STAT406- Methods of Statistical Learning Lecture 23

Matias Salibian-Barrera

UBC - Sep / Dec 2017

Clustering

Dissimilarity measures

- $d(\mathbf{a}, \mathbf{b}) \geq 0$
- $d(\mathbf{a}, \mathbf{b}) = 0$ iff $\mathbf{a} = \mathbf{b}$
- $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$
- $d(\mathbf{a}, \mathbf{b}) \leq d(\mathbf{a}, \mathbf{d}) + d(\mathbf{d}, \mathbf{b})$

Clustering

Dissimilarity measures

- Euclidean distance – L_p distances
  - $d(\mathbf{a}, \mathbf{b}) = \left[\sum_{j=1}^{k} \left| \mathbf{a}_j - \mathbf{b}_j \right|^p \right]^{1/p}$
- L_∞
  - $d(\mathbf{a}, \mathbf{b}) = \max_{1 \leq j \leq k} \left| \mathbf{a}_j - \mathbf{b}_j \right|$

Clustering

When $\mathbf{a}_j \in \{0, 1\}$

- We can use the number of matches / mismatches

	0	1
0	a	b
1	c	d

- $(b + c)/k$ = proportion of mismatches
- $1 - d/k$ = 1 - proportion of 1-1 matches
- Presence is more significant than absence: “person likes Kenneth J. Harvey”

Agglomerative methods

1. Start with n clusters, $C_1, \ldots, C_n$ each with one point
2. Find the pair of closest clusters, C_a, C_b
3. Merge them into $C_{(ab)}$, find $d\left(C_{(ab)}, C_j\right)$ for all other clusters C_j
4. Repeat until all observations belong in one cluster

Agglomerative methods

Different choices for $d\left(C_{(ab)}, C_j\right)$:

- Single linkage
- Complete linkage
- Average linkage
- Ward's "information" criterion

Single linkage

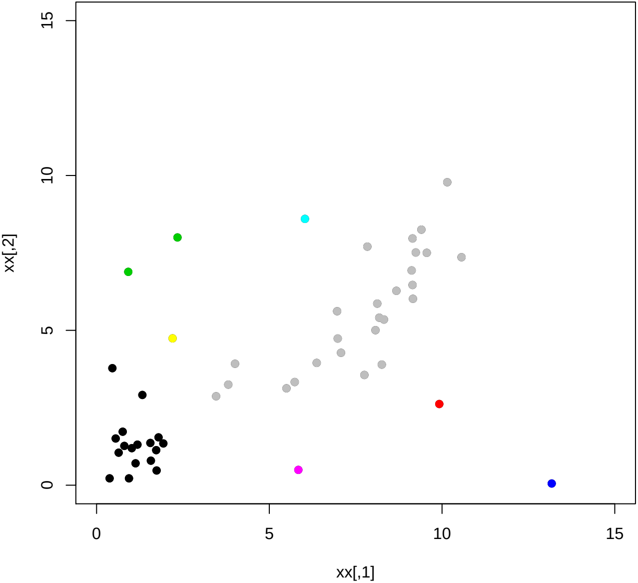
The **distance** between two **clusters** is the **minimum** distance between any **two elements**:

$$\mathcal{C}_1 = \{a_1, \ldots, a_n\} \qquad \mathcal{C}_2 = \{b_1, \ldots, b_m\}$$

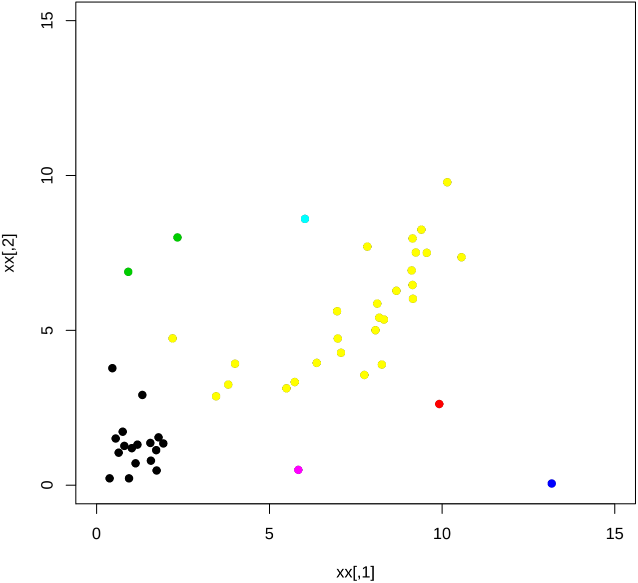
$$d\left(\mathcal{C}_1, \mathcal{C}_2\right) =$$

$$\min \ \{d\left(a_1, b_1\right), d\left(a_1, b_2\right), \ldots, d\left(a_n, b_m\right)\}$$

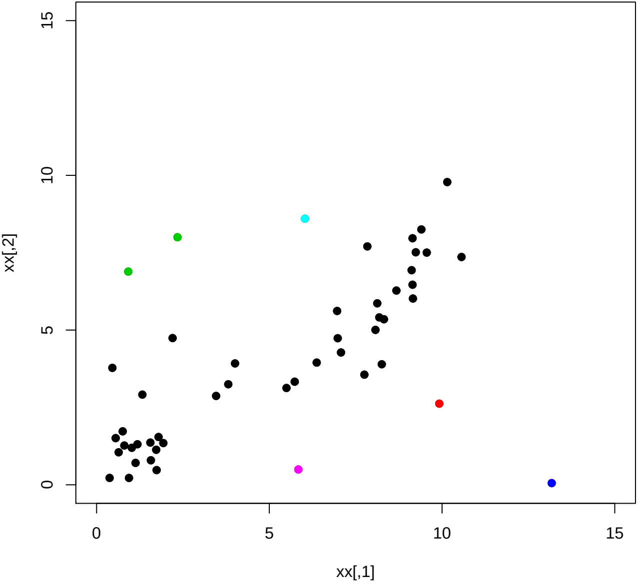
Single linkage



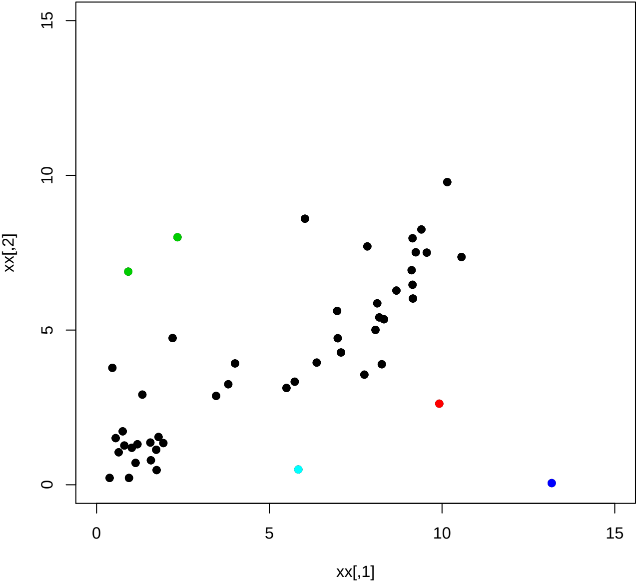
Single linkage



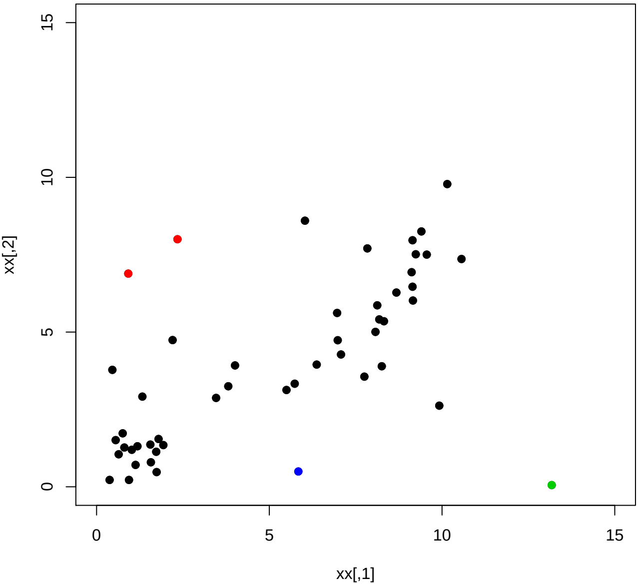
Single linkage



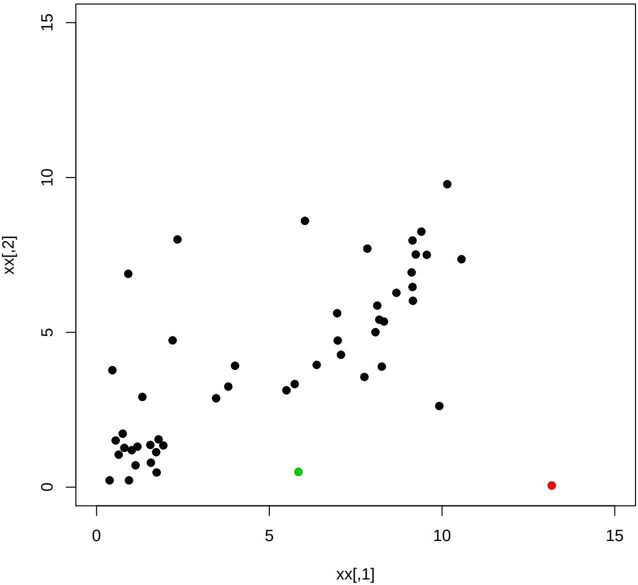
Single linkage



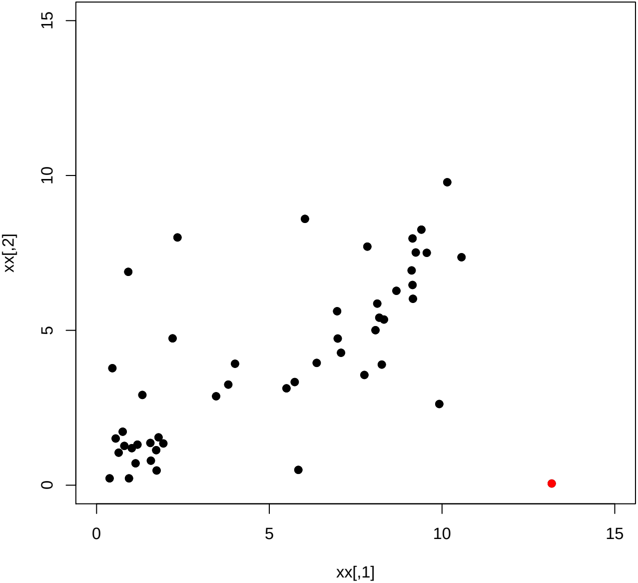
Single linkage



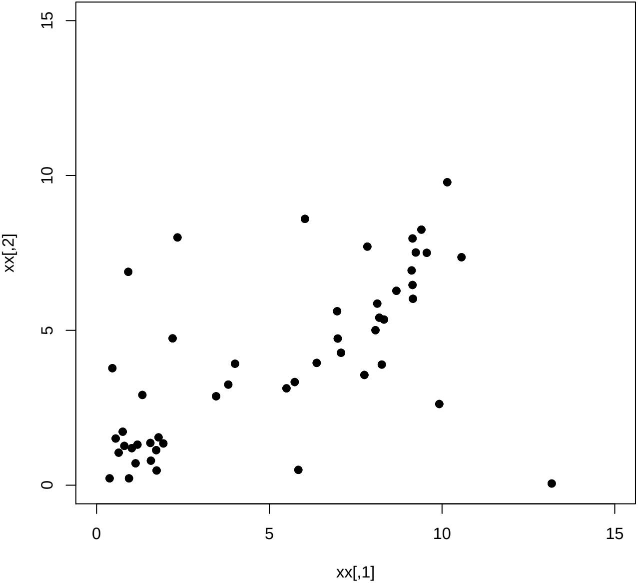
Single linkage



Single linkage



Single linkage



Complete linkage

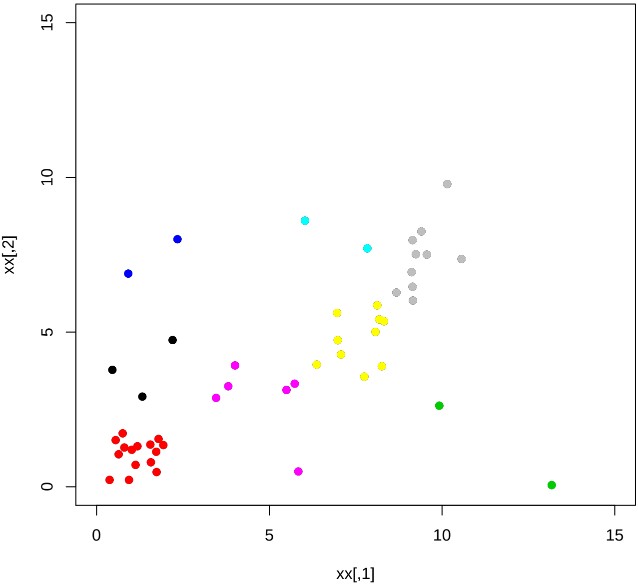
The **distance** between two **clusters** is the **maximum** distance between any **two elements**:

$$\mathcal{C}_1 = \{a_1, \ldots, a_n\} \qquad \mathcal{C}_2 = \{b_1, \ldots, b_m\}$$

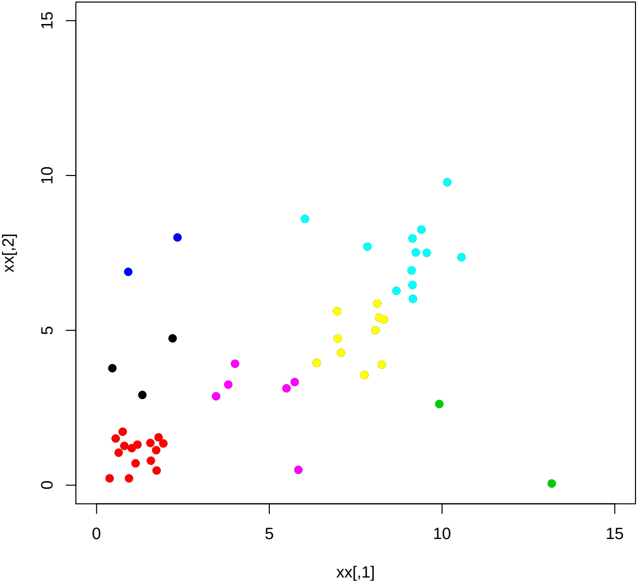
$$d\left(\mathcal{C}_1, \mathcal{C}_2\right) =$$

$$\max \left\{ d\left(a_1, b_1\right), d\left(a_1, b_2\right), \ldots, d\left(a_n, b_m\right) \right\}$$

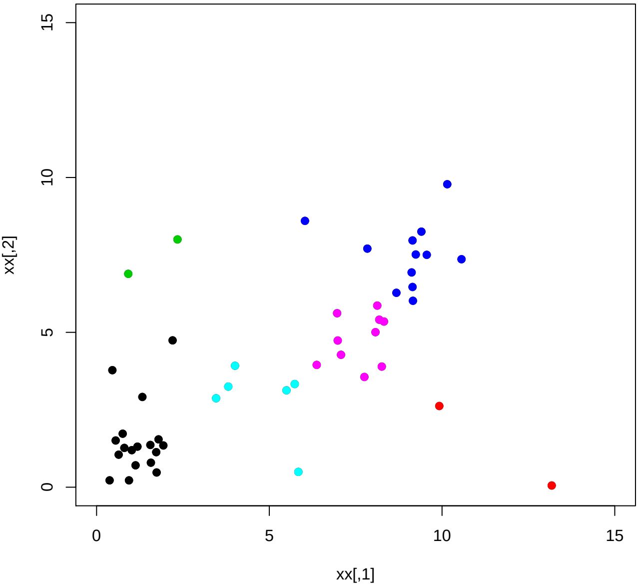
Complete linkage



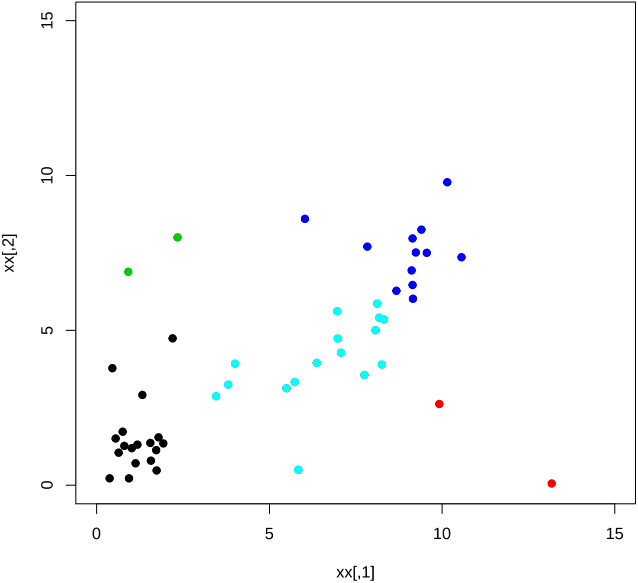
Complete linkage



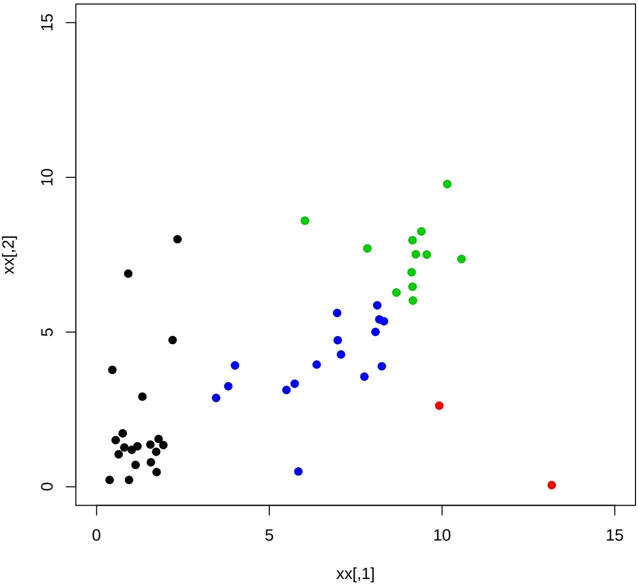
Complete linkage



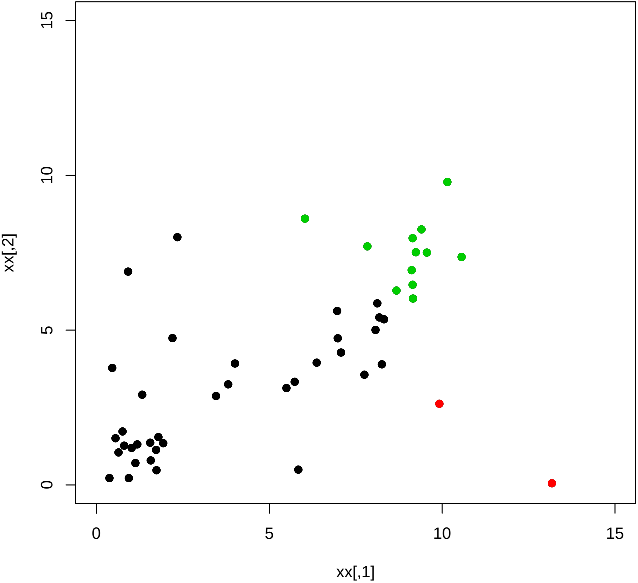
Complete linkage



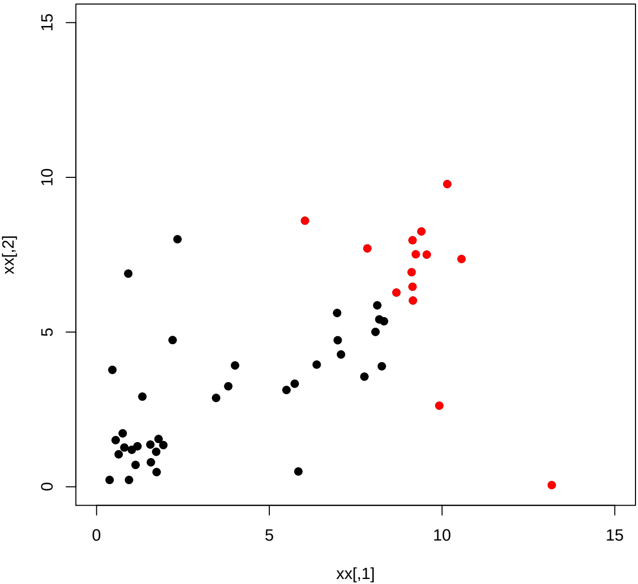
Complete linkage



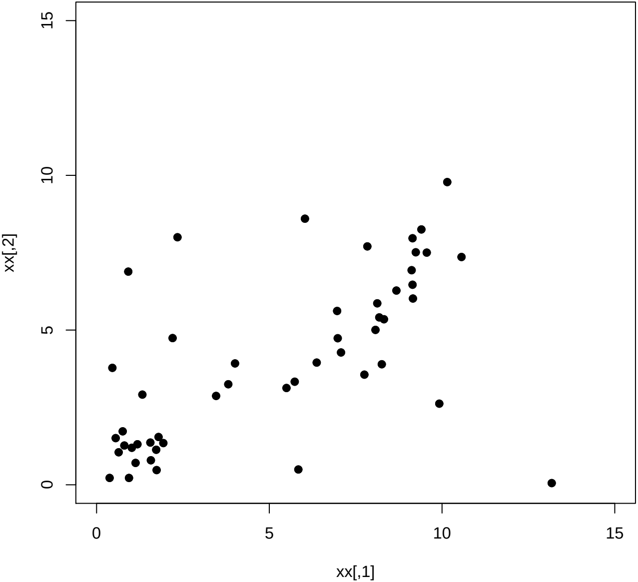
Complete linkage



Complete linkage



Complete linkage



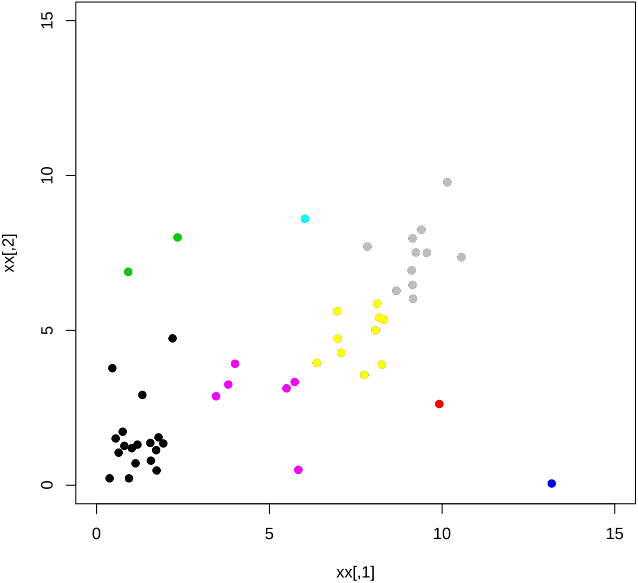
Average linkage

The **distance** between two **clusters** is the **average** of all pairwise distances between any **two elements**:

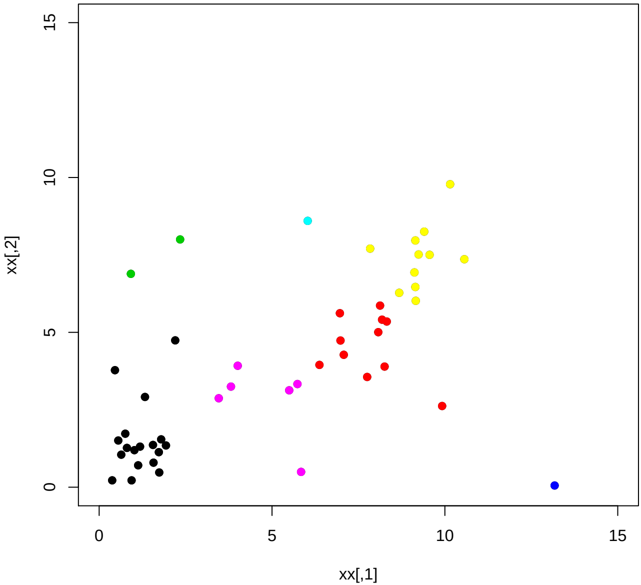
$$\mathcal{C}_1 = \{a_1, \ldots, a_n\} \qquad \mathcal{C}_2 = \{b_1, \ldots, b_m\}$$

$$d\left(\mathcal{C}_1, \mathcal{C}_2\right) = \frac{1}{n\,m} \sum_{i=1}^{n} \sum_{j=1}^{m} d\left(a_i, b_j\right)$$

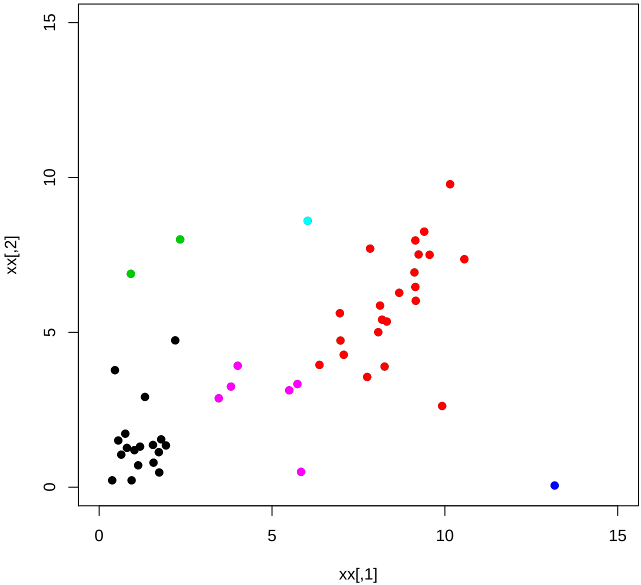
Average linkage



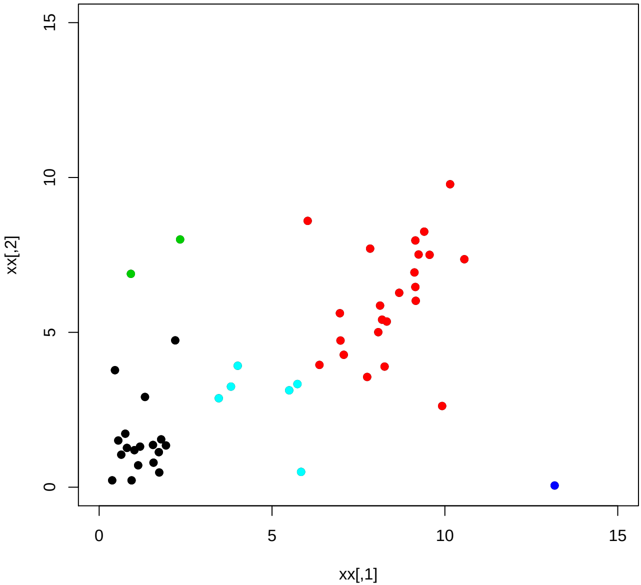
Average linkage



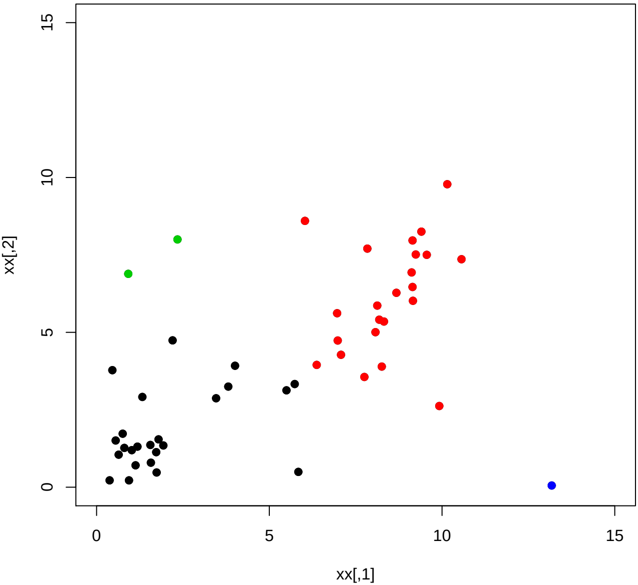
Average linkage



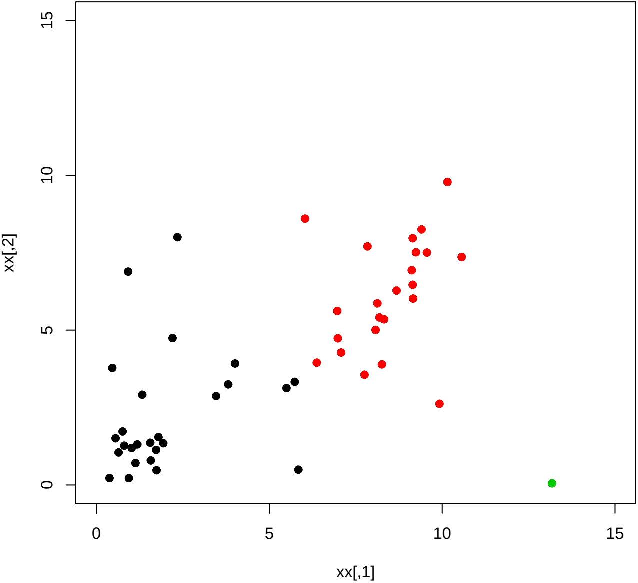
Average linkage



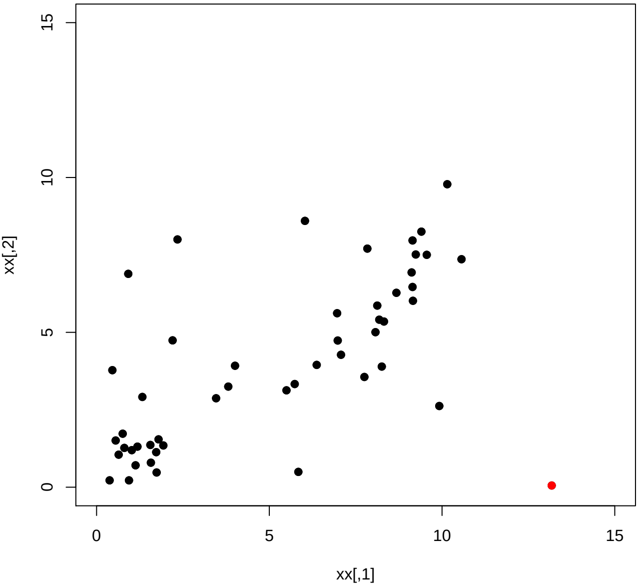
Average linkage



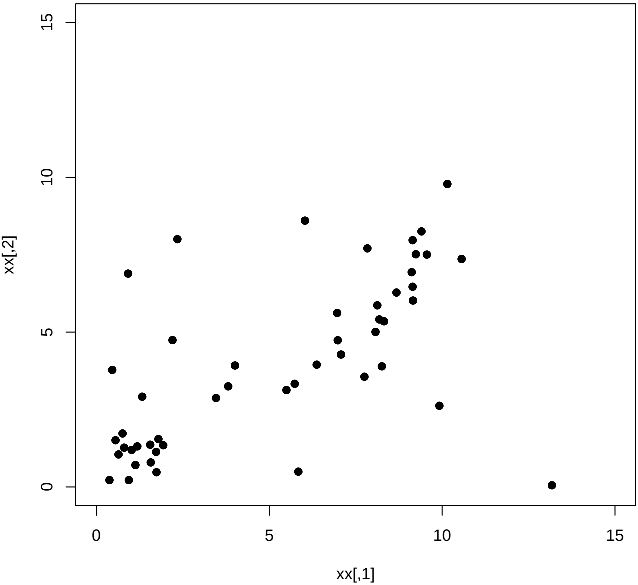
Average linkage



Average linkage



Average linkage



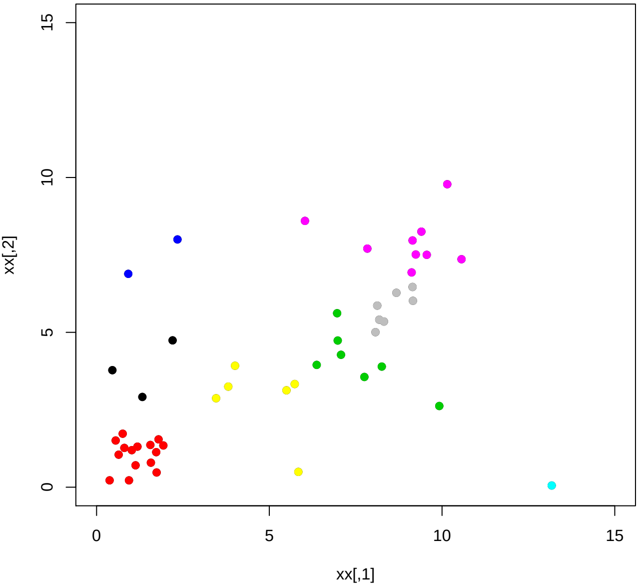
Ward's information criterion

A different merging criterion. Merge those two clusters that would result in the smallest increase in "within cluster sum of squares"

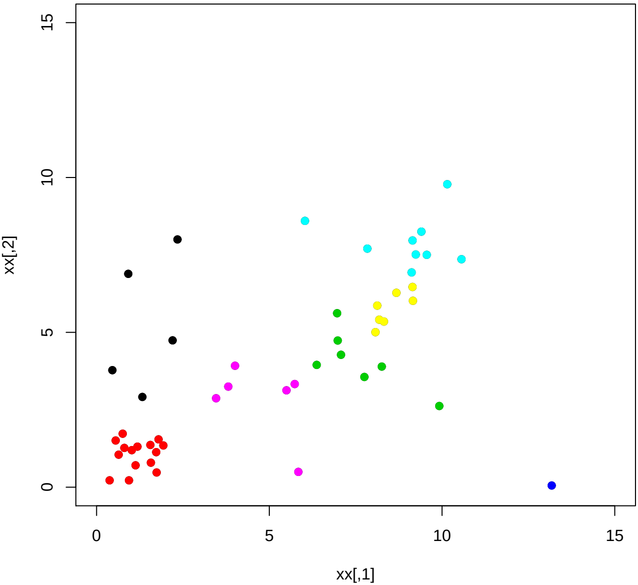
$$\mathrm{SS}\left(\mathcal{C}_r\right) \;=\; \sum_{i \in \mathcal{C}_r} \sum_{j \in \mathcal{C}_r} d^2\left(\mathbf{X}_i, \mathbf{X}_j\right)$$

$$\text{Total SS} \;=\; \sum_{r} \mathrm{SS}\left(\mathcal{C}_r\right)$$

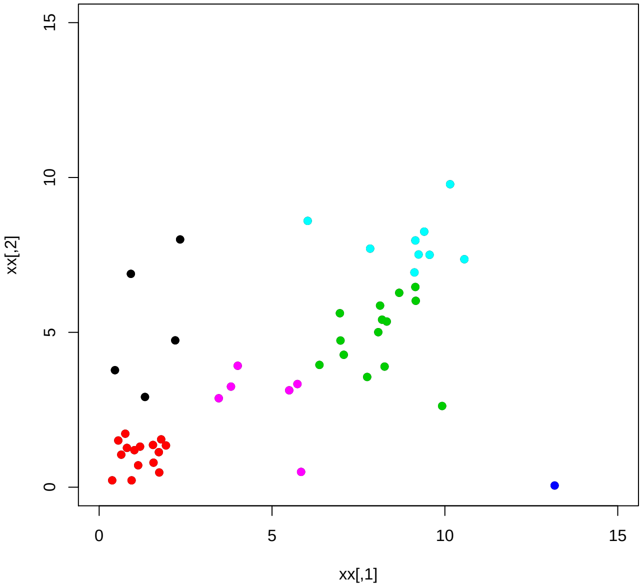
Ward's information criterion



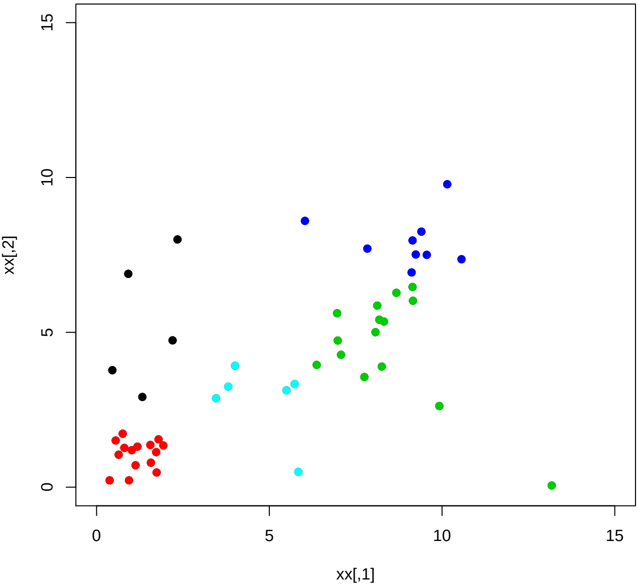
Ward's information criterion



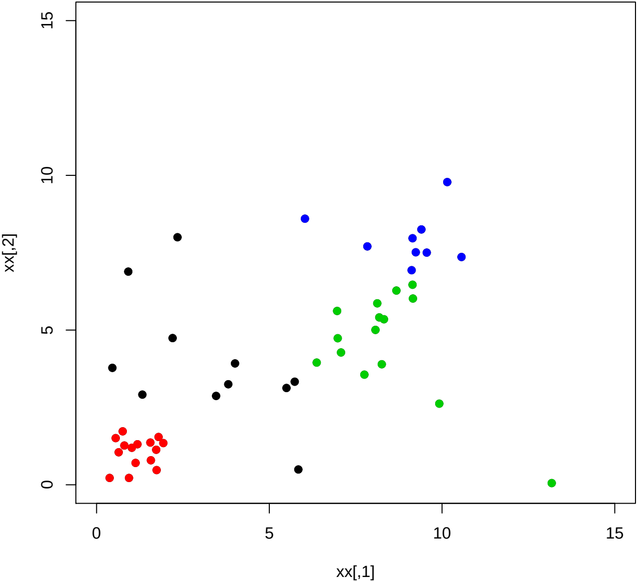
Ward's information criterion



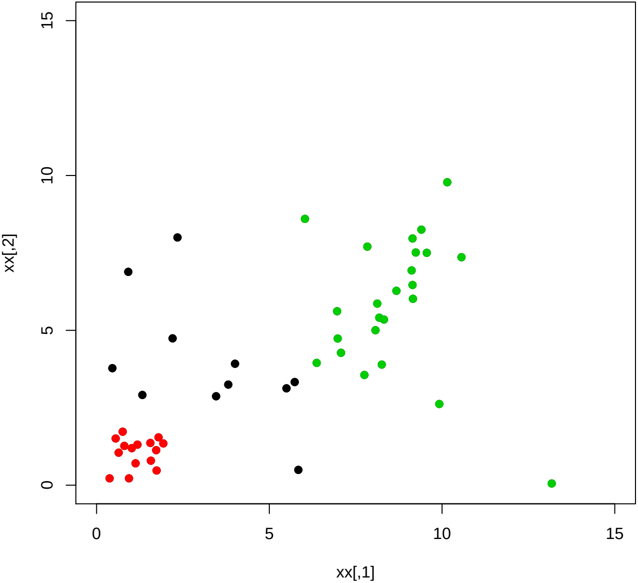
Ward’s information criterion



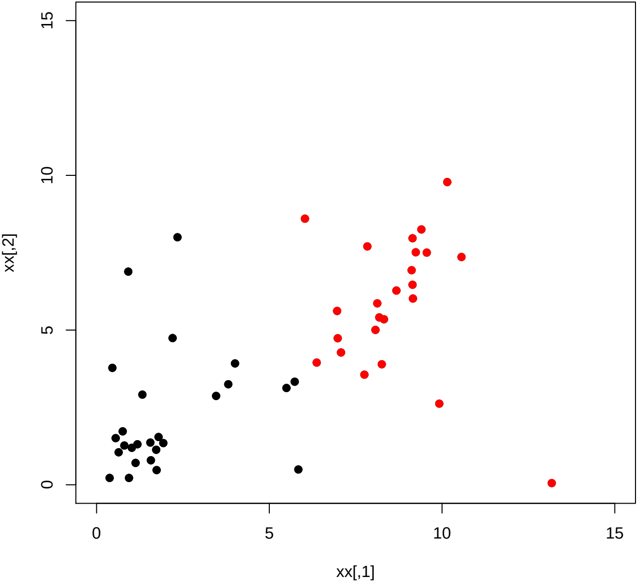
Ward's information criterion



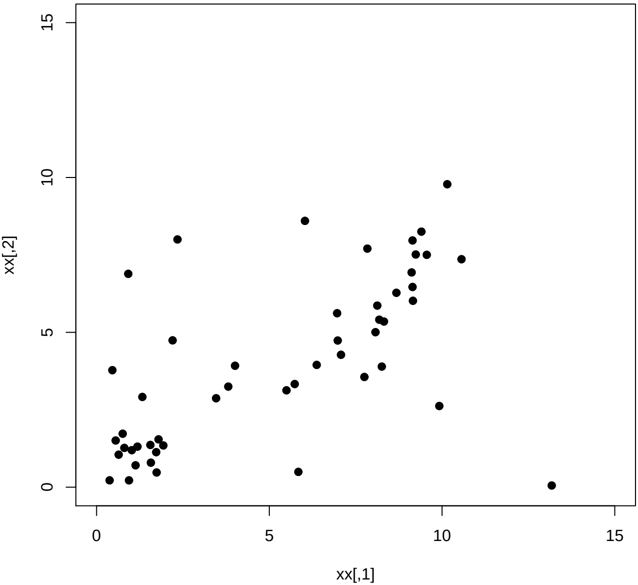
Ward's information criterion



Ward's information criterion



Ward's information criterion



Languages

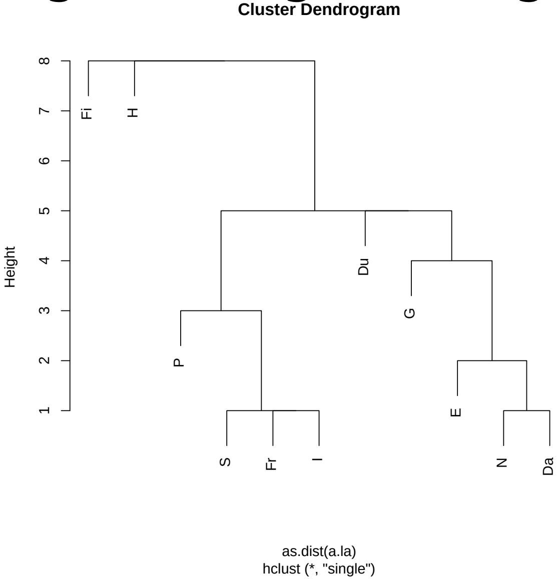
TABLE 12.3 NUMERALS IN 11 LANGUAGES

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

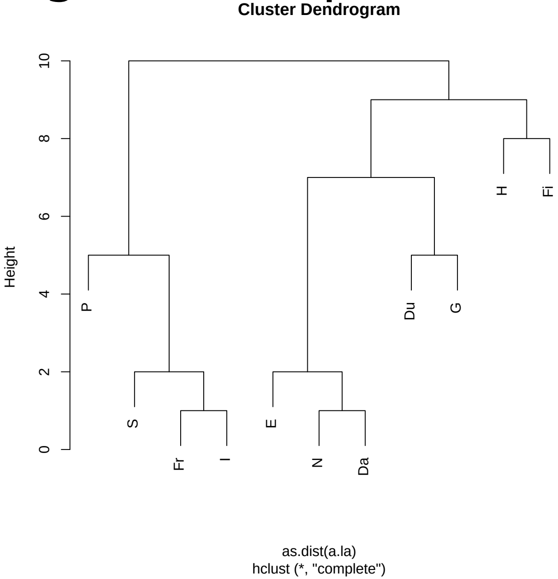
Languages - Dissimilarities

	E	N	Da	Du	G	Fr	S	I	P	H	Fi
E											
N	2										
Da	2	1									
Du	7	5	6								
G	6	4	5	5							
Fr	6	6	6	9	7						
S	6	6	5	9	7	2					
I	6	6	5	9	7	1	1				
P	7	7	6	10	8	5	3	4			
H	9	8	8	8	9	10	10	10	10		
Fi	9	9	9	9	9	9	9	9	9	8	

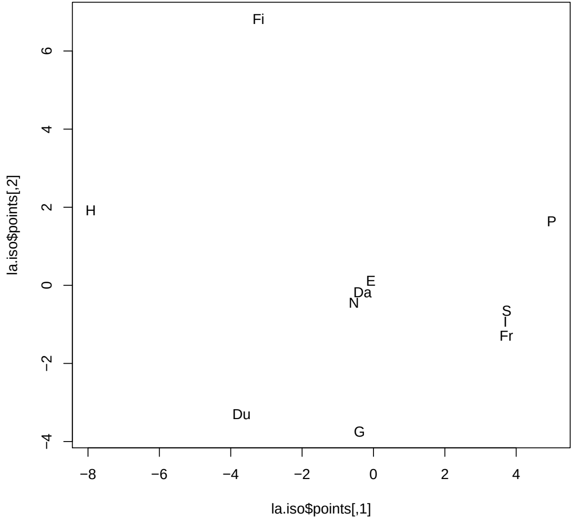
Languages - Single linkage



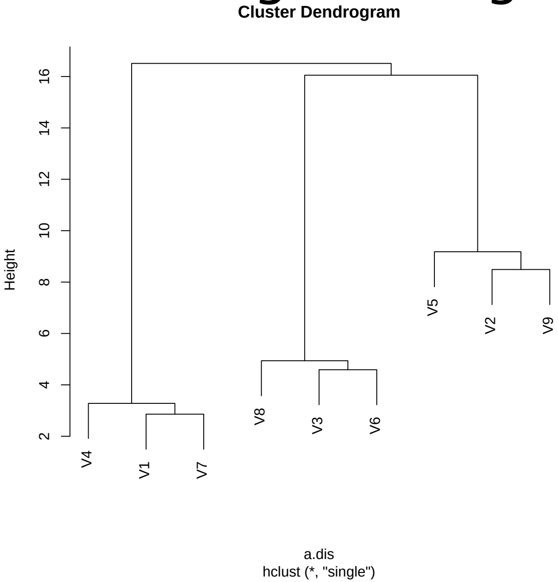
Languages - Complete linkage



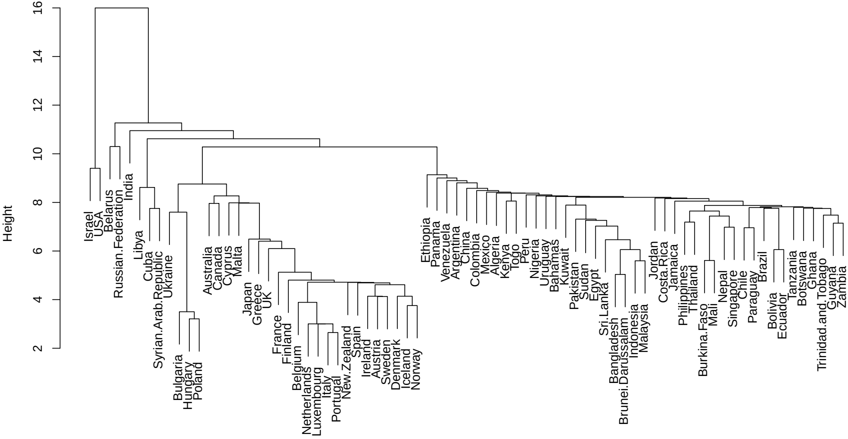
Languages - 2D representation via Multidimensional Scaling



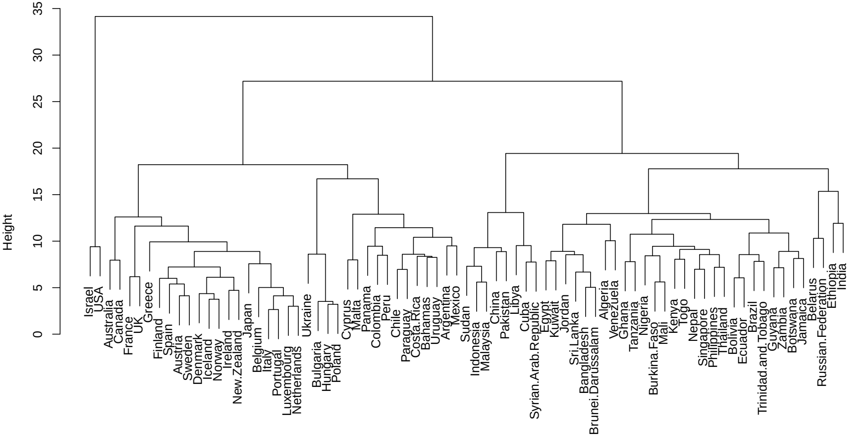
Breweries - Single linkage



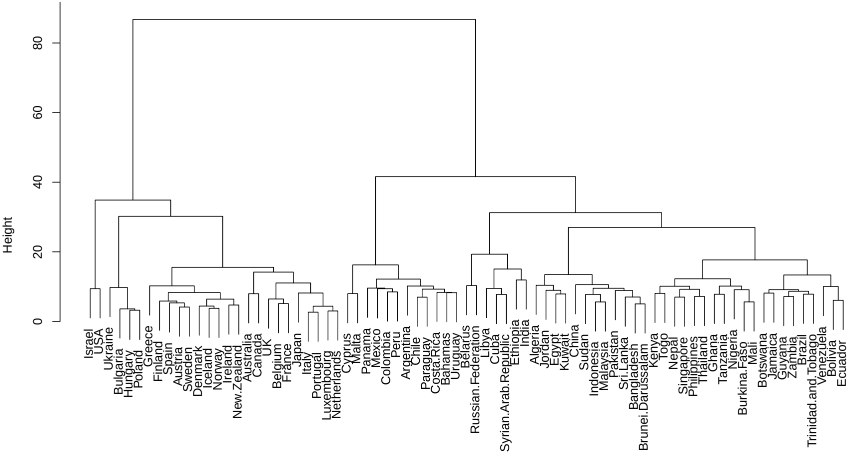
UN Votes - Single linkage



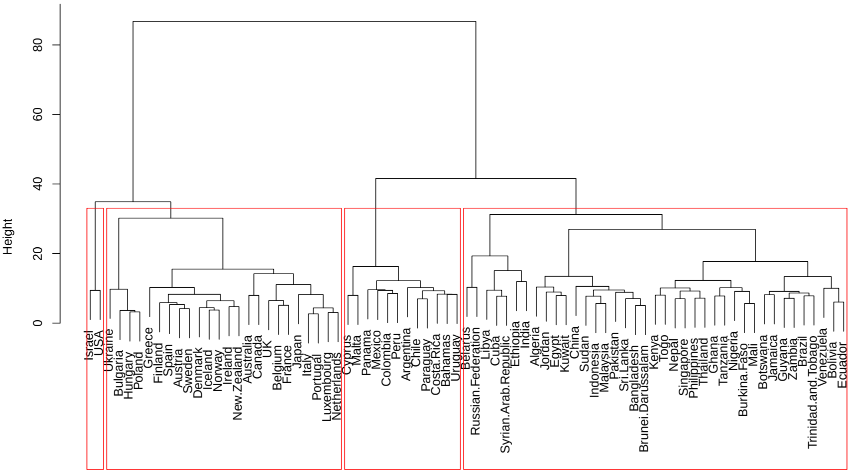
UN Votes - Complete linkage



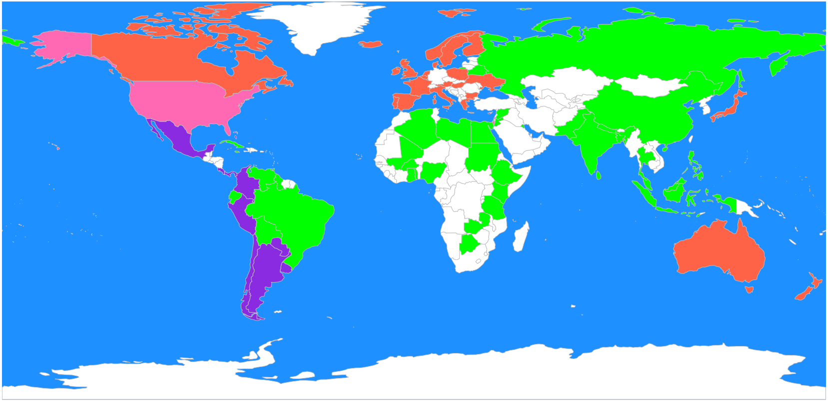
UN Votes - Ward linkage



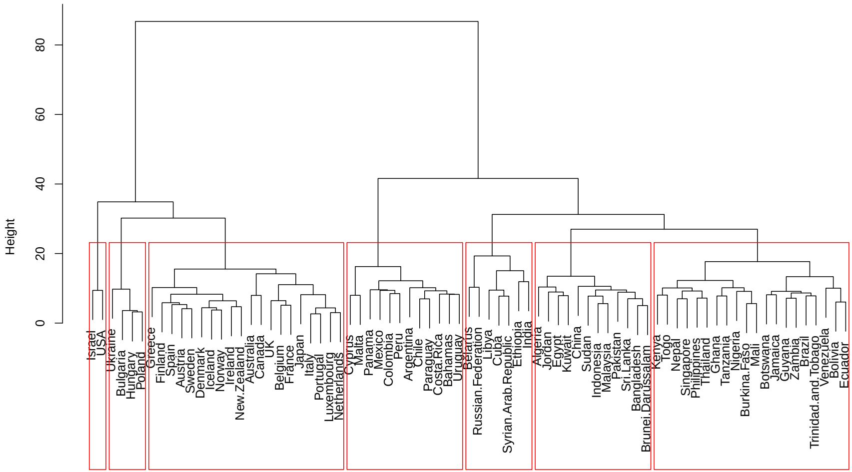
UN Votes - Ward linkage



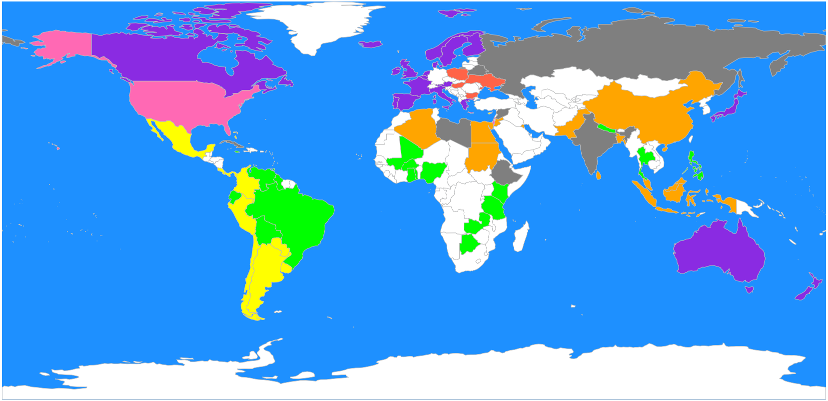
UN Votes - hierarchical - K=4



UN Votes - Ward linkage



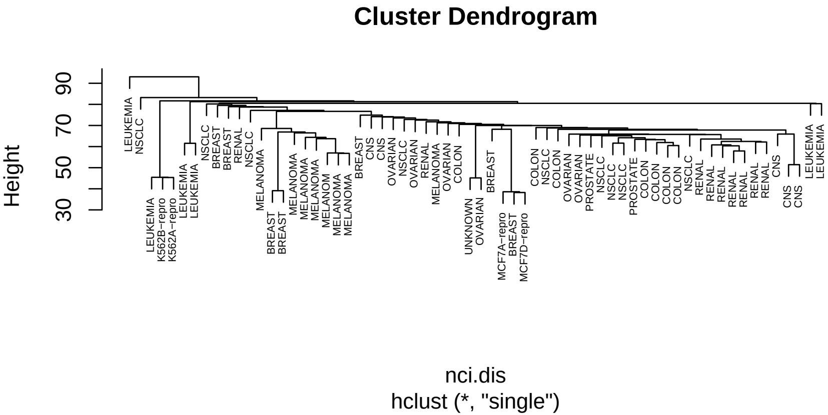
UN Votes - hierarchical - K=7



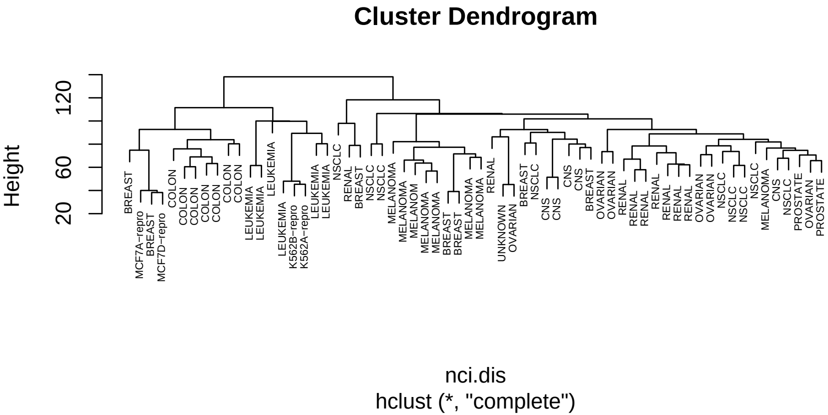
Cancer example

- Gene expression for 64 samples
- There are 6830 genes
- $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{64} \in \mathbb{R}^{6830}$
- We do know the label of each sample (which tissue this sample came from)
- The **real problem** is then “**variable selection**”

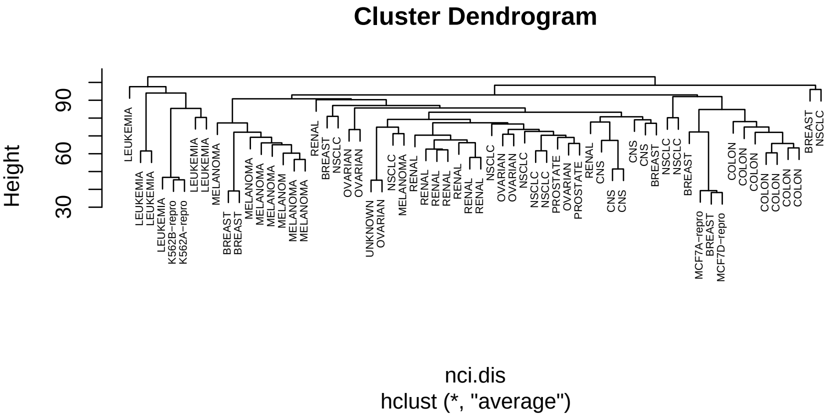
Cancer example - Single



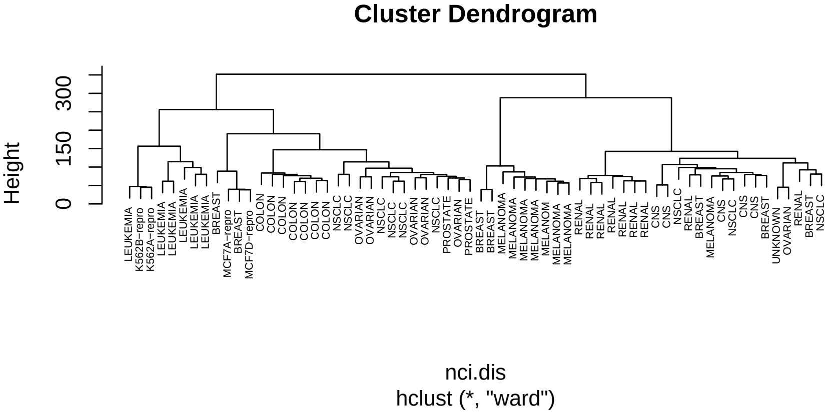
Cancer example - Complete



Cancer example - Average



Cancer example - Ward



Cancer example - Ward - 8 clusters