN (K-Nearest Neighbour)

- store all available instances & classify based on distance metric

Training : store the data (can be regard as no training period)

Predicting :
- most frequent class appeared
- if they are equal, the closest one

* Data should be normalized !

How?

Person	Height(m)	Weight(kg)	Gender
P1	1.79	75	M
P2	1.64	54	F
P3	1.70	63	M
P4	1.88	78	M
P5	1.75	70	??

① normalization

this one we choose min-max normalization

(P1) Height: $\frac{1.79 - 1.64}{1.88 - 1.64} = 0.625$

weight: $\frac{75 - 54}{78 - 54} = 0.875$

as a result

Person	Height	Weight	Gender
P1	0.625	0.875	M
P2	0	0	F
P3	0.25	0.375	M
P4	1	1	M
P5	0.4583	0.6667	??

② Distance calculation of P5 with other

P1 and P5

Euclidean distance $= \sqrt{(0.4583 - 0.625)^2 + (0.875 - 0.6667)^2}$

$= 0.2668$

as a result

Person	P5	Gender
P1	0.267	M
P2	0.809	F
P3	0.358	M
P4	0.636	M
P5	0	??

③ Identify the K most similar data

K = 2

P1 0.267 and P3 0.358

both classify as M

④ conclusion

M