

20) K-Means Clustering and Hierarchical Clustering

Vitor Kamada

March 2018

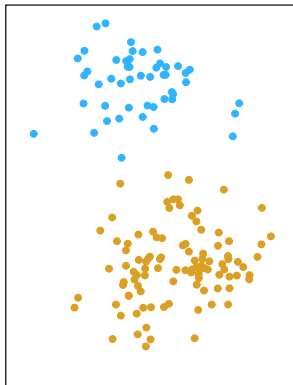
Tables, Graphics, and Figures from:

James et al. (2017): Ch 10.3, 10.5, and 10.6

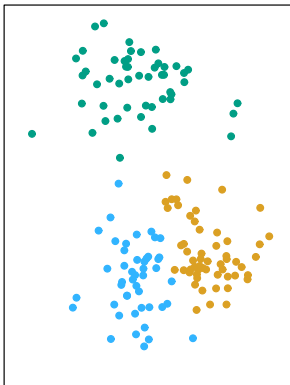
Hastie et al. (2017): Ch14.3

Simulated Data Set with 150 Observations

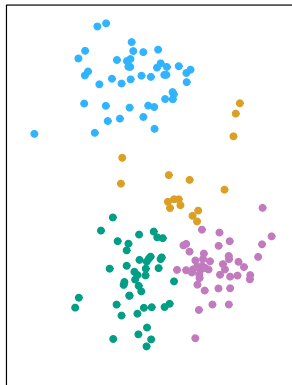
K=2



K=3



K=4



K-Means Clustering

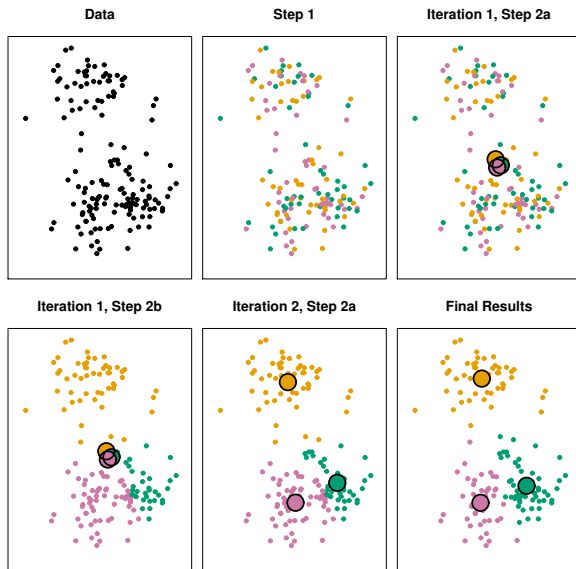
$$C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

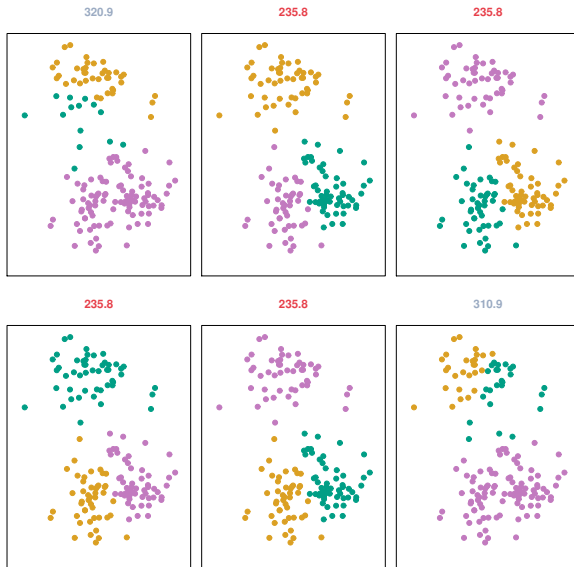
$$\underset{C_1, \dots, C_K}{\text{Minimize}} \left\{ \sum_{k=1}^K W(C_k) \right\}$$

$$W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

K-Means Clustering Algorithm



Different Random Assignment



```
set.seed(2)
```

```
x=matrix(rnorm(50*2), ncol=2)
```

```
x[1:25,1]=x[1:25,1]+3
```

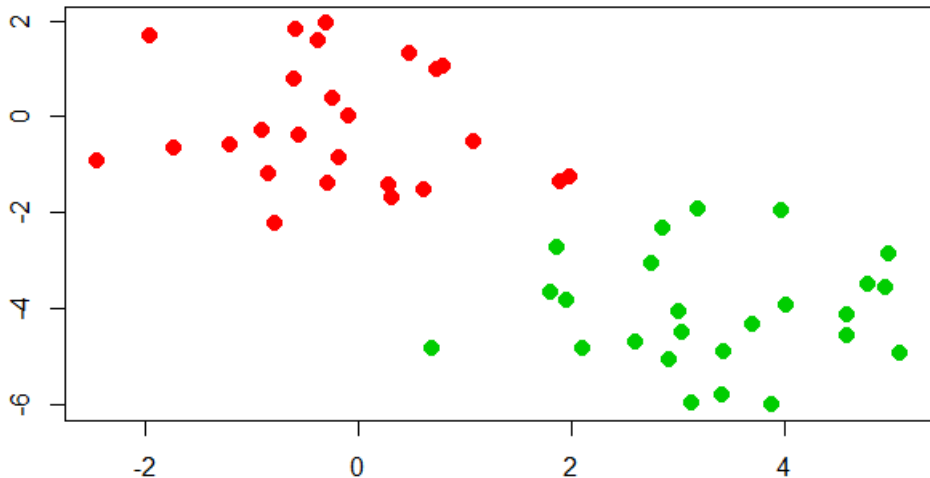
```
x[1:25,2]=x[1:25,2]-4
```

```
km.out=kmeans(x,2,nstart=20)
```

```
km.out$cluster
```

```
[1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
[26] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
plot(x, col=(km.out$cluster+1), xlab="", ylab="",  
pch=20, cex=2)
```




```
set.seed(4); km.out=kmeans(x,3,nstart=20)
```

km.out

K-means clustering with 3 clusters of sizes 10, 23, 17

cluster means:

	[,1]	[,2]
1	2.3001545	-2.69622023
2	-0.3820397	-0.08740753
3	3.7789567	-4.56200798

Clustering vector:

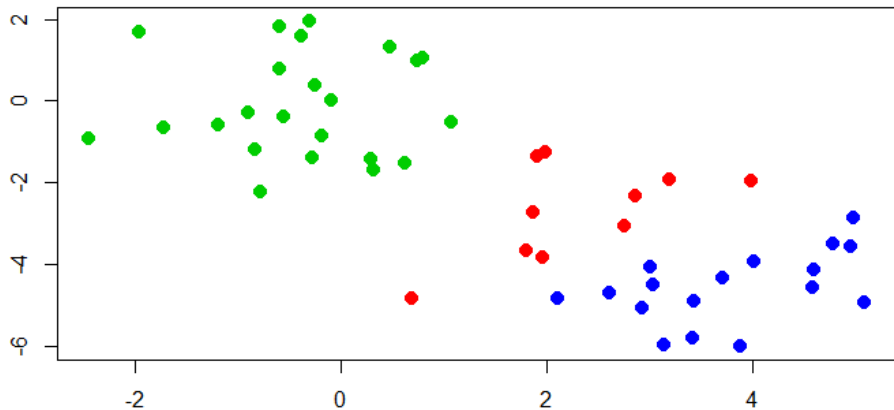
[1]	3	1	3	1	3	3	1	3	1	3	1	3	1	3	3	3	3	3	1	3	3	3	
[26]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	1	2	2	2

within cluster sum of squares by cluster:

[1]	19.56137	52.67700	25.74089
-----	----------	----------	----------

(between_ss / total_ss = 79.3 %)

```
plot(x, col=(km.out$cluster+1), xlab="", ylab="",  
pch=20, cex=2)
```



```
set.seed(3)
```

```
km.out=kmeans(x,3,nstart=1)
```

```
km.out$tot.withinss
```

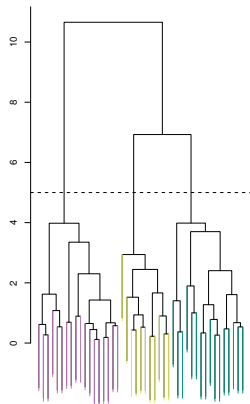
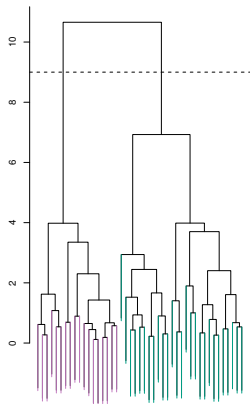
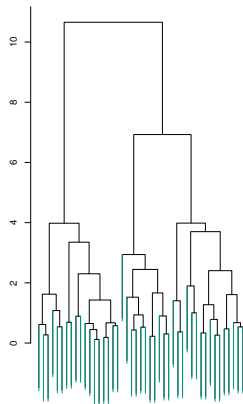
104.33

```
km.out=kmeans(x,3,nstart=20)
```

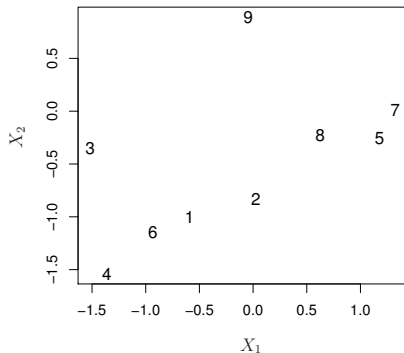
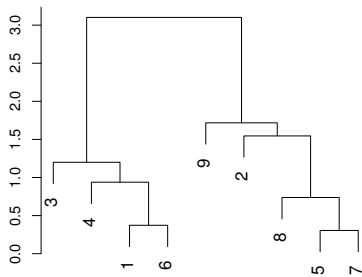
```
km.out$tot.withinss
```

97.97

Hierarchical Clustering - Dendrogram



9 Observations: Euclidean Distance and Complete Linkage



Measure of Dissimilarity ($d(G, H)$)

Single Linkage or Nearest-Neighbor

$$\min_{i \in G, i' \in H} d_{ii'}$$

Complete Linkage or Furthest-Neighbor

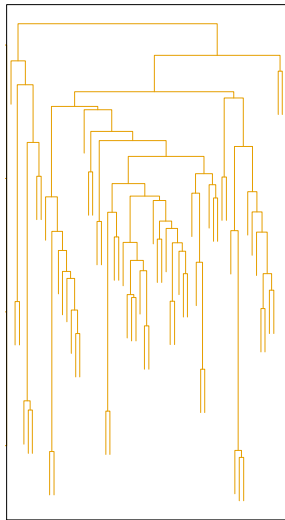
$$\max_{i \in G, i' \in H} d_{ii'}$$

Group Average

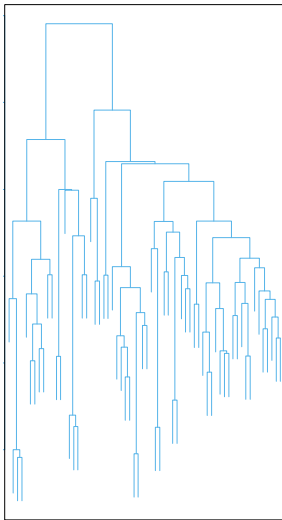
$$\frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

Average, Complete, and Single Linkage

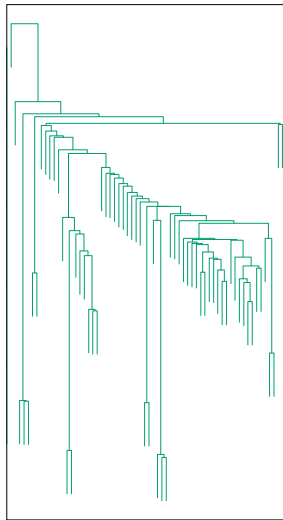
Average Linkage



Complete Linkage



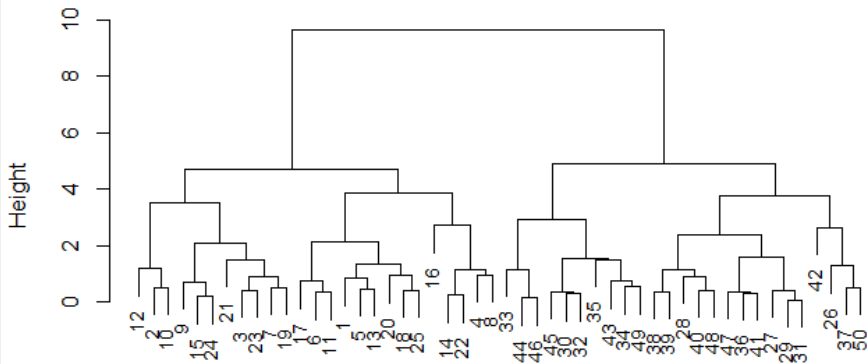
Single Linkage



```
hc.complete=hclust(dist(x), method="complete")
```

```
plot(hc.complete,main="Complete Linkage",  
xlab="", sub="", cex=.9)
```

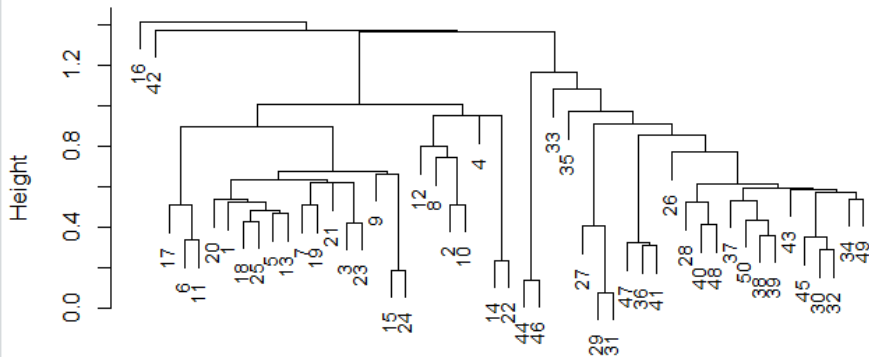
Complete Linkage




```
hc.single=hclust(dist(x), method="single")
```

```
plot(hc.single, main="Single Linkage", xlab="",  
sub="", cex=.9)
```

Single Linkage



```
cutree(hc.complete, 2)
```

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
[26] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
cutree(hc.average, 2)
```

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
[26] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
cutree(hc.single, 2)
```

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
[26] 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2
```

```
cutree(hc.single, 4)
```

```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1  
[26] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3
```

library(ISLR)

```
nci.labs=NCI60$labs
```

```
nci.data=NCI60$data
```

```
dim(nci.data)
```

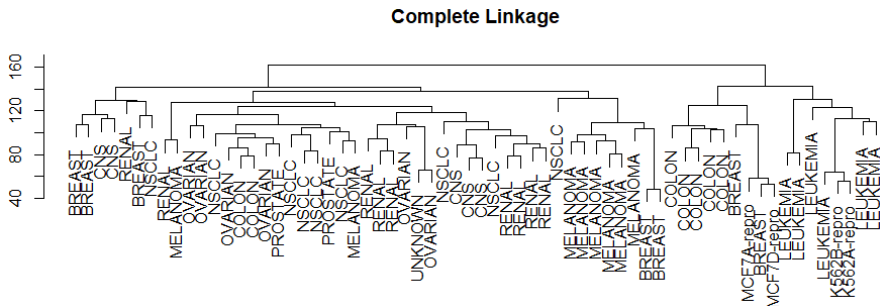
64 6830

```
table(nci.labs)
```

BREAST	CNS	COLON	K562A-repro	K562B-repro
7	5	7	1	1
LEUKEMIA	MCF7A-repro	MCF7D-repro	MELANOMA	NSCLC
6	1	1	8	9
OVARIAN	PROSTATE	RENAL	UNKNOWN	
6	2	9	1	

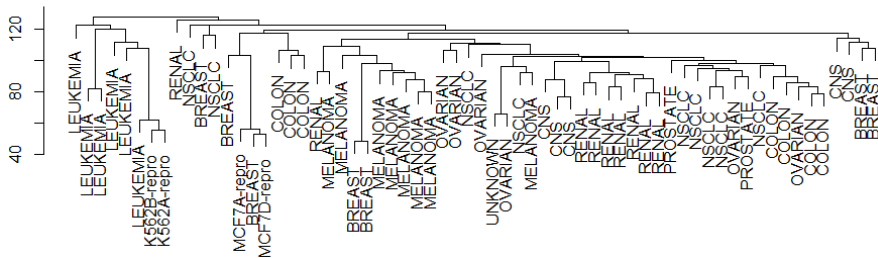
```
sd.data=scale(nci.data); data.dist=dist(sd.data)
```

```
plot(hclust(data.dist), labels=nci.labs, main="Complete  
Linkage", xlab="", sub="", ylab="")
```

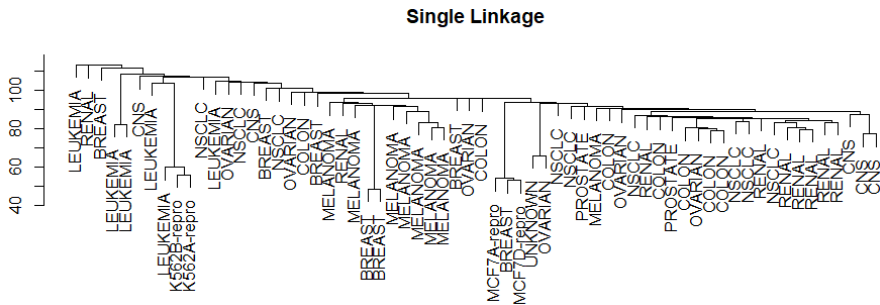


```
plot(hclust(data.dist, method="average"),
labels=nci.labs, main="Average Linkage", xlab="",
sub="",ylab="")
```

Average Linkage



```
plot(hclust(data.dist, method="single"),
labels=nci.labs, main="Single Linkage", xlab="",
sub="",ylab="")
```



```
hc.out=hclust(dist(sd.data))
```

```
hc.clusters=cutree(hc.out,4)
```

```
table(hc.clusters,nci.labs)
```

```
hc.clusters  BREAST  CNS  COLON  K562A-repro  K562B-repro  LEUKEMIA  MCF7A-repro
1           2    3    2           0           0           0           0
2           3    2    0           0           0           0           0
3           0    0    0           1           1           6           0
4           2    0    5           0           0           0           1
```

nci.labs

```
hc.clusters  MCF7D-repro  MELANOMA  NSCLC  OVARIAN  PROSTATE  RENAL  UNKNOWN
1              0          8      8        6        2      8        1
2              0          0      1        0        0      1        0
3              0          0      0        0        0      0        0
4              1          0      0        0        0      0        0
```



```
set.seed(2)
```

```
km.out=kmeans(sd.data, 4, nstart=20)
```

```
km.clusters=km.out$cluster
```

```
table(km.clusters,labels=nci.labs)
```

km.clusters	BREAST	CNS	COLON	K562A-repro	K562B-repro	LEUKEMIA	MCF7A-repro
1	2	0	7	0	0	0	1
2	0	0	0	1	1	6	0
3	2	0	0	0	0	0	0
4	3	5	0	0	0	0	0

labels

km.clusters	MCF7D-repro	MELANOMA	NSCLC	OVARIAN	PROSTATE	RENAL	UNKNOWN
1	1	0	5	3	1	0	0
2	0	0	0	0	0	0	0
3	0	7	0	0	0	0	0
4	0	1	4	3	1	9	1

```
table(km.clusters, hc.clusters)
```

	hc.clusters			
km.clusters	1	2	3	4
1	11	0	0	9
2	0	0	8	0
3	9	0	0	0
4	20	7	0	0