Data Mining
Laboratory

Ewa Figielska

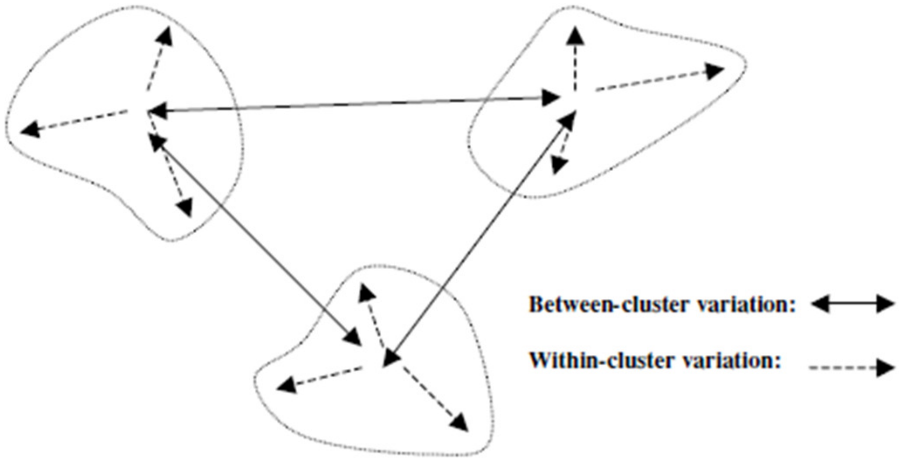
WARSAW SCHOOL
OF COMPUTER SCIENCE

Clustering

- **Clustering** - grouping of records, observations, or cases into classes of similar objects.
- A **cluster** is a collection of records that are similar to one another and dissimilar to records in other clusters.

- Clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized.

- Clustering algorithms seek to construct clusters of records such that the **between-cluster variation** (BCV) is large compared to the **within-cluster variation** (WCV).



BCV and WCV (source: Larose “Discovering knowledge in data” (2015))

Examples

- Examples of clustering tasks in business and research include:
  - For accounting auditing purposes, to segment financial behavior into benign and suspicious categories
  - As a dimension-reduction tool when a data set has hundreds of attributes
  - For gene expression clustering, where very large quantities of genes may exhibit similar behavior

Clustering.Decisions

- Before the cluster analysis is performed some preliminary decisions have to be made:
  - How to measure similarity,
  - How to recode categorical variables,
  - How to standardize or normalize numerical variables,
  - How many clusters we expect to uncover.

- The **Euclidean distance** –simple and often used distance measure (distance between two records $\boldsymbol{x}$ and $\boldsymbol{y}$)

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

where $\boldsymbol{x} = x_1, x_2, \dots, x_m$ and $\boldsymbol{y} = y_1, y_2, \dots, y_m$ represent m attribute values of two records.

- For categorical variables: the “different from” function can be defined for comparing the ith attribute values of a pair of records:

$$different(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{otherwise} \end{cases}$$

where x_i and y_i are categorical values.

Preprocessing for the clustering

- For optimal performance, clustering algorithms, require the data to be normalized so that no particular variable or subset of variables dominates the analysis.

- Analysts may use:
  - **min–max normalization**

$$X^* = \frac{X - \min(X)}{\mathrm{Range}(X)}$$

where $\mathrm{Range}(X) = \max(X) - \min(X)$.

  - **Z-score standardization:**

$$X^* = \frac{X - \mathrm{mean}(X)}{\mathrm{SD}(X)}$$

k-means algoritm for clustering

- The *k*-means clustering algorithm is a straightforward and effective algorithm for finding clusters in data.

- Solution representation: vector of cluster centers (centroids) $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_K})$
- Data set : set of n records $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}, \boldsymbol{a_3}, \dots, \boldsymbol{a_n}\}$

- The algorithm proceeds as follows.

1. Set the number of clusters K;
2. Set randomly the initial cluster center locations $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_K}$;
3. Assign record $\boldsymbol{a_i}$, $i = 1, \dots, n$ to cluster C_k $(k \in \{1,2, \dots, K\})$, if

   $d(\boldsymbol{a_i}, \boldsymbol{x_k}) < d(\boldsymbol{a_i}, \boldsymbol{x_j})$ for $j = 1, \dots, K$ and $j \neq k$;
4. Calculate new cluster centers $\boldsymbol{x'_1}, \boldsymbol{x'_2}, \dots, \boldsymbol{x'_K}$ from to the following expression:

   $$\boldsymbol{x'_k} = \frac{1}{n_k} \sum_{a_i \in C_k} \boldsymbol{a}_i \text{ for } k = 1, \dots, K,$$

   where n_k is the number of records in cluster C_k;
5. If $\boldsymbol{x'_k} = \boldsymbol{x_k}$ for $k = 1, \dots, K$, then stop, otherwise goto Step 3.

Example – generating the initial solution. Steps 1 & 2

- Data set

Record #	v	w
1	6.1	3.2
2	7.8	2.4
3	6.8	2.9
4	1.4	7.9
5	3.2	7.3
6	2.4	7.6
7	2.8	6.5
8	9.1	7.7
9	8.8	7.2
10	8.3	9.5

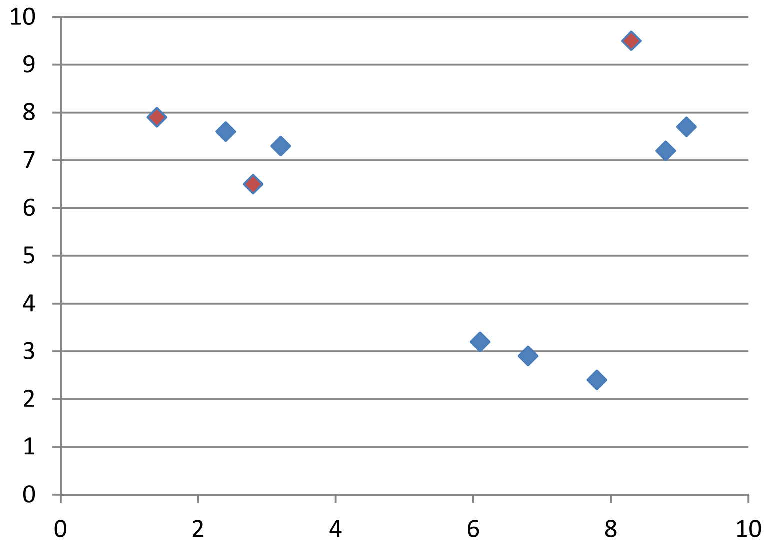
Data set contains $n = 10$ records.
Assume that there are $K = 3$ clusters and the randomly chosen initial solution (three centroids) is as follows:

$$\boldsymbol{x} = ((1.4, 7.9), (2.8, 6.5), (8.3, 9.5))$$

i.e.

$$\begin{aligned} x_{1v} &= 1.4, & x_{1w} &= 7.9 \\ x_{2v} &= 2.8, & x_{2w} &= 6.5 \\ x_{3v} &= 8.3, & x_{3w} &= 9.5 \end{aligned}$$

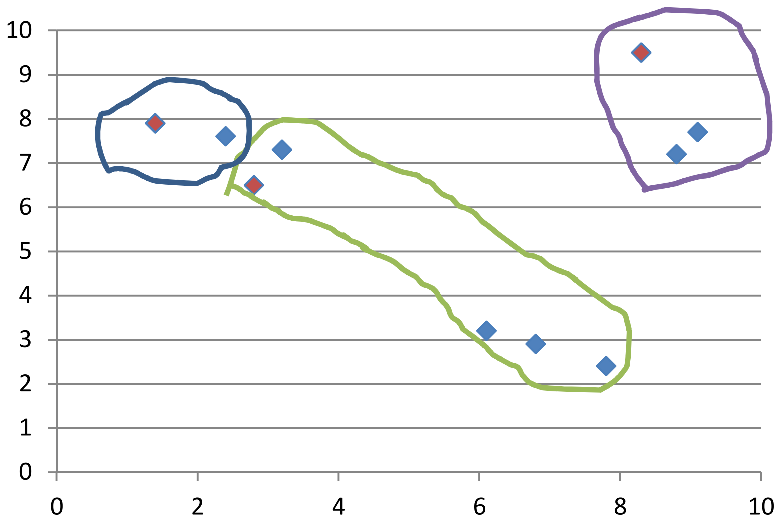
They are randomly chosen from among the data points.



Scatter plot of the data set with randomly chosen centroids

1st iteration. Steps 3 & 4

			x_1			x_2			x_3		
			$x_{1v} = 1.40$	$x_{1w} = 7.9$		$x_{2v} = 2.80$	$x_{2w} = 6.50$		$x_{3v} = 8.40$	$x_{3w} = 9.50$	
					cluster C_1			cluster C_2			cluster C_3
Record i	**v**	**w**	$a_{iv} - x_{1v}$	$a_{iw} - x_{1w}$	$d(a_i, x_1)$	$a_{iv} - x_{2v}$	$a_{iw} - x_{2w}$	$d(a_i, x_2)$	$a_{iv} - x_{3v}$	$a_{iw} - x_{3w}$	$d(a_i, x_3)$
1	6.1	3.2	$6.1 - 1.4 = 4.7$	$3.2 - 7.9 = -4.7$	$\sqrt{4.7^2 + (-4.7)^2} = 6.65$	3.3	-3.3	**4.67**	-2.20	-6.3	**6.67**
2	7.8	2.4	$7.8 - 1.4 = 6.4$	$2.4 - 7.9 = -5.5$	$\sqrt{6.4^2 + (-5.5)^2} = 8.84$	5	-4.1	**6.47**	-0.5	-7.1	**7.12**
3	6.8	2.9	5.4	-5	7.36	4	-3.6	**5.38**	-1.5	-6.6	**6.77**
4	1.4	7.9	0	0	**0.00**	-1.4	1.4	1.98	-6.9	-1.6	**7.08**
5	3.2	7.3	1.8	-0.6	1.90	0.4	0.8	**0.89**	-5.1	-2.2	**5.55**
6	2.4	7.6	1	-0.3	**1.04**	-0.4	1.1	1.17	-5.9	-1.9	**6.20**
7	2.8	6.5	1.4	-1.4	1.98	0	0	**0.00**	-5.5	-3	**6.26**
8	9.1	7.7	7.7	-0.2	7.70	6.3	1.2	6.41	0.8	-1.8	**1.97**
9	8.8	7.2	7.4	-0.7	7.43	6	0.7	6.04	0.5	-2.3	**2.35**
10	8.3	9.5	6.9	1.6	7.08	5.5	3	6.26	0	0	**0.00**



Scatter plot with clusters

Clusters:

$$C_1 = \{a_4, a_6\}$$
$$C_2 = \{a_1, a_2, a_3, a_5, a_7\}$$
$$C_3 = \{a_8, a_9, a_{10}\}$$

New centroids:

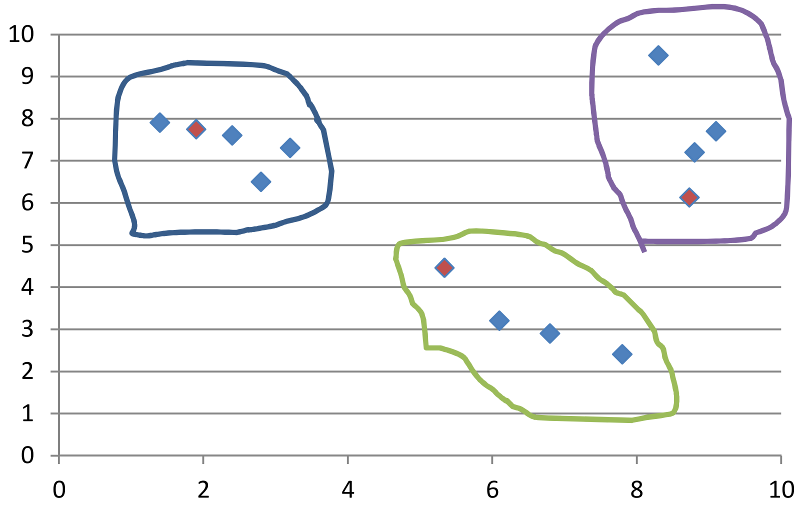
$$x'_{1v} = \frac{1.4 + 2.4}{2} = 1.9 \qquad x'_{1w} = \frac{7.9 + 7.6}{2} = 7.75$$

$$x'_{2v} = \frac{6.1 + 7.8 + 6.8 + 3.2 + 2.8}{5} = 5.34 \qquad x'_{2w} = \frac{3.2 + 2.4 + 2.9 + 7.3 + 6.5}{5} = 4.46$$

$$x'_{3v} = \frac{9.1 + 8.8 + 8.3}{3} = 8.73 \qquad x'_{3w} = \frac{7.7 + 7.2 + 9.5}{3} = 8.13$$

2nd iteration. Steps 3 & 4

			x_1			x_2			x_3		
			$x_{1v} = 1.90$	$x_{1w} = 7.75$		$x_{2v} = 5.34$	$x_{2w} = 4.46$		$x_{3v} = 8.73$	$x_{3w} = 8.13$	
					cluster C_1			cluster C_2			cluster C_3
Record *i*	**v**	**w**	$a_{iv} - x_{1v}$	$a_{iw} - x_{1w}$	$d(a_i, x_1)$	$a_{iv} - x_{2v}$	$a_{iw} - x_{2w}$	$d(a_i, x_2)$	$a_{iv} - x_{3v}$	$a_{iw} - x_{3w}$	$d(a_i, x_3)$
1	6.1	3.2	4.20	-4.55	6.19	0.76	-1.26	**1.47**	-2.63	-4.93	5.59
2	7.8	2.4	5.90	-5.35	7.96	2.46	-2.06	**3.21**	-0.93	-5.73	5.81
3	6.8	2.9	4.90	-4.85	6.89	1.46	-1.56	**2.14**	-1.93	-5.23	5.58
4	1.4	7.9	-0.50	0.15	**0.52**	-3.94	3.44	5.23	-7.33	-0.23	7.34
5	3.2	7.3	1.30	-0.45	**1.38**	-2.14	2.84	3.56	-5.53	-0.83	5.60
6	2.4	7.6	0.50	-0.15	**0.52**	-2.94	3.14	4.30	-6.33	-0.53	6.36
7	2.8	6.5	0.90	-1.25	**1.54**	-2.54	2.04	3.26	-5.93	-1.63	6.15
8	9.1	7.7	7.20	-0.05	7.20	3.76	3.24	4.96	0.37	-0.43	**0.57**
9	8.8	7.2	6.90	-0.55	6.92	3.46	2.74	4.41	0.07	-0.93	**0.94**
10	8.3	9.5	6.40	1.75	6.63	2.96	5.04	5.84	-0.43	1.37	**1.43**



Scatter plot with clusters

Clusters:

$C_1 = \{a_4, a_5, a_6, a_7\}$
$C_2 = \{a_1, a_2, a_3\}$
$C_3 = \{a_8, a_9, a_{10}\}$

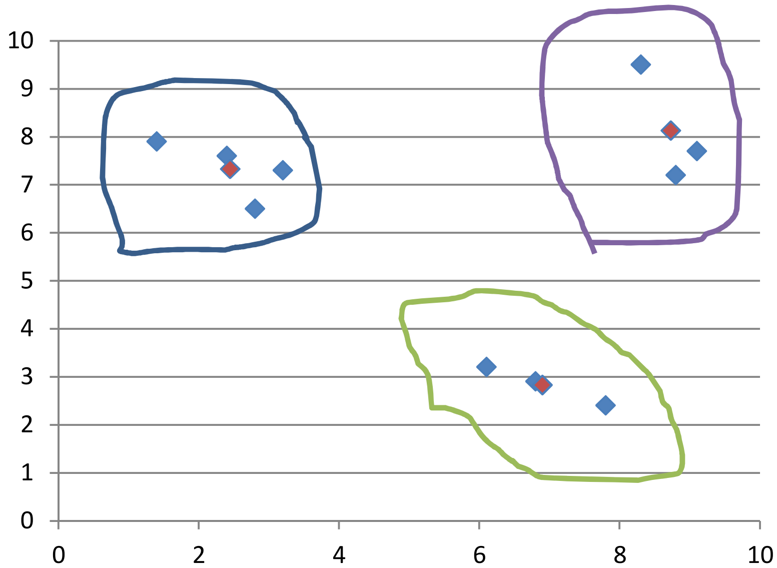
New centroids: $x'_{1v} = 2.45$ $x'_{1w} = 7.33$

$x'_{2v} = 6.90$ $x'_{2w} = 2.83$

$x'_{3v} = 8.73$ $x'_{3w} = 8.13$

3rd iteration. Steps 3 & 4

			x_1			x_2			x_3		
			$x_{1v} = 2.45$	$x_{1w} = 7.33$		$x_{2v} = 6.90$	$x_{2w} = 2.83$		$x_{3v} = 8.73$	$x_{3w} = 8.13$	
					cluster C_1			cluster C_2			cluster C_3
Record i	**v**	**w**	$a_{iv} - x_{1v}$	$a_{iw} - x_{1w}$	$d(a_i, x_1)$	$a_{iv} - x_{2v}$	$a_{iw} - x_{2w}$	$d(a_i, x_2)$	$a_{iv} - x_{3v}$	$a_{iw} - x_{3w}$	$d(a_i, x_3)$
1	6.1	3.2	3.65	-4.13	5.51	-0.80	0.37	**0.88**	-2.63	-4.93	5.59
2	7.8	2.4	5.35	-4.93	7.28	0.90	-0.43	**1.00**	-0.93	-5.73	5.80
3	6.8	2.9	4.35	-4.43	6.21	-0.10	0.07	**0.12**	-1.93	-5.23	5.57
4	1.4	7.9	-1.05	0.57	**1.19**	-5.50	5.07	7.48	-7.33	-0.23	7.33
5	3.2	7.3	0.75	-0.03	**0.75**	-3.70	4.47	5.80	-5.53	-0.83	5.59
6	2.4	7.6	-0.05	0.27	**0.27**	-4.50	4.77	6.56	-6.33	-0.53	6.35
7	2.8	6.5	0.35	-0.83	**0.90**	-4.10	3.67	5.50	-5.93	-1.63	6.15
8	9.1	7.7	6.65	0.37	6.66	2.20	4.87	5.34	0.37	-0.43	**0.57**
9	8.8	7.2	6.35	-0.13	6.35	1.90	4.37	4.77	0.07	-0.93	**0.93**
10	8.3	9.5	5.85	2.17	6.24	1.40	6.67	6.82	-0.43	1.37	**1.44**



Scatter plot with clusters

Clusters:

$$C_1 = \{a_4, a_5, a_6, a_7\}$$
$$C_2 = \{a_1, a_2, a_3\}$$
$$C_3 = \{a_8, a_9, a_{10}\}$$

New centroids: $x'_{1v} = 2.45$ $x'_{1w} = 7.33$

$x'_{2v} = 6.90$ $x'_{2w} = 2.83$

$x'_{3v} = 8.73$ $x'_{3w} = 8.13$

Algorithm stops (centroids no longer change)

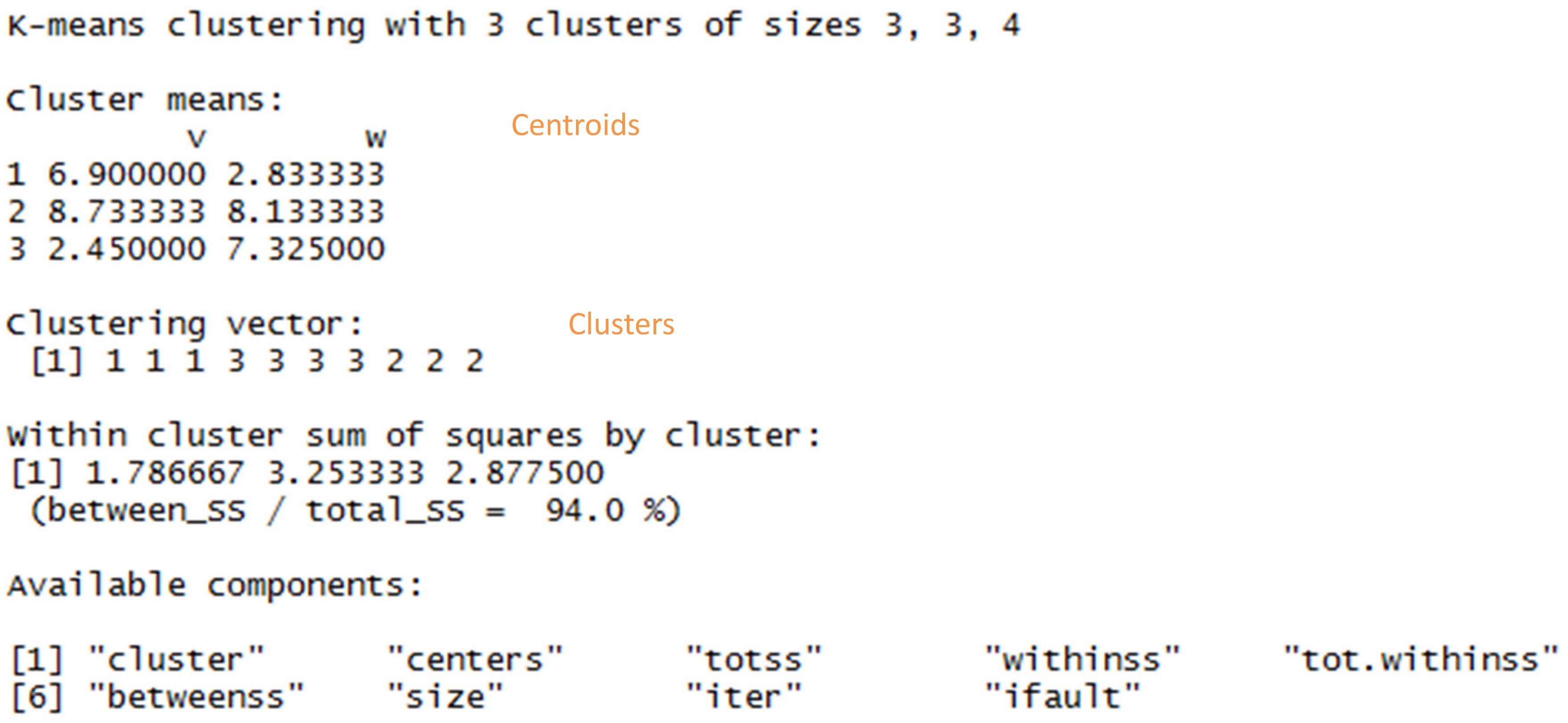
k-means algorithm in R

```
> d<-read.csv(file="data.csv",header=TRUE,sep=',')
> d
> km<-kmeans(d,centers=3)
> km
```

$x'_{1v} = 2.45$ $x'_{1w} = 7.33$

$x'_{2v} = 6.90$ $x'_{2w} = 2.83$

$x'_{3v} = 8.73$ $x'_{3w} = 8.13$



```
K-means clustering with 3 clusters of sizes 3, 3, 4

Cluster means:
         v        w
1 6.900000 2.833333
2 8.733333 8.133333
3 2.450000 7.325000

Clustering vector:
 [1] 1 1 1 3 3 3 3 2 2 2

Within cluster sum of squares by cluster:
[1] 1.786667 3.253333 2.877500
 (between_SS / total_SS =  94.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Centroids

Clusters

Evaluation of the results

- Clustering algorithms seek to construct clusters of records such that the between-cluster variation is large compared to the within cluster variation.
- **Within sum of squares** (WSS) or **Sum of squares error** (SSE):

$$WSS = SSE = \sum_{k=1}^{K} \sum_{a_i \in C_k} d(a_i, x_k)^2$$

where a_i repesents a data point, x_k is a centroid of cluster k.

- **Sum of squares between** clusters (SSB):

$$SSB = \sum_{k=1}^{K} n_k d(x_i, \mu)^2$$

where n_k the number of records in cluster k, μ is the mean of all the data.

- **Total sum of squares** (TSS):

$$TSS = \sum_{i=1}^{n} d(a_i, \mu)^2$$

- **Quality measures:**
  - $\boldsymbol{SSB/TSS}$ should be close to 1
  - $(\boldsymbol{TSS - WSS})/\boldsymbol{TSS}$ should be close to 1 (so called R^2 criterion)

Evaluation of the results for the example

- $SSE = \sum_{k=1}^{K} \sum_{a_i \in C_k} d(a_i, x_k)^2 = 1.19^2 + 0.75^2 + 0.27^2 + 0.9^2 + 0.88^2 + 1^2 + 0.12^2 + 0.57^2 + 0.93^2 + 1.44^2 = 7.91$

```
> km.tot.withins<-kmeans(d,centers=3)$tot.withinss
> km.tot.withins
[1] 7.9175
```

- $\mu = (\mu_v, \mu_w) = (5.67, 6.22)$ - the mean of the data set: for variables v and w
- $SSB = \sum_{k=1}^{K} n_k d(x_i, \mu)^2 = 4((2.45 - 5.67)^2 + (7.33 - 6.22)^2) + 3((6.9 - 5.67)^2 + (2.83 - 6.22)^2) + 3((8.73 - 5.67)^2 + (8.13 - 6.22)^2) = 124.44$

```
> km.betweenss<-kmeans(d,centers=3)$betweenss
> km.betweenss
[1] 124.4395
```

- $TSS = \sum_{i=1}^{n} d(a_i, \mu)^2 = (6.1 - 5.67)^2 + (3.2 - 6.22)^2 + \cdots . + (8.3 - 5.67)^2 + (9.5 - 6.22)^2 = 132.36$

```
> km.totss<-kmeans(d,centers=3)$totss
 > km.totss
[1] 132.357
```

- $\boldsymbol{SSB/TSS}$ **=** $\mathbf{124.44/132.36}$ **= 0.94**
- $(\boldsymbol{TSS} - \boldsymbol{WSS})/\boldsymbol{TSS}$ **= (132.36-7.91)/132.36 = 0.94**

The number of clusters

- Who decides how many clusters to search for?
- Unless the analyst has a priori knowledge of the number of underlying clusters, therefore, an “outer loop” should be added to the algorithm, which cycles through various promising values of *k*.
- Clustering solutions for each value of *k* can therefore be compared, with the value of *k* resulting in the smallest SSE being selected.

Next example. Iris

- Data set constains 150 recods with lengths and widths of 2 kind of petals for 3 types of iris flowers iris setosa, iris versicolor and iris virginica.

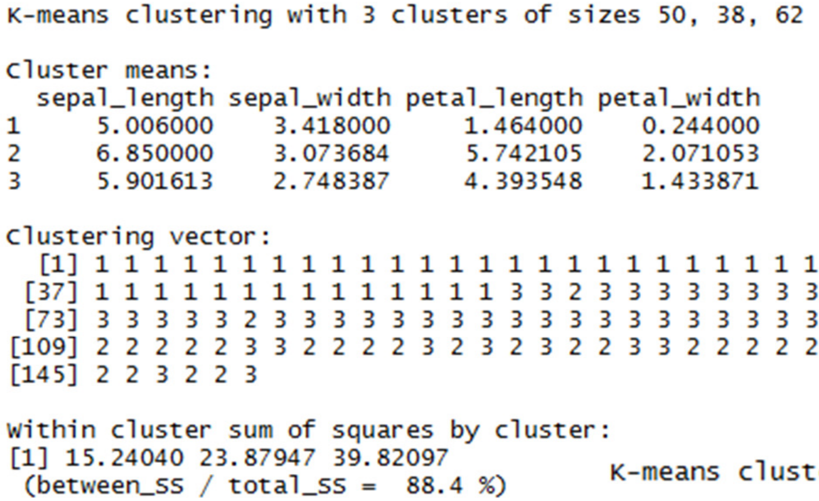
- Clustering is used to group flowers of the same type.

2004 © Peter M. Dziuk

Iris

```
d<-read.csv(file="iris.csv",header=TRUE,sep=',')[c('sepal_length',
'sepal_width','petal_length', 'petal_width')]
d
km<-kmeans(d,centers=3)
km
```

- The results can be different in different runs of the algorithm.
- The performance of the algorithm is affected by the initial solution.

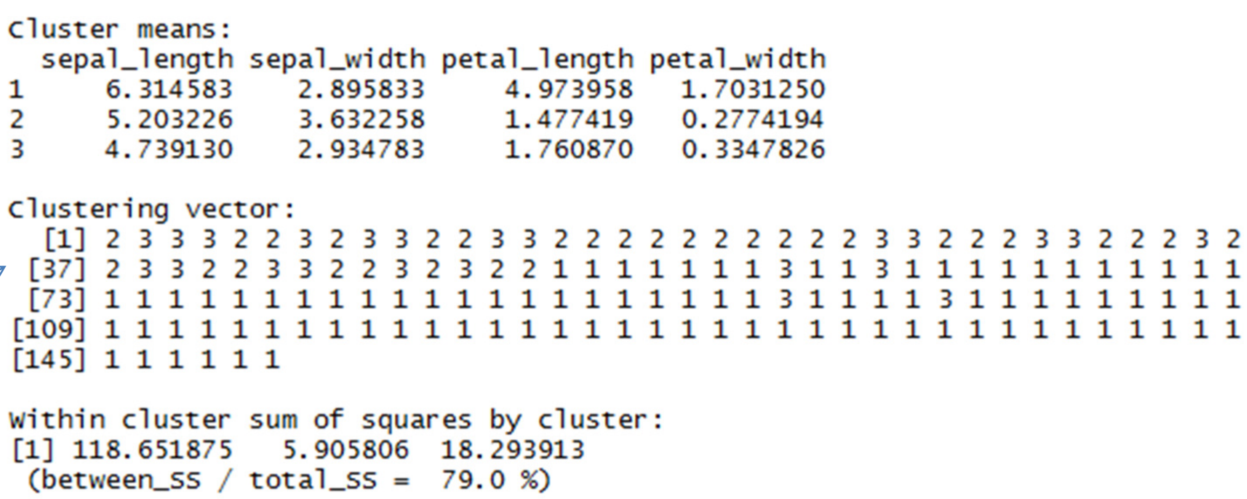


```
K-means clustering with 3 clusters of sizes 50, 38, 62

Cluster means:
  sepal_length sepal_width petal_length petal_width
1     5.006000    3.418000     1.464000    0.244000
2     6.850000    3.073684     5.742105    2.071053
3     5.901613    2.748387     4.393548    1.433871

Clustering vector:
  [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 1 1 1 1 1 1 1 1 1 1 1
 [37] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [73] 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 2 2 2 2 3 2
[109] 2 2 2 2 2 3 3 2 2 2 2 3 2 3 2 3 2 2 3 3 2 2 2 2 2 3 2 2 2 2 3 2 2 2 3 2
[145] 2 2 3 2 2 3

Within cluster sum of squares by cluster:
[1] 15.24040 23.87947 39.82097
 (between_SS / total_SS =  88.4 %)
```



```
K-means clustering with 3 clusters of sizes 96, 31, 23

Cluster means:
  sepal_length sepal_width petal_length petal_width
1     6.314583    2.895833     4.973958   1.7031250
2     5.203226    3.632258     1.477419   0.2774194
3     4.739130    2.934783     1.760870   0.3347826

Clustering vector:
  [1] 2 3 3 3 2 2 3 2 3 3 2 2 3 3 2 2 2 2 2 2 2 2 2 2 3 3 2 2 2 3 3 2 2 2 3 2
 [37] 2 3 3 2 2 3 3 2 2 3 2 3 2 2 1 1 1 1 1 1 1 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1
 [73] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1
[109] 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 1 1 1 1 1 1 1 1 1 1
[145] 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 118.651875   5.905806  18.293913
 (between_SS / total_SS =  79.0 %)
```

A quite great number of misclassification

English-polish dictionary

- **Cluster** – klaster, grupa, skupienie
- **Centroid** – centroid grupy, środek grupy,
- **Sum of squares error (**SSE) - suma kwadratów błędów (reszt), zmienność wewnątrzgrupowa
- **Sum of squares between (**SSB**)** – zmienność międzygrupowa
- **Total sum of squares (**TSS**)** – całkowita suma kwadratów, całkowita zmienność zmiennej