

Experiment-10

Problem Statement: To Implement k-medoids Clustering Algorithm using R Language.

Theoretical Background:

This algorithm can handle outliers well. Initially a random set of k-items is taken to be the set of medoids. Then at each step, all items from the input dataset that are not currently medoids are examined one by one to see if they should be medoids.

- **Within Cluster Sum of Squares :**

$$WSS = \sum_{i=1}^{N_c} \sum_{x \in C_i} d(x, \overline{x_{C_i}})^2$$

- **Between Cluster Sum of Squares :**

$$BSS = \sum_{i=1}^{N_c} |C_i| \cdot d(\overline{x_{C_i}}, \overline{x})^2$$

$C_i = \text{Cluster}$, $N_c = \# \text{ clusters}$, $\overline{x_{C_i}} = \text{Cluster centroid}$, $\overline{x} = \text{Sample Mean}$

Silhouette Coefficient :

$a(o)$: average distance between object o and the objects in its cluster A

$b(o)$: average distance between object o and the objects in its "second closest" cluster B

The silhouette of o is then defined as $s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$

- measures how good the assignment of o to its cluster is
 - $s(o) = -1$: bad, on average closer to members of B
 - $s(o) = 0$: in-between A and B
 - $s(o) = 1$: good assignment of o to its cluster A

Silhouette Coefficient s_c of a clustering: average silhouette of all objects

- $0.7 < s_c \leq 1.0$ strong structure, $0.5 < s_c \leq 0.7$ medium structure
- $0.25 < s_c \leq 0.5$ weak structure, $s_c \leq 0.25$ no structure

Function Description :

Pam function : It is a function available in **cluster** package, used for **clustering**.

pam(x, k, metric = "Euclidean", stand = FALSE)

- o **x** : numeric data matrix or numeric data frame
- o **k** : number of clusters
- o **metric** : the distance metrics to be used
- o **stand** : logical value (if TRUE, the variables in x are standardized before calculating dissimilarities)

Plots : It generates two plots

- i) **Clusplot** – Plot represents the clusters formed.
- ii) **Silhouetteplot** – Plot shows the silhouette lines.

Program :

```
library(cluster)
pam.result<-pam(iris,3)
print(pam.result)
```

```
table(pam.result$clustering,iris$Species)
```

```
pdf("Z.pdf")
```

```
plot(pam.result)
```

dev.off()

Output :

```
>print(pam.result)
```

Medoids:

ID Sepal.Length Sepal.Width Petal.Length Petal.Width Species

[1,] 8 5.0 3.4 1.5 0.2 1

[2,] 113 6.8 3.0 5.5 2.1 3

[3,] 100	5.7	2.8	4.1	1.3	2
----------	-----	-----	-----	-----	---

Clustering vector:

[1] 1

[51] 2 3 2 3

[101] 2 2 2 2 2 2 3 2

Objective function:

build swap

0.6940644 0.6840694

Available components:

```
[1] "medoids" "id.med" "clustering" "objective" "isolation" "clusinfo" "silinfo"
```

```
[8] "diss"    "call"    "data"
```

```
> table(pam.result$clustering,iris$Species)
```

setosa versicolor virginica

1	50	0	0
2	0	3	49
3	0	47	1

