Exercises:

Answer All question in an answer sheet (word) and submit only ONE document in the pdf format to Smartv3. Please state your name and matric number clearly on the front page.

1. Compute the information gain (IG) for feature Count and Total. Based on the IG score, identify which feature is the best feature for learning model?

ID	Count	Total	Class
1	A	A	No
2	A	A	No
3	B	B	No
4	B	C	No
5	B	C	Yes
6	C	D	Yes
7	C	D	Yes
8	C	D	Yes
9	D	D	Yes
10	D	D	No

Answer:

IG for Count feature

Ent(C) = $-(5/10)\log_2(5/10)+(5/10)\log_2(5/10)) = 1$

Ent(C|Count) = $-(2/10(2/2\ \log_2(2/2))+ 3/10(2/3\ \log_2(2/3) + 1/3\ \log_2(1/3)) +$

$3/10\ (3/3\ \log_2(3/3)) + 2/10\ (½\ \log_2(½) + ½\ \log_2(½))) = -(0 -0.28-0-0.2)=0.48$

InfoGain(Count) = Ent(C) – Ent(C|Count) = 1 – 0.48 = 0.52

IG for Total feature

Ent(C) = $-(5/10)\log_2(5/10)+(5/10)\log_2(5/10)) = 1$

Ent(C|Total) = $-(2/10(2/2\ \log_2(2/2))+ 1/10(1/1\ \log_2(1/1)\ ) +2/10(1/2\ \log_2(1/2)\ + 1/2\ \log_2(1/2))$

$5/10\ (4/5\ \log_2(4/5) + 1/5\ \log_2(1/5))\ = -(0-0-0.2-0.36) = 0.56$

InfoGain(Total) = Ent(C) – Ent(C|Total) = 1 – 0.56 = 0.44

InfoGain(Count) > InfoGain(Total), hence Count is the best feature for learning model.