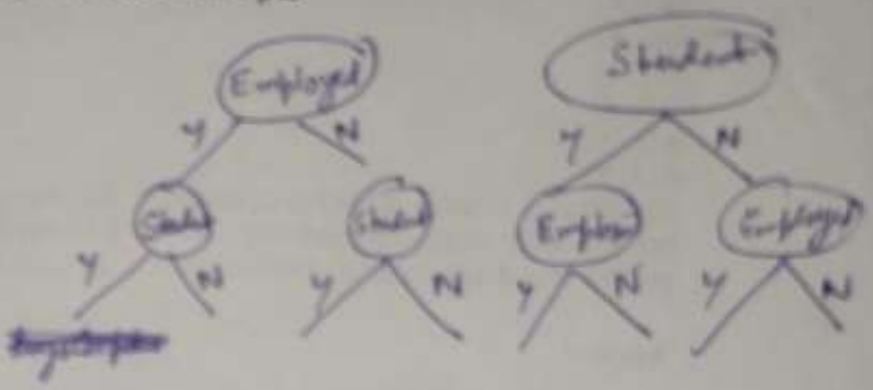
Spring 2009 – Assignment 2
Due Date: 4/20/09

10

Sandipan Dey

1. Consider the following data set for a binary class problem. **3 pts**

Employed	Student	buys computer
T	F	1
T	T	1
T	T	1
T	F	0
T	T	1
F	F	0
F	F	0
F	F	0
T	T	0
T	F	0



Calculate the information gain when splitting on "Employed" and "student". Which attribute would the decision tree induction algorithm choose?

Answer:

$$Info(D) = I(4,6) = -\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right) = 0.97$$

$$Info_{Employed}(D) = \frac{7}{10}I(4,3) + \frac{3}{10}I(3,0)$$

$$= \frac{7}{10}\left(-\frac{4}{7}.\log_2\left(\frac{4}{7}\right) - \frac{3}{7}.\log_2\left(\frac{3}{7}\right)\right) + \frac{3}{10}\left(-\frac{3}{3}.\log_2\left(\frac{3}{3}\right) - \frac{0}{3}.\log_2\left(\frac{0}{3}\right)\right)$$

$$= 0.7 \times 0.985228 + 0.3 \times 0 = 0.69$$

$$Gain_{Employed}(D) = Info(D) - Info_{Employed}(D) = 0.97 - 0.69 = 0.28$$

$$Info_{Student}(D) = \frac{4}{10}I(3,1) + \frac{6}{10}I(1,5)$$

$$= \frac{4}{10}\left(-\frac{3}{4}.\log_2\left(\frac{3}{4}\right) - \frac{1}{4}.\log_2\left(\frac{1}{4}\right)\right) + \frac{6}{10}\left(-\frac{1}{6}.\log_2\left(\frac{1}{6}\right) - \frac{5}{6}.\log_2\left(\frac{5}{6}\right)\right)$$

$$= 0.714525$$

$$Gain_{Student}(D) = Info(D) - Info_{Student}(D) = 0.97 - 0.71 = 0.26$$

We should look for highest gain, hence the decision tree algorithm should choose the attribute 'Employed'.

2. Clustering K-means example **7 pts**

IS 733 Data Warehousing and Mining
Spring 2009 – Assignment 2
Due Date: 4/20/09

Suppose that the data mining task is to cluster the following eight points (with (*x*; *y*) representing location) into three clusters.

$A_1(2; 10)$
$A_2(2; 5)$
$A_3(8; 4)$
$A_4(5; 6)$
$A_5(7; 5)$
$A_6(6; 4)$
$A_7(1; 4)$
$A_8(4; 8)$

The distance function is Euclidean distance. Suppose initially we assign A_1, A_2, and A_3 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only* the three cluster centers after the first round of execution and compute their SSE values.

Answer:

Point	Euclidian Distance			
	$A_1(2, 10)$	$A_2(2, 5)$	$A_3(8, 4)$	Cluste
$A_1(2, 10)$	$\sqrt{(2-2)^2+(10-10)^2}=0$	$\sqrt{(2-2)^2+(5-10)^2}=5$	$\sqrt{(8-2)^2+(4-10)^2}=8.49$	1
$A_2(2, 5)$	$\sqrt{(2-2)^2+(10-5)^2}=5$	$\sqrt{(2-2)^2+(5-5)^2}=0$	$\sqrt{(8-2)^2+(4-5)^2}=6.08$	2
$A_3(8, 4)$	$\sqrt{(2-8)^2+(10-4)^2}=8.49$	$\sqrt{(2-8)^2+(5-4)^2}=6.08$	$\sqrt{(8-8)^2+(4-4)^2}=0$	3
$A_4(5, 6)$	$\sqrt{(2-5)^2+(10-6)^2}=5$	$\sqrt{(2-5)^2+(5-6)^2}=3.16$	$\sqrt{(8-5)^2+(4-6)^2}=3.61$	2
$A_5(7, 5)$	$\sqrt{(2-7)^2+(10-5)^2}=7.07$	$\sqrt{(2-7)^2+(5-5)^2}=5$	$\sqrt{(8-7)^2+(4-5)^2}=1.41$	3
$A_6(6, 4)$	$\sqrt{(2-6)^2+(10-4)^2}=7.21$	$\sqrt{(2-6)^2+(5-6)^2}=4.12$	$\sqrt{(8-6)^2+(4-4)^2}=2$	3
$A_7(1, 4)$	$\sqrt{(2-1)^2+(10-4)^2}=6.08$	$\sqrt{(2-1)^2+(5-4)^2}=1.41$	$\sqrt{(8-1)^2+(4-4)^2}=7$	2
$A_8(4, 8)$	$\sqrt{(2-4)^2+(10-8)^2}=2.28$	$\sqrt{(2-4)^2+(5-8)^2}=3.61$	$\sqrt{(8-4)^2+(4-8)^2}=5.66$	1

Cluster	Point	Center
1	A_1, A_8	(3,9)
2	A_2, A_4, A_7	(2.67, 5)
3	A_3, A_5, A_6	(7,4.33)

$$SSE_{C1} = \sum_{x \in C1} d^2(m_i, x) = (2-3)^2 + (10-9)^2 + (4-3)^2 + (10-9)^2 = 1 + 1 + 1 + 1 = 4$$

$$SSE_{C2} = \sum_{x \in C2} d^2(m_i, x) = (2-2.67)^2 + (5-5)^2 + (5-2.67)^2 + (6-5)^2 + (1-2.67)^2 + (4-5)^2$$

$$= 0.45 + 0 + 5.43 + 1 + 2.79 + 1 = 10.67$$

$$SSE_{C3} = \sum_{x \in C3} d^2(m_i, x) = (8-7)^2 + (4-4.33)^2 + (7-7)^2 + (5-4.33)^2 + (6-7)^2 + (4-4.33)^2$$

$$= 1 + 0.11 + 0 + 0.45 + 1 + 0.11 = 2.67$$

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d^2(m_i, x) = \mathrm{SSE}_{C1} + \mathrm{SSE}_{C2} + \mathrm{SSE}_{C3} = 4 + 10.67 + 2.67 = 17.34$$

IS 733 Data Warehousing and Mining
Spring 2009 - Assignment 1
Due Date: 2/16/08

$$\text{Median} = Q_2 = \frac{13^{th}\,term + 14^{th}\,term}{2} = \frac{0.465125 + 0.474641}{2} = 0.469883$$

$Q_1 = 7^{th}\,term = 0.192108$

$Q_3 = 20^{th}\,term = 600.00$

$IQR = Q_3 - Q_1 = 600.000000 - 0.192108 = 599.917892$

$Q_3 + 1.5 \times IQR = 600 + 1.5 \times 599.917892 = 600 + 899.876838 = 1499.876838$

$Q_1 - 1.5 \times IQR = 0.192108 - 1.5 \times 599.917892 = 0.192108 - 899.876838 = -899.6884730$

Hence, the items falling outside the range $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ are 1600, 2100.The 2 outliers are 1600, 2100.

2.

Observed	Play chess	Not play chess	Total
Like science fiction	280	150	430
Not like science fiction	500	1200	1700
Total	780	1350	2130

Expected	Play chess	Not play chess
Like science fiction	$\frac{430 \times 780}{2130} = 157.465$	$\frac{430 \times 1350}{2130} = 272.535$
Not like science fiction	$\frac{1700 \times 780}{2130} = 622.535$	$\frac{1700 \times 780}{2130} = 1077.465$

Chi-Square $= \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$

$$= \frac{(280-157.465)^2}{157.465} + \frac{(150-272.535)^2}{272.535} + \frac{(500-622.535)^2}{622.535} + \frac{(1200-1077.465)^2}{1077.465}$$

$$= \frac{(122.535)^2}{157.465} + \frac{(-122.535)^2}{272.535} + \frac{(-122.535)^2}{622.535} + \frac{(122.535)^2}{1077.465}$$

$$= \frac{15014.826225}{157.465} + \frac{15014.826225}{272.535} + \frac{15014.826225}{622.535} + \frac{15014.826225}{1077.465}$$

$= 95.353 + 55.093 + 24.119 + 13.935$

$= 95.353 + 55.093 + 24.119 + 13.935$

$=188.5$