School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 2, 2022)
Workshop: week 10

1. Approximately 1% of women aged between 40 and 50 have breast cancer. 80% of mammogram screening tests detect breast cancer when it is there. 90% of mammograms DO NOT show breast cancer when it is **NOT** there[1]. Based on this information, complete the following table.

Cancer	Probability
No	99%
Yes	1%

Cancer	Test	Probability
Yes	Positive	80%
Yes	Negative	? **20%**
No	Positive	? **10%**
No	Negative	90%

P(P | C) = 0.8
P(N | NC) = 0.9
P(N | C) = 1 - P(P | C) = 0.2
P(P | NC) = 1 - P(N | NC) = 0.1

2. Based on the results in question 2, calculate the **marginal probability** of 'positive' results in a Mammogram Screening Test. **P(P) = P(P | C) * P (C) + P(P | NC) * P(NC) = 0.8 * 0.01 + 0.1 * 0.99 = 0.107**
3. Based on the results in question 2, calculate *P(Cancer = 'Yes' | Test = 'Positive')*, using the Bayes Rule.

P(C | P) = P(C) * P(P | C) / P(P) = 0.01 * 0.8 / 0.107 = 0.074

4. For the following dataset:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
TRAINING INSTANCES					
A	s	h	h	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
TEST INSTANCES					
G	o	c	n	T	?
H	s	m	h	F	?

Classify the test instances using the **ID3 Decision Tree** method and **Gain Ratio** as a splitting criterion.

H(R) = 1

H(s) = 0
H(o) = 0
H(r) = - (1/3 * log(1/3) + 2/3 * log(2/3)) = 0.9183
mean-info(outl) = 0 * 1/3 + 0 * 1/6 + H(r) * 1/2 = 0.4592

IG(outl | R) = H(R) - mean-info(outl)
= 1 - 0.4592
= 0.5408

[1] Remember these numbers are not accurate and simplified to ease the calculations in this question.