Continuation from last time:

Stat 571
12-5-13

$\frac{\vec{a}' B \vec{a}}{\vec{a}' \Sigma \vec{a}}$ is to be maximized

Let $\vec{x} = \Sigma^{1/2} \vec{a}$ $\quad (\vec{a} = \Sigma^{-1/2} \vec{x})$

$$\frac{\vec{x}' \Sigma^{-1/2} B \Sigma^{-1/2} \vec{x}}{\vec{x}' \Sigma^{-1/2} \Sigma \Sigma^{-1/2} \vec{x}} = \frac{\vec{x}' (\Sigma^{-1/2} B \Sigma^{-1/2}) \vec{x}}{\vec{x}' \vec{x}}$$

This is max. when $\vec{x} = \vec{e}_1$, the 1st eigenvector of $\Sigma^{-1/2} B \Sigma^{-1/2}$

So $\vec{a}_1 = \Sigma^{-1/2} \vec{x} = \Sigma^{-1/2} \vec{e}_1$

$$\Sigma^{-1/2} B \Sigma^{-1/2} \vec{e}_1 = \lambda_1 \vec{e}_1$$

$$\Sigma^{-1/2} B \vec{a}_1 = \lambda_1 \vec{e}_1$$
$$= \lambda_1 \Sigma^{1/2} \vec{a}_1$$

$$\Sigma^{-1} B \vec{a}_1 = \lambda_1 \vec{a}_1$$

So $\vec{a}_1$ is the 1st eigenvector of $\Sigma^{-1} B$.

Cluster analysis

(1) Distance measures

Euclidean distance $d(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|$

$$= \sqrt{(x_1 - y_1)^2 + \dots + (x_p - y_p)^2}$$

Mahalanobis distance

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})' A (\vec{x} - \vec{y})}$$

↑ for example, S^{-1}

Minkowski distance

$$d(\vec{x}, \vec{y}) = \sqrt[m]{\sum_{i=1}^{p} |x_i - y_i|^m}$$

(2) Similarity measures

Example

Variable	1	2	3	4	5
Item X	1	0	0	1	1
Item Y	1	1	0	1	0

Item X \ Item Y	1	0
1	2	1
0	1	1

5

a	b
c	d

p

Possible similarity coefficients. (5)

1) $\frac{a+d}{p}$ 2) $\frac{2(a+d)}{2(a+d)+b+c}$

3) $\frac{a+d}{a+d+2(b+c)}$ 4) $\frac{a}{p}$

Similarity vs. distance

Let $d(\vec{x},\vec{y})$ be a distance

Then $s(\vec{x},\vec{y}) = \frac{1}{1+d(\vec{x},\vec{y})}$ is a similarity

(6)

Let $s(\vec{x},\vec{y})$ be a similarity measure

Then $d(\vec{x},\vec{y}) = \sqrt{1-s(\vec{x},\vec{y})}$

is a distance

Hierarchical Clustering

1) Agglomerative: each item starts in its own cluster

2) Divisive: start with all items in a single cluster

more popular

Agglomerative methods

How to measure distance from one cluster to another?

a) Single linkage (nearest neighbor)

b) Complete linkage (farthest neighbor)

c) Average linkage (average of all distances)

d) Centroid linkage (distance between the centers of mass)

General algorithm:

Step 1: Start with n items, each in its own cluster, and an nxn matrix of distances or similarities

Step 2: Find the smallest distance & merge those two items into a cluster. Recompute your $(n-1) \times (n-1)$ distance matrix

Repeat until everything is on a single cluster. Draw a tree diagram showing the mergers.

(9)

Chon

	MD	GJ	CW	IW	CH	DI	PD
MD	0						
GJ	2.07	0					
CW	5.81	7.69	0				
IW	3.66	5.49	2.31	0			
CH	1.63	3.45	4.64	3.14	0		
DI	1.68	3.44	4.55	2.37	1.80	0	
PD	.72	2.58	5.52	3.49	1.38	1.84	0

(10)

The algorithm will merge MD & PD at .72

Use nearest neighbor

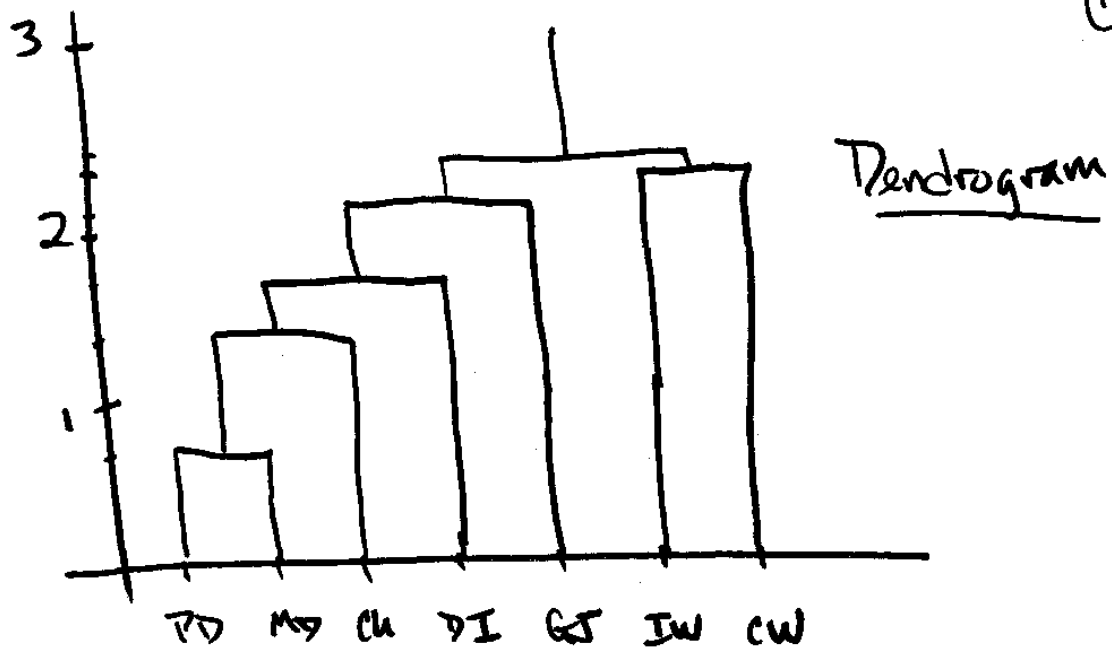
At 1.38, CH joins (MD, PD)

At 1.68, DI joins (MD, PD, CH)

At 2.07, GJ joins (MD, PD, CH, DI)

At 2.31, IW joins CW

At 2.37, you have a single cluster



Dendrogram

Non-hierarchical methods (k-means)

↑

# of klusters

You must know the # of clusters in advance

The algorithm considers every possible partition of the n items into K subgroups, & chooses the one whose center of mass are furthest apart.