PAGE NO.:
DATE: / /

K-means

Each cluster center represented by the mean value of the object in cluster

i/p K - no. of cluster
D - A dataset containing N objects

Methods steps:

Step 1: Arbitarily (randomly) choose K object from D as initial cluster center.

Step 2: repeat a) Re-assign each obj to the cluster to which the obj is most similar based on the mean value of the value in the cluster.
b) Update the cluster mean
(repeat until no change).

K-means centroid Based technique

Data set : 2, 4, 5, 6, 8, 12, 10, 3

K=3

$m_1 = 2 \quad m_2 = 4 \quad m_3 = 5$

$$E = \sum_{i=1}^{K} \sum_{p \in C_i} dist(p, C_i)^2$$

For object "2" dist (P, Ci) = (P, Ci)²

(2,2)

dist (P, C1) = (2-2)² = 0 *

(2,4) = (2-4)² = 4

(2,5) = (2-5)² = 9

For object "4"

dist (4,2) = (4-2)² = 4

(4,4) = (4-4)² = 0 *

(4,5) = (4-5)² = 1

For object = "5"

d(5,2) = (5-2)²

d(5,4) = (5-4) = 1

d(5,5) = 0 *

For object "6"

d(6,2) = (6-2)² = 16

d(6,4) = (6-4)² = 4

d(6,5) = (6-5)² = 1 *

For object "8"

d(8,2) = (8-2)² = 36

d(8,4) = (8-4)² = 16

d(8,5) = (8-5)² = 9 *

For object "12"

d(12,2) = (12-2)² = 100

d(12,4) = (12-4)² = 64

d(12,5) = (12-5)² = 49 *

PAGE NO.:
DATE: / /

For obj "10"

$d(10,2) = 64$

$d(10,4) = 36$

$d(10,5) = 25$ *

For object "3"

$d(3,2) = (3-2)^2 = 1$

$d(3,4) = (3-4)^2 = (-1)^2 = 1$

$d(3,5) = (3-5)^2 = 4$

$K_1 = 2,3$

$K_2 = 4$

$K_3 = 5,6,8,12,10$

update these mean values

$m_1 = \frac{2+3}{2} = \frac{5}{2} = 2.5$

$m_2 = 4$

$m_3 = 5+6+8+12+10/5 = 8.2$

Now again calculate Euclidean distance between center point & data object

$d(2,2.5) = (2-2.5)^2 =$

$d(2,4) = (2-4)^2 =$

$d(2,8.2) = (2-8.2)^2 =$

$d(4,2.5)$

$d(4,4)$

$d(4,8.2)$

PAGE NO.:
DATE: / /

d(5, 2.5)
d(5, 4)
d(5, 8.2)

d(6, 2.5)
d(6, 4)
d(6, 8.2)

d(8, 2.5)
d(8, 4)
d(8, 8.2)

d(12, 2.5)
d(12, 4)
d(12, 8.2)

d(10, 2.5)
d(10, 4)
d(10, 8.2)

d(3, 2.5)
d(3, 4)
d(3, 8.2)

$K_1 = 2, 3$
$K_2 = 4, 5, 6$
$K_3 = 8, 12, 10$

update centre points

$m_1 = 2 + 3/2 = 2.5$
$m_2 = 4 + 5 + 6 / 3 = 5$
$m_3 = 8 + 12 + 10 / 3 = 10$

PAGE NO.:
DATE: / /

Agent calculate the distance between cache pointer & data objects

d(4,2.5)
d(4,5)
d(4,10)

d(5,2.5)
d(5,5)
d(5,10)

d(6,2.5)
d(6,5)
d(6,10)

d(8,2.5)
d(8,5)
d(8,10)

d(12,2.5)
d(12,5)
d(12,10)

d(10,2.5)
d(10,5)
d(10,10)

d(3,2.5)
d(3,5)
d(3,10)

PAGE NO.:
DATE: / /

Recreate cluster

$K_1 = 2,3$

$K_2 = 4,5,6$

$K_3 = 8,12,10$

The two iterations having same cluster values so we need not be proceed further.

At last calculate

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} dist(p, ci)$$

$$= (2-2.5)^2 + (3-2.5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (8-10)^2 + (12-10)^2 + (10-10)^2$$

$$= 0.25 + 0.25 + 1 + 0 + 1 + 4 + 4 + 0$$

$$= 10.5$$

The time complexity of the k means algo is $O(nkt)$

n = the no. of obj.

k = no. of cluster

t = no. of iterations.