Tutorial Letter A3/0/2023

Machine Learning

COS4852

Year module

School of Computing

IMPORTANT INFORMATION

This document contains a discussion on the questions for Assignment 3 for COS4852 for 2023.

BAR CODE

CONTENTS

Page

1 INTRODUCTION

This document contains a discussion on the questions for Assignment 3 for COS4852 for 2023.

2 Assignment 3

Question 1

- https://byjus.com/maths/bayes-theorem/
- https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theo
- https://en.wikipedia.org/wiki/Sensitivity_and_specificity

In the world we are experiencing now, people have become a lot more aware of medical test results. One question that most people would want to be answered: How accurate is a test for an infection?

There are terms often used in medical test results, and are vital in correctly interpreting a test result:

1. **Prevalence**: the ratio of the total population who is infected.
2. **True positive (TP)**: the ratio of the total tests that are accurately labelled as positive. This is dependent on the prevalence.
3. **True negative (TN)**: the ratio of the total tests that are accurately labelled as negative. This is dependent on the prevalence.
4. **Sensitivity**: (true positive rate TPR) the ratio of positive tests that are accurately labelled as positive. This is independent of the prevalence.
5. **Specificity**: (true negative rate TNR) the ratio of negative tests that are accurately labelled as negative. This is independent of the prevalence.
6. **Positive Predictive Value (PPV)**: the probability of an infection given a positive test result.
7. **Negative Predictive Value (NPV)**: the probability of no infection given a negative test result.

Here is a link that will provide more clarity on these terms:

- https://microbenotes.com/sensitivity-specificity-false-positive-false-negative/

Consider the data on a medical test for SUPERBUG:

1. Out of every 10 000 people with a record of possible symptoms, more or less 100 people were diagnosed with SUPERBUG. These are confirmed cases, based on a combination of doctors' diagnoses, CT-scans, several different tests, and post mortem analyses.
2. It is known that for this test, 10 out of 100 positive test results are incorrect. This is the inverse *sensitivity* of the test.
3. It is known that for this test, 10 out of 50 negative test results are incorrect. This is the inverse *specificity* of the test.

Question 1(a)

You have just been tested for SUPERBUG, but are still waiting for your results. Obviously you will want to know what a positive or negative result will tell you about the likelihood that you have been infected with SUPERBUG, so that you can decide how to deal with the result.

Given the data above:

1. Define the variables you will use in your calculations.
2. Calculate the prevalence of SUPERBUG.
3. Calculate the sensitivity and specificity of the test.
4. Calculate the inverse sensitivity and inverse specificity of the test.
5. Calculate the false positive and false negative rate for the test.
6. Calculate the four prior probabilities, using your variable definitions.
7. Use Bayes' theorem to calculate the PPV.
8. Use Bayes' theorem to calculate the NPV.
9. Explain why the probabilities come out this way.
10. Explain what would cause these probabilities to change.
11. What do these results mean in practice.

Mark out of 100.
40 or less for clear indication that student does not understand the topic or evidence of plagiarism
50 for a fair understanding
60-70 for understanding and clear well defined examples
80+ for exceptional detail

Discussion on Question 1(a)

Define the variables. Let:

$$\begin{aligned} \text{Bug} &\leftarrow \text{a person is infected with SUPERBUG} \\ \neg\text{Bug} &\leftarrow \text{a person is not infected with SUPERBUG} \\ \text{Pos} &\leftarrow \text{a positive test result} \\ \text{Neg} &\leftarrow \text{a negative test result} \\ &= \neg\text{Pos} \end{aligned}$$

The actual prevalence (base rate) is unknown. We can only know that number if everybody is tested, with a test that is 100% accurate, and everybody is tested in a very short space of time, or everybody get retested at short intervals. There is no test that is 100% accurate and mass testing is very expensive and complicated. We can therefore only use the frequency of the number of confirmed positive cases against the population size. This will be our best estimate for prevalence. Since we were given a population of 10 000, we will use this number as a convenient population sample to calculate the ratios.

$$\begin{aligned} P(\text{Bug}) &\leftarrow \text{prevalence} \\ &\approx {}^{100}/_{10\,000} \\ &= 0.01 \end{aligned}$$

and therefore:

$$\begin{aligned} P(\neg\text{Bug}) &\leftarrow \text{inverse prevalence} \\ &= 1 - 0.01 \\ &= 0.99 \\ &= {}^{9\,900}/_{10\,000} \end{aligned}$$

The known *sensitivity* and *specificity* tells us that:

$$\begin{aligned} P(\text{Pos}|\text{Bug}) &\leftarrow \text{sensitivity} \\ &= {}^{(100-10)}/_{100} \\ &= 0.9 \\ &= {}^{90}/_{100} \\ P(\text{Neg}|\neg\text{Bug}) &\leftarrow \text{specificity} \\ &= {}^{(50-10)}/_{50} \\ &= 0.8 \\ &= {}^{7\,920}/_{9\,900} \end{aligned}$$

We can also write down their inverses:

$$\begin{aligned} P(\text{Neg}|\text{Bug}) &\leftarrow 1 - P(\text{Pos}|\text{Bug}) \\ &= 1 - 0.9 \\ &= 0.1 \\ &= {}^{10}/_{100} \\ P(\text{Pos}|\neg\text{Bug}) &\leftarrow 1 - P(\text{Neg}|\neg\text{Bug}) \\ &= 1 - 0.8 \\ &= 0.2 \\ &= {}^{1\,980}/_{9\,900} \end{aligned}$$

Therefore, of the 100 infected people (out of the population of 10 000), 90 get positive results, and 10 get negative results. Of the 9 900 people not infected (out of the population of 10 000), $0.8 \times 9\,900 = 7\,920$ get negative results, and $9\,900 - 7\,920 = 1980$ get positive results. Therefore $90 + 1\,980 = 2\,070$ came back positive, and $7\,920 + 10 = 7\,930$ came back negative.

We can also now write down the following probabilities:

$$\begin{aligned} \text{True positive} &\leftarrow \text{ratio of all tests correctly labelled as positive} \\ &= {}^{90}/_{10\,000} \\ &= 0.009 \\ \text{True negative} &\leftarrow \text{ratio of all tests correctly labelled as negative} \\ &= {}^{7\,920}/_{10\,000} \\ &= 0.792 \\ \text{False positive} &\leftarrow \text{ratio of all tests incorrectly labelled as positive} \\ &= {}^{1\,980}/_{10\,000} \\ &= 0.198 \\ \text{False negative} &\leftarrow \text{ratio of all tests incorrectly labelled as negative} \\ &= {}^{10}/_{10\,000} \\ &= 0.001 \end{aligned}$$

We can represent these values in table form, as shown in Tables 1 and 2. Figure 1 shows this visually, but the small ratios make this not so easy to see. Figure 2 shows another set of ratios, where the prevalence becomes 25%, which gives a better idea. As an exercise, repeat the calculations for a prevalence of 25%, to see if you can get the same numbers as in the figure.

	Bug	¬Bug	totals
Pos	90	1980	2070
Neg	10	7920	7930
totals	100	9900	10 000

Table 1: Ratios (out of 10 000) for SUPERBUG and it's test.

	Bug	¬Bug	totals
Pos	0.009	0.198	0.207
Neg	0.001	0.792	0.793
totals	0.010	0.990	1.000

Table 2: probabilities for SUPERBUG and it's test.

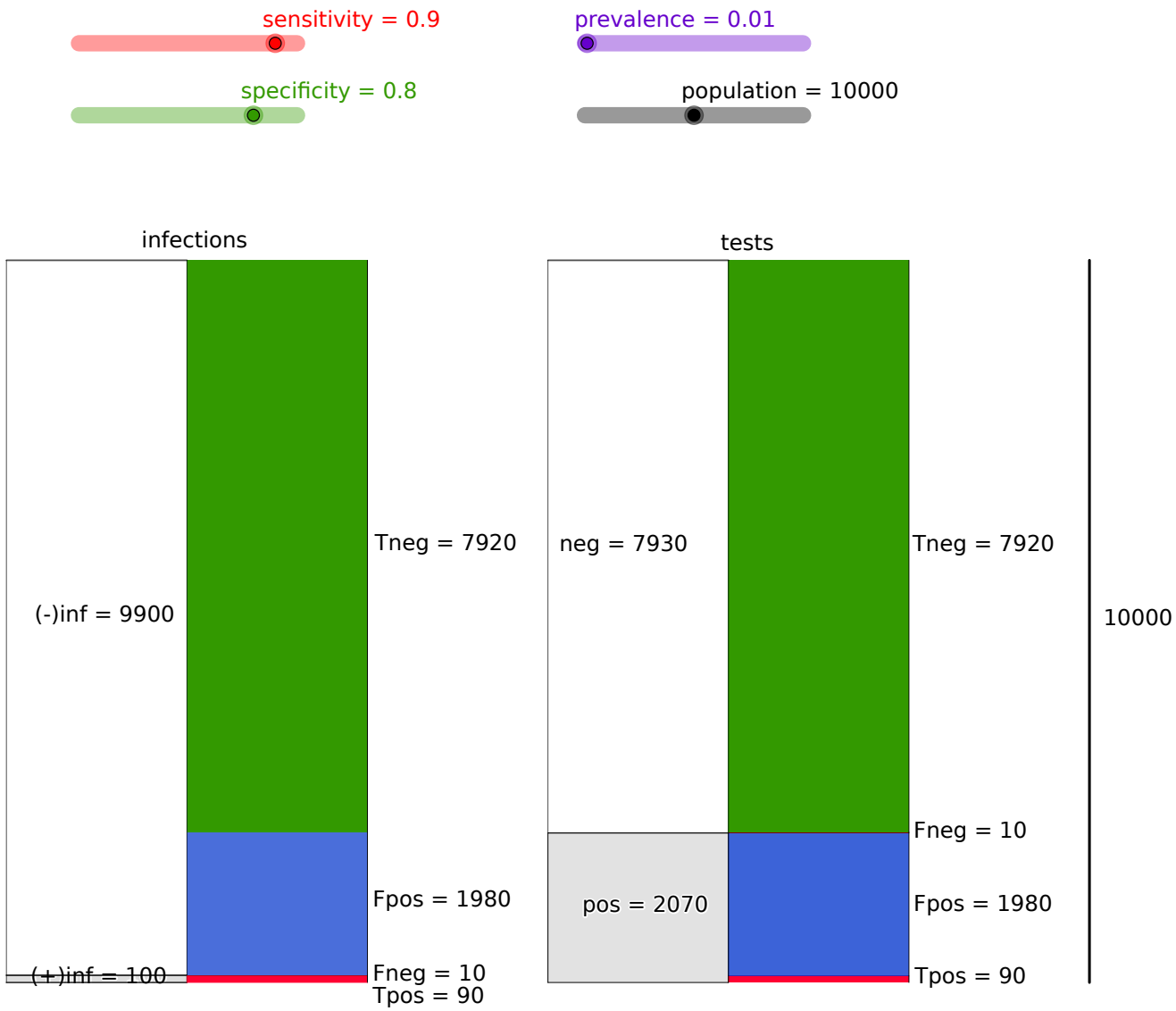

Figure 1: Bug test ratios shown visually for a prevalence of 0.01.

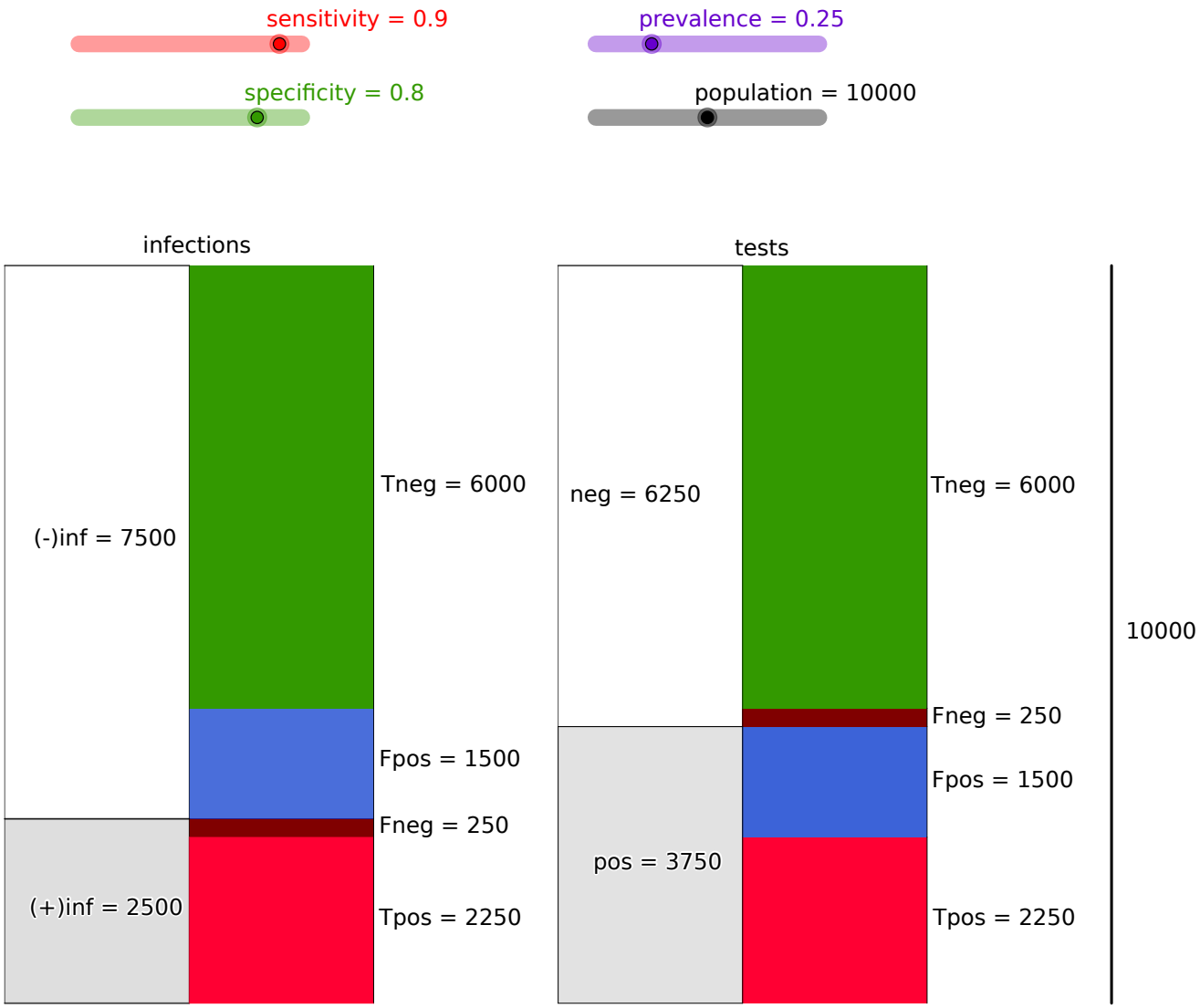

Figure 2: Bug test ratios shown visually for a prevalence of 0.25.

What we cannot immediately calculate, are the following:

$$\begin{aligned} P(\text{Bug}|\text{Pos}) &\leftarrow \text{chance of having the bug given a positive test} \\ P(\text{Bug}|\text{Neg}) &\leftarrow \text{chance of having the bug given a negative test} \\ P(\neg\text{Bug}|\text{Pos}) &\leftarrow \text{chance of not having the bug given a positive test} \\ P(\neg\text{Bug}|\text{Neg}) &\leftarrow \text{chance of not having the bug given a negative test} \end{aligned}$$

Bayes' rule can be used to calculate these probabilities, and is expressed using the prior (known) probabilities $P(A)$, $P(B)$, and the probability of B given that A is true, $P(B|A)$. Bayes' theorem gives us the probability of A given B, $P(A|B)$ as:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

where

$$P(B) = P(A)P(B|A) + P(\neg A)P(B|\neg A)$$

To summarise the known probabilities:

$$\begin{aligned} P(\text{Bug}) &= 0.01 \\ P(\neg\text{Bug}) &= 0.99 \\ P(\text{Pos}|\text{Bug}) &= 0.90 \\ P(\text{Neg}|\text{Bug}) &= 0.10 \\ P(\text{Neg}|\neg\text{Bug}) &= 0.80 \\ P(\text{Pos}|\neg\text{Bug}) &= 0.20 \end{aligned}$$

To test the posterior probability of having SUPERBUG, given that the test results came back positive, let:

$$\begin{aligned} A &= \text{Bug} \\ B &= \text{Pos} \end{aligned}$$

Plugging these into Bayes' theorem, we get:

$$\begin{aligned} P(B) &= P(A)P(B|A) + P(\neg A)P(B|\neg A) \\ P(\text{Pos}) &= P(\text{Bug})P(\text{Pos}|\text{Bug}) + P(\neg\text{Bug})P(\text{Pos}|\neg\text{Bug}) \\ &= (0.01 \times 0.90) + (0.99 \times 0.20) \\ &= 0.009 + 0.198 \\ &= 0.207 \end{aligned}$$

$$\begin{aligned} P(\text{A}|\text{B}) &= \frac{P(\text{A}) \times P(\text{B}|\text{A})}{P(\text{B})} \\ P(\text{Bug}|\text{Pos}) &= \frac{P(\text{Bug}) \times P(\text{Pos}|\text{Bug})}{P(\text{Pos})} \\ &= \frac{0.01 \times 0.90}{0.207} \\ &= \frac{0.009}{0.207} \\ &= 0.04348 \end{aligned}$$

To test the posterior probability of *not* having SUPERBUG, given that the test results came back negative, let:

$$\begin{aligned} A &= \neg\text{Bug} \\ B &= \text{Neg} \end{aligned}$$

$$\begin{aligned} P(B) &= P(A)P(B|A) + P(\neg A)P(B|\neg A) \\ P(\text{Neg}) &= P(\neg\text{Bug})P(\text{Neg}|\neg\text{Bug}) + P(\text{Bug})P(\text{Neg}|\text{Bug}) \\ &= (0.99 \times 0.80) + (0.01 \times 0.10) \\ &= 0.792 + 0.001 \\ &= 0.793 \end{aligned}$$

$$\begin{aligned} P(\text{A}|\text{B}) &= \frac{P(\text{A}) \times P(\text{B}|\text{A})}{P(\text{B})} \\ P(\text{Bug}|\text{Neg}) &= \frac{P(\neg\text{Bug}) \times P(\text{Neg}|\neg\text{Bug})}{P(\text{Neg})} \\ &= \frac{0.99 \times 0.80}{0.000792} \\ &= \frac{0.792}{0.793} \\ &= 0.9987 \end{aligned}$$

This means that a positive test results indicated that you only have a 4.3% likelihood of being infected with SUPERBUG. The reason for this very low number is related to the low figure we have for the prevalence of the infection in the population. This means that most people who are testes, and gets a positive result have not been infected. In a pandemic this may be acceptable, since you want to prevent the spread of the infection. In a situation where the majority of people are negative, mass testing will not be helpful. When the testing regime is changed to only test people with a high likelihood of being positive (such as having known symptoms, or having been exposed to a known person with the disease), the prevalence in the tested population will go up, and the PPV will also increase.

Something else to consider is that the prevalence is related to the population being *tested*. If the tests are done randomly, you would see the sort of figures shown here. However, in reality people are only tested when there is a valid suspicion that they may be positive, such as having most of the related symptoms, or if they have been exposed to a known positive person. In such a population, the prevalence will go up significantly, with a 25% prevalence being about right. Even in such a case, there are still very many false positives, but the false negatives are very low.

Similarly, we can calculate the posterior probability of not having SUPERBUG, given a negative test result:

$$\begin{aligned} A &= \neg\text{Bug} \\ B &= \text{Neg} \end{aligned}$$

Bayes' theorem therefore tells us:

$$\begin{aligned} P(\neg\text{Bug}|\text{Neg}) &= \frac{P(\neg\text{Bug}) \times P(\text{Neg}|\neg\text{Bug}))}{P(\text{Neg})} \\ &= \frac{0.9891 \times 0.98}{0.9694} \\ &= 0.9999 \end{aligned}$$

which means that a negative test results indicates that you have an almost 100% likelihood of not being infected with SUPERBUG.

Keep in mind that the numbers used here are purely hypothetical, and does not reflect any real figures. False positive and false negative rates are as problematic in a new test as the prevalence numbers would be. In the real world, you don't look at the test results in isolation. First, there is the uncertainty of the actual frequency of the population who has the infection. When looking at something like cancer tests, where their is a long history of medical data to work with, the figures about prevalence is very accurate, and becomes more so as more data is collected. With a developing pandemic, such as we see now, there simply is not enough data to get the prevalence accurately enough, and estimates are used, with various models developed specifically for this.

The timing of the test for a viral infection has an important part to play in the accuracy of test results. Viral tests, for example, are dependant on the patient having enough viral particles for the test to extract sufficient material for the test, and a day makes a huge difference in the viral load. Further, RNA tests are highly sensitive, due to the RNA being unique for every entity. Such a test usually has a very low false positive rate, but the false negative could be very high due to timing, mishandling of the sample, lab errors, and so on.

The prevalence would also change as the outbreak progresses, but will get more accurate in time. In the case of some outbreaks, there are people who are positive, but does not experience any symptoms. Most of these people will not get diagnosed, nor tested. This would therefore mean that the prevalence rate is underestimated. Antibody tests may indicate whether somebody have had the infection in the past, which could then be added to the prevalence numbers. Again, such tests will not be done on everybody, for practical and economic reasons, and people may lose their immunity in time. In short, the prevalence rate is at bet an estimate.

In a pandemic situation, where the bulk of the population does not have the infection, we find that the probability of a positive test being correct is much lower than a negative test being correct. Here is an excellent page that explains why it is important to use a test that has high specificity.

- `https://towardsdatascience.com/bayes-rule-with-a-simple-and-practical-example-2bce3d0`
- `https://www.aruplab.com/news/4-21-2023/How-Accurate-Are-COVID-19-Tests`
- `https://www.bmj.com/content/369/bmj.m1808`

Question 2

Some information about the causes and symptoms of *Lung Cancer* are available. The factors that affect a patient's chances of having lung cancer are:

- *Pollution* (measured as *High* or *Low*)
- *Smoker* (measured as *Yes* or *No*)

It is also know that if a patient has *Lung Cancer* (measured as *True* or *False*) it will affect the patient's symptoms, namely:

- The patient having an abnormal chest *X-ray* result (measured as *Abnormal* or *Normal*)
- The patient experiencing *Dyspnoea* (difficult breathing in English, measured as *Present* or *Absent*)

The following evidence is available about patients in general:

- 91% of patients are not exposed to high levels of *Air Pollution*
- 33% of patients are *Smokers*

Of patients that have *Lung Cancer* the following is known:

- 4% have been exposed to high levels of *Air Pollution* and are *Smokers*
- 3% have been exposed to high levels of *Air Pollution*, but are not *Smokers*
- 2% have had low exposure to *Air Pollution* and are *Smokers*
- 1 out of every 1000 have had low exposure to *Air Pollution* and are not *Smokers*

The following is known about the *X-ray* results of patients tested for *Lung Cancer*:

- 19 out of 20 patients with *Cancer* have *abnormal X-ray* results
- 1 out of 6 patients who do not have *Cancer* have *abnormal X-ray* results

The following is known about the symptoms of *Dyspnoea*:

- 70% of patients with *Cancer* have symptoms of *Dyspnoea*
- 25% of patients who do not have *Cancer* have symptoms of *Dyspnoea*

Answer the following questions:

a) Construct a Bayesian Belief Network that illustrates the conditional dependencies between the five variables, and draw a diagram to illustrate the network.

b) On your belief network diagram, show the conditional probability tables for each of the five variables.

Here is a links you can reference to give you the basic background behind Bayesian Belief Networks:

- `https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/`
- `https://www.cs.ubc.ca/~murphyk/Bayes/Charniak_91.pdf`

Mark out of 100.
40 or less for clear indication that student does not understand the topic or evidence of plagiarism
50 for a fair understanding
60-70 for understanding and clear well defined examples
80+ for exceptional detail

Discussion on Question 2(*a*))

Define the variables and their values:

Variable	shortened	values	shortened in set notation
Lung Cancer	*LC*	*True*, *False*	$\{T, F\}$
Air Pollution	*Air*	*High*, *Low*	$\{H, L\}$
Smoker	*Sm*	*Yes*, *No*	$\{Y, N\}$
X-ray	*Xr*	*Abnormal*, *Normal*	$\{A, N\}$
Dyspnoea	*Dys*	*Present*, *Absent*	$\{P, A\}$

The network contains five nodes, *Lung Cancer*, *Air Pollution*, *Smoker*, *X-ray* and *Dyspnoea*. From the information provided *Lung Cancer* is conditionally dependent on *Air Pollution* and *Smoker*, while the probabilities of the *X-ray* results and *Dyspnoea* as a symptom is dependent on whether a patient has *Lung Cancer* or not.

Figure 3 shows the structure of these dependencies in the Bayesian lBelief Network, as nodes and arrows.

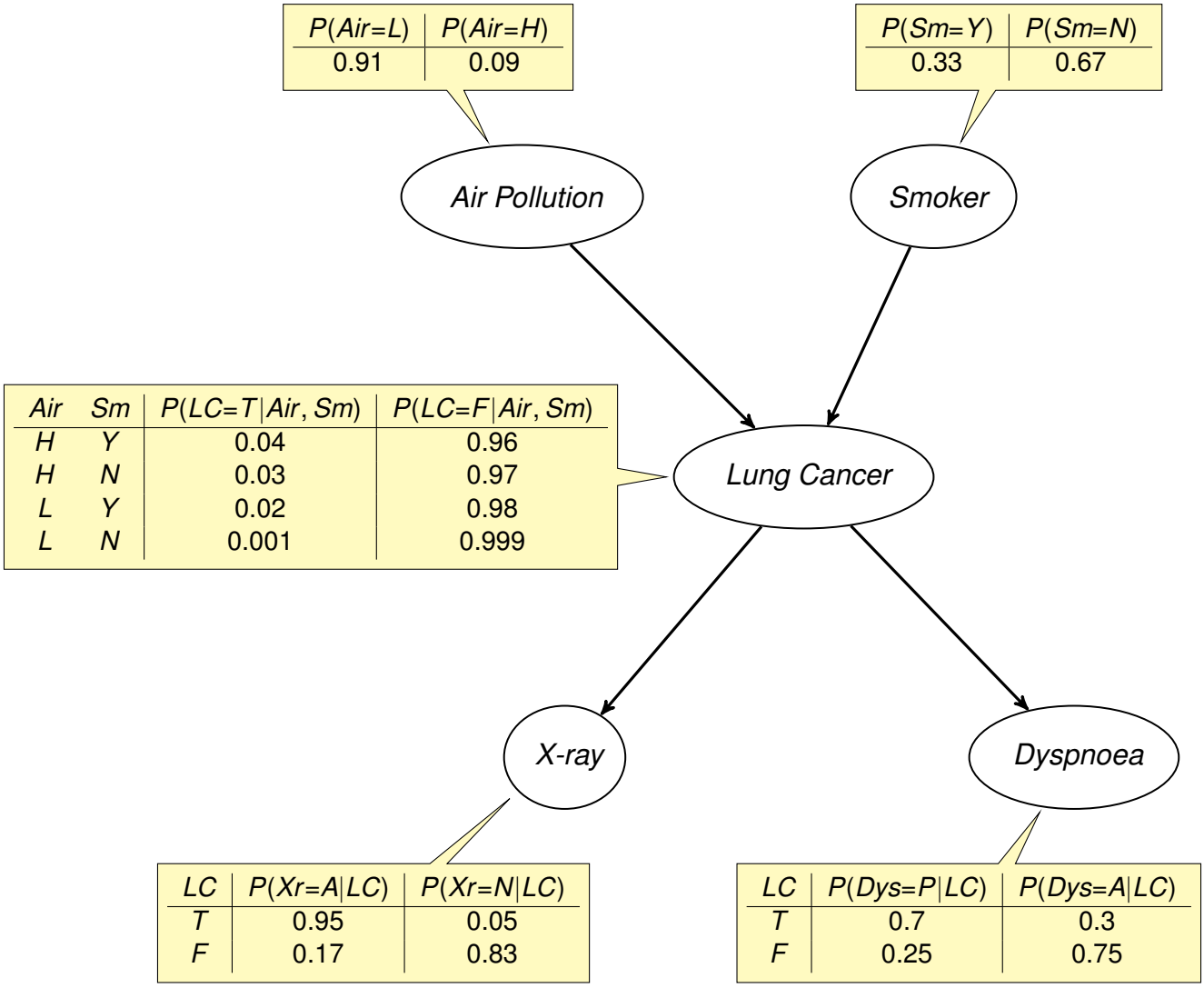

Figure 3: A Bayesian Belief Network showing the (a) structure of the network and (b) the conditional dependencies of the *Lung Cancer* data.

Discussion on Question 2(*b*))

Probability tables are shown in Figure 3.

Question 3

a) Write a short research report (roughly 10 pages) and give a detailed explanation of the basic GA algorithm. Explain the terminology used, the operators, how to represent a solution to the task as a chromosome, and how to determine/design a good fitness function, as well as what criteria defines a good fitness function. Briefly discuss some variations on the basic GA algorithm, and what they are used for.

The source of literature should ideally use textbooks, or articles published in scientific journals or conferences. Review articles are a good option. Use the Harvard Referencing method.

The book *'An Introduction to Genetic Algorithms'* by *Melanie Mitchell* is available on the Internet as a PDF document. This book provides good detail on the basics of Genetic Algorithms.

- `https://svn-d1.mpi-inf.mpg.de/AG1/MultiCoreLab/papers/ebook-fuzzy-mitchell-99.pdf`

Other useful links:

- `http://www.cs.cmu.edu/~02317/slides/lec_8.pdf`
- `https://towardsdatascience.com/how-to-define-a-fitness-function-in-a-genetic-algorith`

b) Consider the following optimisation tasks:

(i) Optimise the function $f(x) = x^2$ over the interval $[0, 31]$
(ii) Find the square root of a given number without using the square root calculation.
(iii) Find values for x and y that will satisfy the equation $2x^2 + y = 23$
(iv) Find the intersection point(s) of $y = 0.4^{x-2} - 3$ and $2x^2 + y = 23$

For each of these tasks:

- Define the limitations of the task.
- Design an appropriate chromosome to represent solutions to the task.
- Define a fitness function, keeping the criteria for good fitness functions in mind.
- Generate an initial population of 10 members, apply the fitness function to these, and choose the top 5 for reproduction.
- Calculate the exact solution(s) without using a Genetic Algorithm (using normal mathematical techniques).

Mark out of 100.
40 or less for clear indication that student does not understand the topic or evidence of plagiarism
50 for a fair understanding
60-70 for understanding and clear well defined examples
80+ for exceptional detail

Discussion on Question 3

a) The research report is marked individually.

b) Genetic Algorithms create huge populations of possible solutions. Each of these members will have a number of operations applied to them, from fitness testing to the generation of the new population. That means the the chromosome encoding and the fitness function need to be as storage and computationally efficient as possible. The fitness function also should also be clearly understood in relation to the optimisation problem, and quantitatively measure the fitness of a particular chromosome/solution. These fitness scores should also be intuitive - better scores produce better solutions.

(i) Optimise the function $f(x) = x^2$ over the interval $[0, 31]$.

- The problem domain is limited to the range $[0, 31]$. It is also assumed that optimisation here implies maximising the function value (if you are trying to minimise the function value the fitness function is inverted). It is also assumed that the domain is limited to integer x-values.
- It is possible to encode the chromosome using a 5-bit string by simply using the binary encoding of the integer. For example, $x = 1$ is encoded as 00001, $x = 23$ is encoded as 11101.
- The fitness function is simply testing the chromosome against the function. Let c be the integer value associated with a chromosome.

$$\text{fit}(c) = c^2$$

  Higher fitness values are directly correlated to higher function values.
- 10 randomly generated chromosomes make up the initial population:

chromosome	c	fit(c)	top 5?
00101	20	400	yes
10010	9	81	no
11101	23	529	yes
10000	1	1	no
10001	17	289	yes
10101	21	441	yes
01110	14	196	no
11000	3	9	no
11110	15	225	no
01001	18	324	yes

- For reference, $f(31) = 31^2 = 961$.

(ii) Find the square root of a given number without using the square root calculation.

- The task is to find $y = \sqrt{x}$ for a given x. By definition the square root of a positive real number will be smaller than the number. We can therefore limit the domain to $y \in [0, x]$. This does mean that the bulk of the chromosomes will be unfit, though. Can you think of a better mechanism to limit the domain?

- The chromosome in this case could simply be a real number in the valid domain: $c \in [0, x]$, with $c, x \in \mathbb{R}$. This does mean that the binary crossover and mutation operations need to be replaced by real-valued versions, for which there are a number of techniques to implement. Of course, real valued, floating-point numbers can also be represented as binary numbers (as is done in the hardware of a computer already).
- Since multiplication is a much less computationally intensive operation than calculation the square root, fitness can be tested by simply squaring a potential solution and comparing this with the input value, x. The closer the potential solution is to being correct, the smaller this difference. Therefore, to correlate higher fitness values with solutions that are closer to correct, we need to invert the difference. The fitness function now becomes:

$$\text{fit}(c) = \frac{1}{|x - c^2|}$$

- Let $x = 100$. Then, 10 randomly generated chromosomes make up the initial population:

chromosome	fit(c)	top 5?
60	$1/3500$	no
8	$1/36$	yes
45	$1/1925$	no
14	$1/96$	yes
47	$1/2109$	no
97	$1/9309$	no
24	$1/476$	yes
23	$1/429$	yes
11	$1/21$	yes
63	$1/3869$	no

- For reference, $\sqrt{100} = 10$.

(iii) Find values for x and y that will satisfy the equation $2x^2 + y = 23$

- The task is to find values for x and y that will satisfy $2x^2 + y = 23$. It is immediately obvious that there are infinitely many solutions, but since this is a quadratic equation, $y = f(x) = 23 - x^2$, it tells us that $f(x)$ has a maximum value of $y = 23$ at $x = 0$. This allows us to limit the $y \in [-\infty, 23]$. x is unlimited, $x \in [-\infty, \infty]$.
- The chromosome in this case could simply be a set of real numbers $C = \{c_1, c_2\}$ in the valid domains: $c_1 \in [-\infty, 23]$, $c_2 \in [-\infty, \infty]$ and $c_1, c_2 \in \mathbb{R}$. Again, this means that the binary crossover and mutation operations need to be replaced by real-valued versions.
- The fitness function values will be larger the closer $2x^2 + y - 23$ gets to 0. Invert the difference to get the fitness function:

$$\text{fit}(C) = \frac{1}{2c_1^2 + c_2 - 23}$$

- To limit the initial search for solution, choose 10 randomly generated chromosomes from the ranges, $c_1 \in \{-10, 10\}$ and $c_2 \in \{-10, 23\}$:

chromosome	fit(C)	top 5?
$\{7, 6\}$	$1/81$	no
$\{-9, 18\}$	$1/157$	no
$\{-8, -4\}$	$1/101$	no
$\{-3, 16\}$	$1/11$	yes
$\{-9, -7\}$	$1/132$	no
$\{-7, 9\}$	$1/84$	no
$\{6, 10\}$	$1/59$	yes
$\{-4, 18\}$	$1/27$	yes
$\{-5, -3\}$	$1/24$	yes
$\{-7, 2\}$	$1/77$	yes

- There are infinitely many solutions. For reference, the following three points are known to satisfy the equation:

$$(\pm\sqrt{23/2}, 0)$$

$$(0, 23)$$

(iv) Find the intersection point(s) of $y = 0.4^{x-2} - 3$ and $2x^2 + y = 23$

- This task is to find value(s) for x where $0.4^{x-2} - 3 = 23 - 2x^2$. The value(s) of y can then be calculated by plugging the x values into either of the equations and solving for y. These are a quadratic function and a logarithmic function, so we can infer that there could be 0, 1, or 2 intersection points. As in the previous task, one of the equation has a maximum at $(0, 23)$, so there will not be any intersection points at values of y larger than 23. This allows us to limit the $y \in [-\infty, 23]$. x is unlimited, $x \in [-\infty, \infty]$.
- The chromosome in this case could simply be a single real numbers c in the valid domain for x: $c \in [-\infty, \infty]$. Again, this means that the binary crossover and mutation operations need to be replaced by real-valued versions.
- The fitness function values will be larger the smaller the difference between the two equations. Again, invert the difference to get larger fitness values for smaller differences:

$$\begin{aligned} \text{fit}(c) &= \frac{1}{|(0.4^{x-2} - 3) - (23 - 2x^2)|} \\ &= \frac{1}{|2x^2 + 0.4^{x-2} - 26|} \end{aligned}$$

- To limit the initial search for solution, choose 10 randomly generated chromosomes from the range, $c \in \{-10, 10\}$:

chromosome	fit(c)	top 5?
18	0.00161	no
-3	0.01115	yes
20	0.00236	no
14	0.00273	no
3	0.13158	yes
6	0.02173	yes
-2	0.04748	yes
-7	0.00026	yes
-1	0.11940	yes
4	0.00321	no

- For reference the intersection points are at:

$$(-1.46211, 20.86244)$$

and

$$(5.09325, -2.94124)$$
