CZ4041: Tutorial Week 5

Due on February 11, 2021 at 8:30am

Assoc Prof Pan, Sinno Jialin - CS4

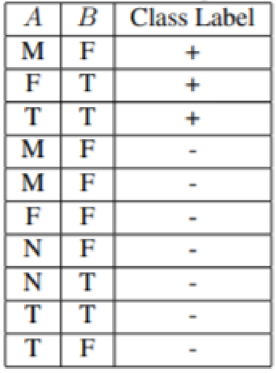
Pang Yu Shao
U1721680D

11/02/2021

Problem 1

Consider the data set shown in Table 1 for a binary classification problem

Table 1: Data set for Question 1.



A	B	Class Label
M	**F**	+
F	**T**	+
T	**T**	+
M	**F**	-
M	**F**	-
F	**F**	-
N	**F**	-
N	**T**	-
T	**T**	-
T	**F**	-

1. Calculate the information gain when splitting on A and B (using multi-way split on A). Which feature would the decision tree induction algorithm choose?
2. Calculate the gain ratio when splitting on A and B (using multi-way split on A). Which feature would the decision tree induction algorithm choose?

Solution

1. Information Gain: $\Delta_{info} = Entropy(parent\ node) - Entropy(children\ nodes)$

$$\begin{aligned} E_p &= -(3/10)log_2(3/10) - (7/10)log_2(7/10) \\ &= 0.88129 \end{aligned}$$

$$\begin{aligned} E_A &= (3/10)(-(1/3)log_2(1/3) - (2/3)log_2(2/3)) + (2/10)(1) + (3/10)(-(1/3)log_2(1/3) - (2/3)log_2(2/3)) + (2/10)(0) \\ &= 0.75097 \end{aligned}$$

$$\begin{aligned} E_B &= (4/10)(1) + (6/10)(-(1/6)log_2(1/6) - (5/6)log_2(5/6)) \\ &= 0.79001 \end{aligned}$$

$$\begin{aligned} \Delta_{info(A)} &= 0.88129 - 0.75097 \\ &= \mathbf{0.13032} \\ \Delta_{info(B)} &= 0.88129 - 0.79001 \\ &= 0.09128 \end{aligned}$$

Therefore, the decision tree induction algorithm would choose to split on **A**.

Problem 1 continued on next page. . .

2.
Gain Ratio: $\Delta_{InfoR} = \frac{\Delta_{info}}{SplitINFO}$

Where $SplitINFO = -\sum_{i=1}^{p} \frac{n_i}{n} log_2(\frac{n_i}{n})$

$$\begin{aligned}
SplitINFO_A &= -((3/10)log_2(3/10) + (3/10)log_2(3/10) + (2/10)log_2(2/10) + (2/10)log_2(2/10)) \\
&= -((6/10)log_2(3/10) + (4/10)log_2(2/10)) \\
&= 1.97095
\end{aligned}$$

$$\begin{aligned}
SplitINFO_B &= -((4/10)log_2(4/10) + (6/10)log_2(6/10)) \\
&= 0.97095
\end{aligned}$$

$$\begin{aligned}
\Delta_{infoR(A)} &= \Delta_{info(A)}/SplitINFO_A \\
&= 0.13032/1.97095 \\
&= 0.06612 \\
\Delta_{infoR(B)} &= \Delta_{info(B)}/SplitINFO_B \\
&= 0.09128/0.97095 \\
&= \mathbf{0.09401}
\end{aligned}$$

Therefore, the decision tree induction algorithm would choose to split on **B**.