**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Trees**

**Part 2**

**Name: Balasundaram Avudai Nayagam**

**Student ID: 801104903**

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?

A) Entropy is given as,

Entropy(t) = -∑p(c)log2p(c)

E(T) = -(4/9) \* log2(4/9) – (5/9) \* log2(5/9)

= 1.184\*0.44 + 0.862\*0.55

= 0.991

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

examples?

A) Entropy and gain formulae are given below:

I= -∑p(c)log2p(c)

Ires= -∑p(v) ∑p(c/v) log2 p(c/v)

Gain= I - Ires.

We know from the first problem entropy of the collection is, I = 0.991.

Ires(a1) = [(4/9) \* [-(3/4) log2(3/4) - (1/4) log2(1/4)]] + [(5/9) \* [-(1/5) log2 (1/5) -(4/5) log2(4/5)]]

= [(0.44 \* (0.311 + 0.5)] + [(0.55 \* (0.464 + 0.258))]

= 0.44 \* 0.8113 + 0.55 \* 0.7219

Ires(a1) = 0.7608

Gain (a1) = 0.9911 – 0.7608

= 0.23

Ires(a2) = [(5/9) \* [-(2/5) log2(2/5) - (3/5) log2(3/5)]] + [(4/9) \* [-(2/4) log2 (2/4) -(2/4) log2(2/4)]]

= [0.55 \* 0.971] + [0.44 \* 1]

Ires(a2) = 0.9829

Gain(a2) = 0.9911 – 0.9829

= 0.007

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

for every possible split.

A)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Spotted values | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | |
| Split position | 0.5 | | 2 | | 3.5 | | 4.5 | | 5.5 | | 6.5 | | 7.5 | | 8.5 | |
|  | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > |
| + | 0 | 4 | 1 | 3 | 1 | 3 | 2 | 2 | 2 | 2 | 3 | 1 | 4 | 0 | 4 | 0 |
| - | 0 | 5 | 0 | 5 | 1 | 4 | 1 | 4 | 3 | 2 | 3 | 2 | 4 | 1 | 5 | 0 |

**Split 1:**

I (<=0.5) = 0

I (>0.5) = -(4/9) log2(4/9) - (5/9) log2(5/9)

= 0.9911

Ires (0.5) = 0.9911

Gain= 0

**Split 2:**

I (<=2) = 0

I (>2) = -(3/8) log2(3/8) - (5/8) log2(5/8)

= 0.5306 + 0.4237

= 0.9543

Ires (2) = (1/9) \*0 + (8/9) \*0.9543 = 0.8483

Gain (2) = 0.1428

**Split 3:**

I (<=3.5) = -(1/2) log2(1/2) - (1/2) log2(1/2) = 1

I (>3.5) = -(3/7) log2(3/7) - (4/7) log2(4/7) = 0.5240 + 0.4613 = 0.9853

Ires(3.5) = (2/9) \* 1 + (7/9) \* 0.9853 = 0.988

Gain (3.5) = 0.003

**Split 4:**

I (<= 4.5) = -(2/3) log2(2/3) - (1/3) log2(1/3) = 0.3956 + 0.5278

I (<= 4.5) = 0.9234

I (> 4.5) = -(2/6) log2(2/6) - (4/6) log2(4/6) = 0.5278 + 0.3956

I (>4.5) = 0.9234

Ires(4.5) = (3/9) \* 0.9234 + (6/9) \* 0.9234 = 0.92

Gain (4.5) = 0.9911 – 0.92 = 0.07

**Split 5:**

I (<=5.5) = -(2/5) log2(2/5) - (3/5) log2(3/5) = 0.5287 + 0.4421 =0.9709

I (>5.5) = -(2/4) log2(2/4) - (2/4) log2(2/4) = 1

Ires(5.5)= (5/9) \* 0.9709 + (4/9) \* 1= 0.9838

Gain (5.5) = 0.007

**Split 6:**

I (<=6.5) = -(3/6) log2(3/6) - (3/6) log2(3/6) = 1

I (>6.5) = -(1/3) log2(1/3) - (2/3) log2(2/3) = 0.9234

Ires(6.5)= (6/9) \* 1 + (3/9) \* 0.9234 = 0.9727

Gain (6.5) = 0.02

**Split 7:**

I (<=7.5) = -(4/8) log2(4/8) - (4/8) log2(4/8) = 1

I (>7.5) = 0

Ires(7.5)= (8/9) = 0.8889

Gain (7.5) = 0.102

**Split 8:**

I (<=8.5) = -(4/9) log2(4/9) - (5/9) log2(5/9)

= 0.521136 + 0.47432 = 0.9910

I (>8.5) =0

Ires(8.5)= 0.9954

Gain (8.5) = 0

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

gain?

A) The best split among *a*1, *a*2, and *a*3 based on the information gain would be at a1, because the gain is higher than a2 and a3.

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

error rate?

A) Classification error rate is given as **1- max[p(i/t)]**

**For a1:**

Classification error (T) = 1- max[(1/4), (3/4)] = 1-(3/4) = 0.25

Classification error (F) = 1- max[(1/5),(4/5)] = 1- (4/5) = 0.20

Classification error (a1) = (4/9) \* 0.25 + (5/9) \* 0.2 = 0.111 + 0.111= 0.222.

**For a2:**

Classification error (T) = 1- max [(2/5), (3/5)] = 1- 3/5 =0.4

Classification error (F) = 1- max [(2/4), (2/4)] = 0.5

Classification error (a2) = (5/9) \* 0.4 + (4/9) \* 0.5 = 0.222 + 0.222 = 0.444

**a1** is the best split because of its lowest error value.

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

A )Gini index is given by, Gini(t)= 1-∑[p(i/t)]2.

**For a1:**

Gini index (T) = 1- [(1/4)2 + (3/4)2] = 1- [0.0625 + 0.5625] = 0.375

Gini index (F) = 1- [(1/5)2 + (4/5)2] = 1- [0.04 + 0.64] = 0.320

Gini index (a1) = (4/9) \* 0.375 + (5/9) \* 0.320= 0.1665 + 0.1776 = 0.3441

**For a2:**

Gini index (T) = 1- [(2/5)2 + (3/5)2] = 1- [0.16 + 0.36] = 0.48

Gini index (F) = 1- [(2/4)2 + (2/4)2] = 0.5

Gini index (a2) = (5/9) \* 0.48 + (4/9) \* 0.5 = 0.266 + 0.222 = 0.488

**a1** is the best split because of its lowest Gini index.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 | − |