CSCI 620: INTRODUCTION TO BIG DATA

Activity 5 : Data Classification

Ameya Nagnur (an4920)
Siddarth Sargunaraj (sxs2469)
Ketan Kokane (kk7471)
Kavya Kotian (kk2014)

April 4, 2019

1.

a. What is the entropy of this collection of training examples with respect to the positive class

class	Count
+	4
-	5

$$Entropy = \frac{-4}{9}\log\frac{-4}{9} - \frac{5}{5}\log\frac{5}{9} = 0.99$$

b. What are the information gains of splitting on a1 and splitting on a2 relative to these training examples?

A1	True	False
+	3	1
-	1	4

$$A1_{TRUE} = \frac{-3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.81$$

$$A1_{FALSE} = \frac{-1}{5}\log\frac{-1}{5} - \frac{4}{5}\log\frac{4}{5} = 0.72$$

$$I.GA1 = 1 - \frac{4}{9} * A1_{TRUE} - \frac{5}{9} * A1_{FALSE} = 0.23$$

A2	True	False
+	2	2
-	3	2

$$A2_{TRUE} = \frac{-2}{5}\log\frac{-2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

$$A2_{FALSE} = \frac{-2}{4}\log\frac{-2}{4} - \frac{-2}{4}\log\frac{-2}{4} = 1$$

$I.GA2 = 1 - \frac{5}{9} * A2_{TRUE} - \frac{4}{9} * A2_{FALSE} = 0.016$

c. For a3, which is a continuous attribute, compute the information gainfor every possible split

a_3	Split Point	branch(pos,neg)	Entropy	IG
1.0	2.0	left(1,0) right(3,5)	0, 0.95	0.15
3.0	3.5	left(1,1) right(3,4)	1, 0.98	0.01
4.0	4.5	left(2,1) right(2,4)	0.91, 0.91	0.08
5.0	5.5	left(2,3) right(2,2)	0.97, 1	0.01
5.0	6.5	left(3,3) right(1,2)	1, 0.91	0.02
6.0	7.5	left(4,4) right(0,1)	1,0	0.1
7.0 7.0 8.0				

d. What is the best split (among a1, a2, and a3) according to the information gain?

According to the information gain, the best split would be with attribute a1 as it has highest information gain of 0.23.

e. What is the best split (between a1 and a2) according to the classification error rate

class	Count
+	4
-	5

$E_{Origin} = 1 - max(\frac{4}{9}, \frac{5}{9}) = \frac{4}{9}$

A1	True	False
+	3	1
-	1	4

$EA1_{True} = 1 - max(\frac{3}{4}, \frac{1}{4}) = \frac{1}{4}$

$EA1_{False} = 1 - max(\frac{1}{5}, \frac{4}{5}) = \frac{1}{5}$

$gainbyA1 = E_{Origin} - \frac{4}{9}EA1_{True}\frac{5}{9}EA1_{False} = \frac{2}{9}$

A2	True	False
+	2	2
-	3	2

$EA2_{True} = 1 - max(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}$

$EA2_{False} = 1 - max(\frac{2}{4}, \frac{2}{4}) = \frac{1}{2}$

$gainbyA2 = E_{Origin} - \frac{5}{9}EA2_{True}\frac{4}{9}EA2_{False} = 0$

Choose attribute A1 because it has the highest gain.