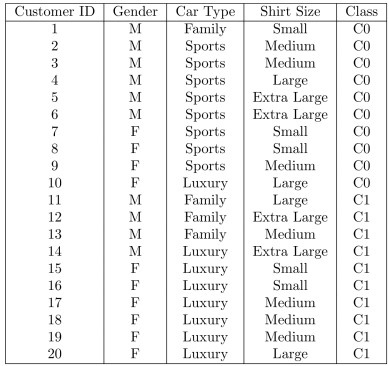
1. Consider the training examples shown in Table for a binary classification problem. (Treat the “Class” field as a target.)

(a) Compute the Gini index for the overall collection of training examples.

C1 = 10個 ,C0 = 10個 => P1 = 10/20, P2 = 10/20

Gini index = 1 – ΣP2k

=1 – [P12+P22+P23+…]

= 1 – [(10/20)2 + (10/20)2]

= 1 – 1/2

= 0.5

=> gini index of overall collection of training examples of C0 & C1 is 0.5

(b) Compute the Gini index for the Car Type attribute.

三種: Family or Sports or Luxury

Family:

C0 = 1/4, C1 = 3/4

1 – [P12 + P22]

= 1 – [(1/4)2 + (3/4)2]

= 1 – (10/16)

= 0.375

Sports:

C0 = 8/8, C1 = 0/8

1 – [P12 + P22]

= 1 – [(8/8)2 + (0/8)2]

= 1 – 1

= 0

Luxury:

C0 = 1/8, C1 = 7/8

1 – [P12 + P22]

= 1 – [(1/8)2+(7/8)2]

= 1 – (50/64)

= 0.21875

=> Total:4/20 x 0.375 + 8/20 x 0 + 8/20 x 0.21875 = 3.25/20 = 0.1625

(c) Compute the Gini index for the Gender attribute.

兩種:男or女

女(Female):

C0 = 4/10, C1 = 6/10

1 – [P12 + P22]

= 1 - [(4/10)2 + (6/10)2]

= 1 – [16/100 + 36/100]

= 1 – 52/100

= 0.48

男(Male):

C0 = 6, C1 = 4

1 – [P12 + P22]

= 1 - [(4/10)2 + (6/10)2]

= 1 – [16/100 + 36/100]

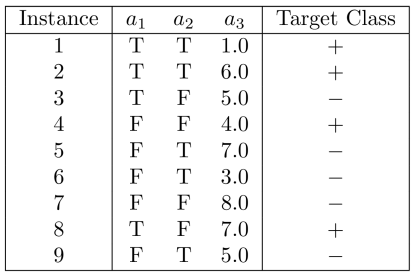
= 1 – 52/100

= 0.48

=> Total: 0.48 x 1/2 + 0.48 x 1/2 = 0.48

=>gini index og Gender attribute is 0.48

2. Consider the training examples shown in Table for a binary classification problem.

(a) What is the entropy of this collection of training examples with respect to the positive class?

Ans:

There are four positive examples andfivenegative examples.Thus, P(+)=4/9andP(−)=5/9.Theentropyofthetrainingexamples is −4/9log2(4/9)−5/9log2(5/9)=0.9911.

(b)What is the best split (among a1, a2, and a3) according to the information gain?

a1 :

entropy = 4/9[-(3/4)log2(3/4) – (1/4)log2(1/4)] + 5/9[-(1/5)log2(1/5) – (4/5)log2(4/5)] = 0.7616

* Information gain = 0.9911 – 0.7616 = 0.2294

a2 :

entropy = 5/9[-(2/5)log2(2/5) – (3/5)log2(3/5)] + 4/9[-(2/4)log2(2/4) – (2/4)log2(2/4)] = 0.9839

* Information gain = 0.9911 – 0.9839 = 0.0072

a3:

case1: a3 = 1.0

split point = 2.0, entropy = 0.8484, information gain = 0.1427

case2: a3 = 3.0

split point = 3.5, entropy = 0.9885, information gain = 0.0026

case3: a3 = 4.0

split point = 4.5, entropy = 0.9183, information gain = 0.0728

case4: a3 = 5.0

split point = 5.5, entropy = 0.9839, information gain = 0.0072

case5: a3 = 6.0

split point = 6.5, entropy = 0.9728, information gain = 0.0183

case6: a3 = 7.0

split point = 7.5, entropy = 0.8889, information gain = 0.1022

=> a1 has best split

(c) According to “Target Class”, use gain ratio to decide which feature should be used to split the first level of the decision tree.

(Calculate to the third decimal place, and then round off

unconditionally.)

原始:

I(p,n) = -4/9 log2 4/9 – 5/9 log2 5/9 = 0.9911

a1:

gains = 0.2294

split gain = -4/9 log2 4/9 – 5/9 log2 5/9 = 0.9911

gain ratio = 0.2294/0.9911 = 0.231 (最大)

a2:

gains = 0.0072

split gain = -5/9 log2 5/9 – 4/9 log2 4/9 = 0.9911

gain ratio = 0.0072/0.9911 = 0.007

a3:

gains = 0.1427

split gain =5(-1/9 log2 1/9) – 2(2/9 log2 2/9) = 2.72548

gain ratio = 0.1427/2.72548 = 0.052

