**Name:** Bala Guna Teja Karlapudi

**Student ID:** 801044799

**Email:** [bkarlapu@uncc.edu](mailto:bkarlapu@uncc.edu)

**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Trees**

**Part 2**

**Consider the training examples shown in below table for a binary classification**

**problem.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Instance | a1 | a2 | a3 | Target Class |
| 1 | T | T | 1 | + |
| 2 | T | T | 6 | + |
| 3 | T | F | 5 | − |
| 4 | F | F | 4 | + |
| 5 | F | T | 7 | − |
| 6 | F | T | 3 | − |
| 7 | F | F | 8 | − |
| 8 | T | F | 7 | + |
| 9 | F | T | 5 | − |

**(a) What is the entropy of this collection of training examples with respect**

**to the positive class?**

Class (+)=4

Class(-)=5

Total class=9

Entropy= -[(4/9)\*log2(4/9)] – [(5/9)\*log2(5/9)] = 0.9911

**(b) What are the information gains of *a*1 and *a*2 relative to these training**

**examples?**

For the attribute a1,

|  |  |  |
| --- | --- | --- |
| a1 | + | - |
| T | 3 | 1 |
| F | 1 | 4 |

Therefore, the entropy for a1 is

[4/9 (− (3/4)\*log2(3/4) − (1/4)\*log2(1/4) )] +[5/9 ( − (1/5)\*log2(1/5) − (4/5)\*log2(4/5) )]

= 0.7616

The information gain for a1= 0.9911 − 0.7616 = 0.2294

For the attribute a2,

|  |  |  |
| --- | --- | --- |
| a2 | + | - |
| T | 2 | 3 |
| F | 2 | 2 |

Therefore, the entropy for a2 is

[5/9 (− (2/5)\*log2(2/5) − (3/5)\*log2(3/5) )] +[4/9 ( − (2/4)\*log2(2/4) − (2/4)\*log2(2/4) )]

= 0.9839

The information gain for a2 =0.9911 − 0.9839 = 0.0072

**(c) For *a*3, which is a continuous attribute, compute the information gain**

**for every possible split.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | <=1 | >1 | <=3 | >3 | <=4 | >4 | <=5 | >5 | <=6 | >6 | <=7 | >7 | <=8 | >8 |
| CLASS (+) | 1 | 3 | 1 | 3 | 2 | 2 | 2 | 2 | 3 | 1 | 4 | 0 | 4 | 0 |
| CLASS  (-) | 0 | 5 | 1 | 4 | 1 | 4 | 3 | 2 | 3 | 2 | 4 | 1 | 5 | 0 |
| Total | 1 | 8 | 2 | 7 | 3 | 6 | 5 | 4 | 6 | 3 | 8 | 1 | 9 | 0 |

For, I(<=1.0)=-((1/1)\*LOG2(1/1))-((0/1)\*log2(0/1))=0

I(>1.0)=-((3/8)\*log2(3/8))-((5/8)\*log2(5/8))=0.95444

Information gain = .991 - (1/9)\*0 - (8/9)\*.9544 = 0.1427

Next split,

I(<=3.0)=-((1/2)\*log2(1/2))-((1/2)\*log2(1/2))=1

I(>3.0)=-((3/7)\*log2(3/7))-((4/7)\*log2(4/7))=0.9852

Information gain = .991 - (2/9)\*1 - (7/9)\*.9852 = 0.0026

Next split,

I(<=4.0)=-((2/3)\*log2(2/3))-((1/3)\*log2(1/3))=0.9183

I(>4.0)=-((2/6)\*log2(2/6))-((2/6)\*log2(2/6))=0.9183

Information gain=0.991-(3/9)\*0.9183-(6/9)\*0.9183=0.728

For next split,

I(<=5.0)=-((2/5)\*log2(2/5))-((3/5)\*log2(3/5))=0.971

I(>5.0)=-((2/4)\*log2(2/4))-((2/4)\*log2(2/4))=1

Information gain= .991 - (5/9)\*.971 - (4/9)\*1 = 0.0072

Next split,

I(<=6.0)=-((3/6)\*log2(3/6))-((3/6)\*log2(3/6))=1

I(>6.0)=-((1/3)\*log2(1/3))-((2/3)\*log2(2/3))=0.9183

Information gain=0.991-(6/9)\*1-(3/9)\*0.39=0.0183

Next split,

I(<=7.0)=-((4/8)\*log2(4/8))-((4/8)\*log2(4/8))=1

I(>7.0)=-((0/1)\*log2(0/1))-((1/1)\*log2(1/1))=0  
Information gain = .991 - (8/9)\*1 - (1/9)\*0 = 0.1021

Next,

I(<=8.0)=-((4/9)\*log2(4/9))-((5/9)\*log2(5/9))=0.9911

I(>8.0)=0

Information gain= .991 - (9/9)\*1 - (0/9)\*0= 0

**(d) What is the best split (among *a*1, *a*2, and *a*3) according to the information**

**gain?**

The best split among a1,a2,a3 is a1 because the difference in entropy is large which is 0.11427

**(e) What is the best split (between *a*1 and *a*2) according to the classification**

**error rate?**

For a1, according to classification error rate,

CE(T) = 1 - max[p(T+|Total), p(T-|Total)] = 1 - max [3/4, 1/4] = 1 - 3/4 = 0.25

CE(F) = 1 - max[p(F+|Total), p(F-|Total)] = 1 - max [2/4, 2/4] = 1 - 4/5 = 0.2

Total CE(a1) = [Total CE(T)/Total(a1)] \*CE(T) + [Total CE(F)/Total(a1)] \*CE(F)=

(4/9) \*0.25 + (5/9) \*0.48 = 0.2222

For a2,

CE(T) = 1 - max[p(T+|Total), p(T-|Total)] = 1 - max [2/5, 3/5] = 1 - 3/5 = 0.4

CE(F) = 1 - max[p(F+|Total), p(F-|Total)] = 1 - max [1/5, 4/5] = 1 - 1/2= 0.5

Total CE(a2) = [Total CE(T)/Total(a2)] \*CE(T) + [Total CE(F)/Total(a2)] \*CE(F) =

(5/9) \*0.4 + (4/9) \*0.5 = 0.4444

The best split is a1 since it has the lowest classification error rate.

**(f)** **What is the best split (between *a*1 and *a*2) according to the Gini index?**

For a1, the Gini Index considering T and F is

Gini(T) = 1 – (p(+|T)^2) – (p(-|T)^2) = 1 – ((3/4)^2)- ((1/4)^2) = 0.375

Gini(F) = 1 – p((+|F)^2) – (p(-|F)^2) = 1 - ((1/5)^2) – ((4/5)^2) = 0.32

Gini(a1) = [[(Total(T)/Total(a1)] \*Gini(T) ]+ [[(Total(F)/Total(a1)] \*Gini(F]])

= ((4/9) \*0.375) + ((5/9) \*0.32) = 0.3444

For a2,

Gini(T) = 1 – p((+|T)^2) – p((-|T)^2) = 1 –( (2/5)^2) – ((3/5)^2) = 0.48

Gini(F) = 1 – p((+|F)^2) – p((-|F)^2) = 1 – ((2/4)^2) – ((2/4)^2) = 0.5

Gini(a2) = [[(Total(T)/Total(a2)] \*Gini(T)] + [[(Total(F)/Total(a2)] \*Gini(F)] =( (

5/9) \*0.48) + ((4/9) \*0.5) = 0.4889

The best split is A1 as the subsets for a1 attribute have smaller Gini index.