**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Trees**

**Part 2**

**NAME: Ashish Bisht**

**Student ID: 801168390**

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

**problem.**

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

**to the positive class?**

Ans:

Class+ = 4  
Class-  = 5  
Total = 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?**

Ans:

Delta = I(Parent) - SUM[N(vj)/N]Iv(j)  
  
Using a1 for the split gives l(a|t) with 3 + and 1 - and l(a|f) with 1 + and 4 -  
  
l(a|t) = -(3/4)\*log2(3/4) - (1/4)\*log2(1/4)

= 0.8113   
  
l(a|f) = -(1/5)\*log2(1/5) - (4/5)\*log2(4/5)

= 0.7219

>

Delta(a1) = 0.9911 - (4/9)\*0.8113 - (5/9)\*0.7129 = 0.2345   
  
Using a2 for the split gives l(a|t) with 2 + and 3 -, and l(a|f) with 2 + and 2 -  
  
l(a|t) = -(2/5)\*log2(2/5) - (3/5)\*log<2(3/5)

= 0.971  
  
l(a|f) = -(2/4)\*log2(2/4) - (2/4)\*log2(2/4)

= 1

Delta(a2) = 0.9911 - (5/9)\*0.971 - (4/9)\*1

= 0.007211

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

**for every possible split.**

Ans:

  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

Total = 9  
l(T) = .9911  
  
l(<=1.0) = -(1/1)log2(1/1) - (0/1)log2(0/1) = 0  
l(>1.0) = -(3/8)log2(3/8) - (5/8)log2(5/8) = .95444  
Delta(1.0) = .991 - (1/9)\*0 - (8/9)\*.9544 = .1427  
  
  
l(<=3.0) = -(1/2)log2(1/2) - (1/2)log2(1/2) = 1  
l(>3.0) = -(3/7)log2(3/7) - (4/7)log2(4/7) = .9852  
Delta(3.0) = .991 - (2/9)\*1 - (7/9)\*.9852 = .0026  
  
ll(<=4.0) = -(2/3)log2(2/3) - (1/3)log2(1/3) = .9183  
l(>4.0) = -(2/6)log2(2/6) - (2/6)log2(2/6) = .9183  
Delta(4.0) = .991 - (3/9)\*.9183 - (6/9)\*.9183 = .0728  
  
  
l(<=5.0) = -(2/5)log2(2/5) - (3/5)log2(3/5) = .971  
l(>5.0) = -(2/4)log2(2/4) - (2/4)log2(2/4) = 1  
Delta(5.0) = .991 - (5/9)\*.971 - (4/9)\*1 = .0072  
  
  
l(<=6.0) = -(3/6)log2(3/6) - (3/6)log2(3/6) = 1  
l(>6.0) = -(1/3)log2(1/3) - (2/3)log2(2/3) =  .9183  
Delta(6.0) = .991 - (6/9)\*1 - (3/9)\*.39 = .0183  
  
l(<=7.0) = -(4/8)log2(4/8) - (4/8)log2(4/8) = 1  
>l(>7.0) = -(0/1)log2(0/1) - (1/1)log2(1/1) = 0  
Delta(7.0) = .991 - (8/9)\*1 - (1/9)\*0 = .1021  
  
l(<=8.0)= -(4/9)log2(4/9) - (5/9)log2(5/9) = .9911  
l(>8.0)= -(0/0)log2(0/0) - (0/0)log2(0/0) = 0  
Delta(8.0) = .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?**

Ans: The best split(among *a*1, *a*2, and *a*3) as stated by the information

gain is using attribute a1 because it has the largest delta difference in entropy with 0.11427.

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

**error rate?**

Ans:

The best split (between *a*1 and *a*2) according to the classification

error rate is,

a1 -

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

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

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

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 as it has the lower classification error. < /P > < /P >

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

−

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

Ans:

The best split (between *a*1 and *a*2) according to the Gini index is,

a1-

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

TGini(a1) = [(Total(T)/Total(a1)]\*Gini(T) + [(Total(F)/Total(a1)]\*Gini(F) =(4/9)\*0.375 + (5/9)\*0.32 = 0.3444

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

TGini(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 because the subsets for attribute a1 have a smaller Gini index.