School of Computing and Information Systems
The University of Melbourne
COMP90049 Introduction to Machine Learning (Semester 1, 2021)
Week 3

1. For the following dataset:

apple	*ibm*	*lemon*	*sun*	CLASS
TRAINING INSTANCES				
4	0	1	1	FRUIT
5	0	5	2	FRUIT
2	5	0	0	COMPUTER
1	2	1	7	COMPUTER
TEST INSTANCES				
2	0	3	1	?
1	2	1	0	?

   (i). Using the **Euclidean distance** measure, classify the test instances using the 1-NN method.

   (ii). Using the **Manhattan distance** measure, classify the test instances using the 3-NN method, for the three weightings we discussed in the lectures: *majority class*, *inverse distance*, *inverse linear distance*.

   (iii). Can we do weighted k-NN using **cosine similarity**?

2. 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.

TP = True Pos
TN = True Neg
FP = False Pos
FN = False Neg

Cancer	Probability
No	99%
Yes	1%

Cancer	Test	Probability
Yes	Positive	80%
Yes	Negative	? 0.2
No	Positive	? 0.1
No	Negative	90%

3. Based on the results in question 1, calculate the **marginal probability** of 'positive' results in a Mammogram Screening Test.

$$P(P) = \sum_{i \in \{NC, C\}} P(P|i)P(i) = \sum_{i \in \{NC, C\}} P(P, i)$$

$$= P(P|C)P(C) + P(P|NC)P(NC)$$

$$= 0.8 \times 0.1 + 0.1 \times 0.99 = 0.107$$

4. Based on the results in question 1, calculate *P(Cancer* = *'Yes'* | *Test* = *'Positive')*, using the Bayes Rule.

$P(P|C) = 0.8$

$P(N|C) = 1 - P(P|C) = 1 - 0.8 = 0.2$

$P(P|NC) = 1 - P(N|NC) = 1 - 0.9 = 0.1$

$P(N|NC) = 0.9$

$$P(C|P) = \frac{P(P|C)P(C)}{P(P)} = \frac{0.8 \times 0.1}{0.107} = 0.75$$

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