1. Chapter 3, Exercise 3

|  |  |  |  |
| --- | --- | --- | --- |
| A1 | A2 | A3 | Target Class |
| T | T | 1 | + |
| T | T | 6 | + |
| T | F | 5 | - |
| F | F | 4 | + |
| F | T | 7 | - |
| F | T | 3 | - |
| F | F | 8 | - |
| T | F | 7 | + |
| F | T | 5 | - |

* 1. What is the entropy of this collection of training examples with respect to the class attribute?  
     = - ((4/9) \* log2(4/9)) – (5/9 \* log2(5/9)) = 0.99
  2. What are the information gains of a1 and a2 relative to these training examples?

I(A1=T) = (4/9) \* 0.811 + (5/9) \* .72198 = .761  
Gain = .991 - .761 =**0.2293**  
  
Gain for A2: 0.991 – [ (4/9) \* 1 + (5/9) \* 0.970591 = 0.9838] **= 0.007139**

* 1. For a3, a continuous attribute, compute the information gain for every possible split.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | |  | |  |
| 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 | - |
| 0.991 | | 0.8483 | | 0.9885 | | 0.918 | | 0.9838 | | 0.97276 | | 0.8888 | | 0.9991 | | Entropy |
| 0 | | .143 | | .0025 | | .0727 | | .0071 | | .0182 | | .10211 | | 0 | | GAIN |

* 1. What is the best split among a1-a3 according to the information gain?

A1 has the highest possible gain – and is therefore the best split.

* 1. What is the best split between a1 and a2 according to the misclassification error rate?

A1 is still optimal as it’s classification rate is 1 – max((2/9),(7/9)) = 0.22.

* 1. What is the best split between a1 and a2 according to the gini index?

**The gini index is calculated here as 1 – [(i/n)^2 + (j/n)^2]**

Gini for A1: (A1=T) \* (4/9) + (A1=F) \* (5/9) = ~0.34

Gini for A2: (A2=T) \* (5/9) + (A2=F) \* (4/9) = ~0.49

By this metric, the best split is by a1 since it has a lower gini index.