TEXT MINING DA 1

OM ASHISH MISHRA
16BCB0789

(1) K-Means

Iteration I

	C(2,8) C1	C(4,5) C4	C(2,8) C7	Belong to
C1 (2,8)	0	$\sqrt{13}$	0	C1
C2 (2,5)	3	2	3	C4
C3 (5,4)	5	$\sqrt{2}$	5	C4
C4 (4,5)	$\sqrt{13}$	0	$\sqrt{13}$	C4
C5 (3,4)	$\sqrt{17}$	$\sqrt{2}$	$\sqrt{17}$	C4
C6 (4,2)	$\sqrt{40}$	3	$\sqrt{40}$	C4
C7 (2,8)	0	$\sqrt{13}$	0	C1
C8 (3,6)	$\sqrt{5}$	$\sqrt{2}$	$\sqrt{5}$	C4

Since C1 = C4,

∴ C1's centroid = (2, 8)

C4's centroid = $\left(\frac{2+5+4+3+4+3}{6}, \frac{5+4+5+4+2+6}{6}\right)$

= (3.5, 4.33)

Iteration II

	(2,6) C1	(3.5, 4.33) C4	Belongs to
C1 (2,8)	0	4.9689	C1
C2 (2,5)	3	1.6489	C4
C3 (5,4)	5	1.5358	C4
C4 (4,5)	$\sqrt{13} = 3.6$	0.836	C4
C5 (3,4)	$\sqrt{17} = 4.1$	0.599	C4
C6 (4,2)	$\sqrt{10} = 6.32$	2.383	C4
C7 (2,8)	0	2.464	C1
C8 (3,6)	$\sqrt{5} = 2.23$	1.7	C4

i) The 3 centroids after the two iterations:-

C1: (2,8)

C4: (3.5, 4.33)

C1: (2,8)

ii) The 3 clusters' members after two iterations

C1 = C7 = (2,8)

C4 = { (2,5), (5,4), (4,5), (3,4), (4,2), (3,6) }

(2) Decision Tree

TF - IDF → Entropy (I.G)

Term 1	Term 2	Term 3	Class
2.1	10.4	12.1	X
2.5	12.5	18.5	Y
2.6	9.8	11.0	X
1.1	12.5	16.2	Y
1.4	12.1	11.2	X
1.6	10.6	10.5	Y

Information Gain

$$-\left[\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right] = 1$$

In order to calculate the entropy for preference

Term Since the X and Y are equal divided (each 3 in numbers). Therefore we take the mean of the sum of entropies in order to divide and form tree structure.

Term 1

$$\frac{2.1 + 2.5 + 2.6 + 1.1 + 1.4 + 1.6}{6} = 1.8$$

$\therefore$ (1.8 <) 2✗ 1✓

$$-\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$
$$= 0.918$$

(1.8 >) 2✓ 1✗

$$-\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$
$$= 0.918$$

Term 2

$$\frac{10.4 + 12.5 + 9.8 + 12.5 + 12.1 + 10.5}{6} = 11.3$$

$$-\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$
$$= 0.918$$

$\therefore$ (11.3 <) 2✓ 1✗

(11.3 >) 2✗ 1✓

$$-\left[\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]$$
$$= 0.918$$

Term 3

$$\frac{12.1 + 18.5 + 11.0 + 16.2 + 11.2 + 10.5}{6} = \frac{79.5}{6} = 13.25$$

$\therefore$ (13.25 <) 2✓ 0✗

$$-\left[\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] = 0$$

13.25 < 3 x
1 y

$$-\left[\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]$$

$$= 0.811$$

∴ The division Tree is :-