**Cloud Computing for Data Analysis**

**Exercise 09 : Decision Trees**

**Part 2**

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

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | |  | Count | | Class + | 4 | | Class - | 5 | | Total | 9 |   Entropy = - (4/9)log2(4/9) - (5/9)log2(5/9) = 0.99107 |

1. What are the information gains of *a*1 and *a*2 relative to these training examples?

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | a1 | T | F | Total | | Class + | 3 | 1 | 4 | | Class - | 1 | 4 | 5 | | Total | 4 | 5 | 9 |   Entropy(T) = - (3/4)log2(3/4) – (1/4)log2(1/4) = 0.81128  Entropy(F) = - (1/5)log2(1/5) – (4/5)log2(4/5) = 0.72193  Information gain of a1 = 0.99107 – (4/9) \* 0.81128 – (5/9) \* 0.72193 = 0.229   |  |  |  |  | | --- | --- | --- | --- | | a2 | T | F | Total | | Class + | 2 | 2 | 4 | | Class - | 3 | 2 | 5 | | Total | 5 | 4 | 9 |   Entropy(T) = - (2/5)log2(2/5) – (3/5)log2(3/5) = 0.97095  Entropy(F) = - (2/4)log2(2/4) – (2/4)log2(2/4) = 1.0  Information gain of a2 = 0.99107 – (5/9) \* 0.97095 – (4/9) \* 1.0 = 0.007 |

1. For *a*3, which is a continuous attribute, compute the information gain for every possible split.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Values | 1 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | |  | | | Split Point | 0.5 | | 2.0 | | 3.5 | | 4.5 | | 5.5 | | 6.5 | | 7.5 | | 8.5 | | | Operator | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | | + | 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 | | Gain | 0 | | 0.14269 | | 0.002558 | | 0.07277 | | 0.00721 | | 0.018305 | | 0.10217 | | 0 | |   Split point: 0.5  Entropy(>0.5) = - (4/9)log2(4/9) - (5/9)log2(5/9) = 0.99107  Information gain = 0.99107 - 0.99107 = 0  Split point: 2.0  Entropy(<=2.0) = - (1/1)log2(1/1) – (0)\*log2(0) = 0  Entropy(>2.0) = - (3/8)log2(3/8) – (5/8)log2(5/8) = 0.95443  Weighted average of entropy = (1/9)\*0 + (8/9)\* 0.95443 = 0.84838  Information gain = 0.99107 - 0.84838 = 0.14269  Split point: 3.5  Entropy(<=3.5) = - (1/2)log2(1/2) – (1/2)\*log2(1/2) = 1.0  Entropy(>3.5) = - (3/7)log2(3/7) – (4/7)\*log2(4/7) = 0.98523  Weighted average of entropy = (2/9)\*1.0 + (7/9)\*0.98523 = 0.988512  Information gain = 0.99107 - 0.988512 = 0.002558  Split point: 4.5  Entropy(<=4.5) = - (2/3)log2(2/3) – (1/3)\*log2(1/3) = 0.9183  Entropy(>4.5) = - (2/6)log2(2/6) – (4/6)\*log2(4/6) = 0.9183  Weighted average of entropy = (3/9)\*0.9183 + (6/9)\*0.9183 = 0.9183  Information gain = 0.99107 - 0.9183 = 0.07277  Split point: 5.5  Entropy(<=5.5) = - (2/5)log2(2/5) – (3/5)\*log2(3/5) = 0.97095  Entropy(>5.5) = - (2/4)log2(2/4) – (2/4)\*log2(2/4) = 1.0  Weighted average of entropy = (5/9)\*0.97095 + (4/9)\*1.0 = 0.98386  Information gain = 0.99107 - 0.98386 = 0.00721  Split point: 6.5  Entropy(<=6.5) = - (3/6)log2(3/6) – (3/6)\*log2(3/6) = 1.0  Entropy(>6.5) = - (1/3)log2(1/3) – (2/3)\*log2(2/3) = 0.9183  Weighted average of entropy = (6/9)\*1.0 + (3/9)\* 0.9183 = 0.972765  Information gain = 0.99107 - 0.972765 = 0.018305  Split point: 7.5  Entropy(<=7.5) = - (4/8)log2(4/8) – (4/8)\*log2(4/8) = 1.0  Entropy(>7.5) = - (0/1)log2(0/1) – (1/1)\*log2(1/1) = 0  Weighted average of entropy = (8/9)\*1.0+ (1/9)\*0 = 0.8889  Information gain = 0.99107 - 0.8889 = 0.10217  Split point: 8.5  Entropy(<=8.5) = - (4/9)log2(4/9) - (5/9)log2(5/9) = 0.99107  Information gain = 0.99107 - 0.99107 = 0 |

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

|  |
| --- |
| According to the information gain, **a1** is the best split as it achieves the highest gain (i.e. 0.229) |

1. What is the best split (between *a*1 and *a*2) according to the classification error rate?

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | a1 | T | F | Total | | Class + | 3 | 1 | 4 | | Class - | 1 | 4 | 5 | | Total | 4 | 5 | 9 |   Error(T): 1 – max(3/4, 1/4) = ¼ = 0.25  Error(F): 1 – max(1/5, 4/5) = 1/5 = 0.20  Weighted average of error in a1 = (4/9 \* 0.25) + (5/9 \* 0.20) = 0.22   |  |  |  |  | | --- | --- | --- | --- | | a2 | T | F | Total | | Class + | 2 | 2 | 4 | | Class - | 3 | 2 | 5 | | Total | 5 | 4 | 9 |   Error(T): 1 – max(2/5, 3/5) = 2/5 = 0.40  Error(F): 1 – max(2/4, 2/4) = 1/2 = 0.50  Weighted average of error in a2 = (5/9 \* 0.40) + (4/9 \* 0.50) = 0.44  Hence, the best split is **a1** as it has the lower classification error rate. |

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | a1 | T | F | Total | | Class + | 3 | 1 | 4 | | Class - | 1 | 4 | 5 | | Total | 4 | 5 | 9 |   Gini(T): 1 – (3/4)^2 – (1/4)^2 = 0.375  Gini(F): 1 – (1/5)^2 – (4/5)^2 = 0.32  Weighted average of Gini in a1: (4/9) \* 0.375 + (5/9) \* 0.32 = 0.3467   |  |  |  |  | | --- | --- | --- | --- | | a2 | T | F | Total | | Class + | 2 | 2 | 4 | | Class - | 3 | 2 | 5 | | Total | 5 | 4 | 9 |   Gini(T): 1 – (2/5)^2 – (3/5)^2 = 0.48  Gini(F): 1 – (2/4)^2 – (2/4)^2 = 0.50  Weighted average of Gini in a1: (5/9) \* 0.48 + (4/9) \* 0.50 = 0.4867  Hence, the best split is **a1** as it has the lower Gini index (i.e. 0.3467) |

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