2020SM2 Workshop Week 9 Exercise 1

Comp20008

1- Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

- Write a formula for the information gain when splitting on feature A.

- Contingency Table after splitting on feature A

	A = T	A = F
+	4	0
-	3	3

- The overall entropy before splitting :

$$E_{Orig} = -\,0.4 \log 0.4 \;-0.6 \log 0.6 = 0.9710$$

- The information gain after splitting on A is:

$$E_{A=T} = -\frac{4}{7}\log\frac{4}{7} - \frac{3}{7}\log\frac{3}{7} = 0.9852$$

$$E_{A=F} = -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3} = 0$$

$$\Delta = E_{Orig} - \frac{7}{10}E_{A=T} - \frac{3}{10}E_{A=F} = 0.2813$$

1- Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

- Write a formula for the information gain when splitting on feature B.

- Contingency Table after splitting on feature B

	B = T	B = F
+	?	?
-	?	?

- The overall entropy before splitting :

$$E_{Orig} = -\,0.4 \log 0.4 \; - 0.6 \log 0.6 = 0.9710$$

- The information gain after splitting on B is:

$$E_{B=T} = -\,?\log ? - ?\log ? = ?$$

$$E_{B=F} = -\,?\log ? - ?\log ? = ?$$

$$\Delta = \; E_{Orig} - ?\,E_{B=T} - ?\,E_{B=F} \; = ?$$

1- Consider the following data set for a binary class problem and consider building a decision tree using this data.

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

$$\Delta = E_{Orig} - \frac{4}{10} E_{B=T} - \frac{6}{10} E_{B=F} = 0.2565$$

- Write a formula for the information gain when splitting on feature B.

- Contingency Table after splitting on feature B

	B = T	B = F
+	3	1
-	1	5

- The overall entropy before splitting :

$$E_{Orig} = -0.4 \log 0.4 \ -0.6 \log 0.6 = 0.9710$$

- The information gain after splitting on B is:

$$E_{B=T} = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113$$

$$E_{B=F} = -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} = 0.6500$$

- The information gain after splitting on A is:

$$\Delta = E_{Orig} - \frac{7}{10} E_{A=T} - \frac{3}{10} E_{A=F} = 0.2813$$

- The information gain after splitting on B is:

$$\Delta = E_{Orig} - \frac{4}{10} E_{B=T} - \frac{6}{10} E_{B=F} = 0.2565$$

- Therefore attribute ? will be chosen to split the node

- Therefore attribute A will be chosen to split the node