**Name: Nikhil Jadhav (801075504)**

**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Trees**

**Part 2**

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?

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

examples?

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

for every possible split.

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

gain?

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

error rate?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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?

Solution:

There are four positive examples and five negative examples. The Entropy is calculated using

Entropy =

Therefore,

Therefore, the entropy for the given example is as follows:

1. For attribute ‘a1’ –

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

Using probability for + and – values found in previous example i.e.

The entropy for a1 is –

Therefore, the information gain for a1 is - .

For attribute ‘a2’ –

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

The entropy for a2 is –

Therefore, the information gain for a2 is – .

1. Computing information gain for every possible split for continues attribute a3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A3 | Class Label | Split Point | Entropy | Information Gain |
| 1.0 | + | 2.0 | 0.84 | 0.1427 |
| 3.0 | - | 3.5 | 0.98 | 0.0026 |
| 4.0 | + | 4.5 | 0.91 | 0.0728 |
| 5.0  5.0 | -  - | 5.5 | 0.98 | 0.0072 |
| 6.0 | + | 6.5 | 0.97 | 0.0183 |
| 7.0  7.0 | +  - | 7.5 | 0.88 | 0.1022 |

We have maximum information gain for split point 2. Thus, we can see that the best split occurs at split point = 2.

1. According to information gain calculated in previous examples we can conclude that a1 has the highest information gain, which in turn produces the best split
2. For attribute a1: error rate = 2/9.

For attribute a2: error rate = 4/9.

Therefore, according to error rate, a1 produces the best split.

1. For a1, the Gini Index is -

For a2, the Gini Index is –

Since the Gini Index for A1 is lesser than that of A2, A1 will produce a better split.