Discussion Document A1-S/0/2023

Machine Learning

COS4852

Year module

Department of Computer Science

School of Computing

CONTENTS

This document discusses the questions in Assignment 1 for COS4852 for 2023.

Define tomorrow.

UNISA university of south africa

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

This document discusses the questions in Assignment 1 for COS4852 for 2023.

Each question (except Q1 = 10 marks) will be assigned a mark out of 100 and the total mark for the assignment is then calculated out of (10 + (5 × 100)) = 510.

When we mark the question we want to see that YOU understand the work. Simply copying or regurgitating other peoples' work (from the web, previous solutions, other students' work) does not show that YOU understand the work. Show ALL your assumption, definitions, variables, and full calculations.

2 Assignment 1

Question 1

Find and download the following online textbooks on Machine Learning:

- Introduction to Machine Learning, Nils J. Nilsson, 1998.
- A first encounter with Machine Learning, Max Welling, 2011.

Give the complete URL where you found these textbooks, as well as the file size of the PDF you've downloaded.

Here are the links to the books:

`http://ai.stanford.edu/~nilsson/MLBOOK.pdf`	1855 Kb
`https://www.ics.uci.edu/~welling/teaching/273ASpring10/IntroMLBook.pdf`	416 Kb

10 marks for complete and correct URL and size

Question 2

Read Nilsson's book, Chapter 2. Summarise the chapter in 2-4 pages in such a way that you can show that you thoroughly understand the concepts described there. Use different example functions from the ones in the book to show that you understand the concepts.

Answers are marked individually.

Decision lists are a way to partition a space using Boolean expressions in Disjunctive Normal Form (DNF), and form the basis of understanding binary decision trees, which give a more restrictive division than decision lists. The mapping of DNF to decision lists are also relevant to Question 5.

- `http://www.cs.utexas.edu/~klivans/f07lec3.pdf`
  Discusses DNF, decision trees, and decision lists.
- `http://www.cdam.lse.ac.uk/Reports/Files/cdam-2005-23.pdf`
  Discusses the mapping between Boolean functions and decision lists, as well as some theory.

Linear separability is an important concept in many machine learning algorithms, but especially so in neural networks, the subject of your next task and assignment 2. See the following resources on this topic:

- `http://www.ece.utep.edu/research/webfuzzy/docs/kk-thesis/kk-thesis-html/node19.html`
  Short and concise discussion.
- `https://stackoverflow.com/questions/13976565/neural-networks-what-does-linearly-separa`
  Informal discussion.
- `https://onlinecourses.science.psu.edu/stat857/node/240`
  Adds the notions of the hyperplane and support vectors.

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

Question 3

Read Chapter 5 of Welling's book. Do some research on the k-nearest neighbour classification algorithm and write a 2-page report on how the algorithm works. Your report should include a detailed example, with all calculations shown.

A brief summary of the *k*-Nearest Neighbours algorithm

The *k*-Nearest Neighbours (*k*NN) algorithm is one of the simplest, though widely useful, classification algorithms. It works on the principle that instances of the same class tend to cluster together. In other words, a new instance is very likely to be of the same class as those closest to it.

A target function $f : X \rightarrow Y$, is represented by a set of n instances $\langle X_i, Y_i \rangle$, where $X = \{X_1, X_2, \dots, X_n\}$ are a set of attribute values. These attribute values could be coordinates, or any combination of values that belong to a specific instance. Y_i typically represent a single class value that matches the attribute values of X_i. When a new instance $X_j = \{X_{j1}, X_{j2}, \dots, X_{jn}\}$, of unknown class has to be classified, *k*NN calculates the distance between X_j and each of the other instances. The k nearest neighbours are selected, and their class values counted to determine the majority class. This majority class is then assigned to the new instance X_j.

The distance measure is selected to match the data types of the instance attributes. These include the Euclidean distance, and the Manhattan distance. There are several others that are used. For example, if the attributes are coordinate values, the Euclidean distance measure works well.

The value of k is also critical to the algorithm. With $k = 1$ the new instance will be assigned the class of the nearest neighbour, which may be an outlier, and therefore not be an accurate representation of the classes. Small values of k may lead to overfitting. Larger values of k can lead to underfitting. If $k = n$, the class value of all the instances are used, and there is no point in calculating the distances. Clearly there are values of k that are close to optimal. Statistical methods such as cross-validation can be used for this. A simple heuristic value, that is often used is $k = \sqrt{n}$, or more specifically the nearest uneven integer to $\sqrt{n}$. The value of k should be uneven so that there is always a majority outcome.

An example will illustrate the workings of the algorithm. Consider the instance set in Figure 1, showing 8 instances of two classes A and B. A new instance P_9 at $(2, 1)$ has an unknown class.

Use *k*NN to determine the new class for C. Using the heuristic k should be chosen as $k = 3$, but to illustrate the effect of different distance measures, use $k = 5$. In other words find the 5 nearest neighbours to C.

Use the Euclidian distance measure

$$d_{Euclidian}(p, q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

to calculate the distance between P_9 and the other 8 instances, and rank them according to the closest distance.

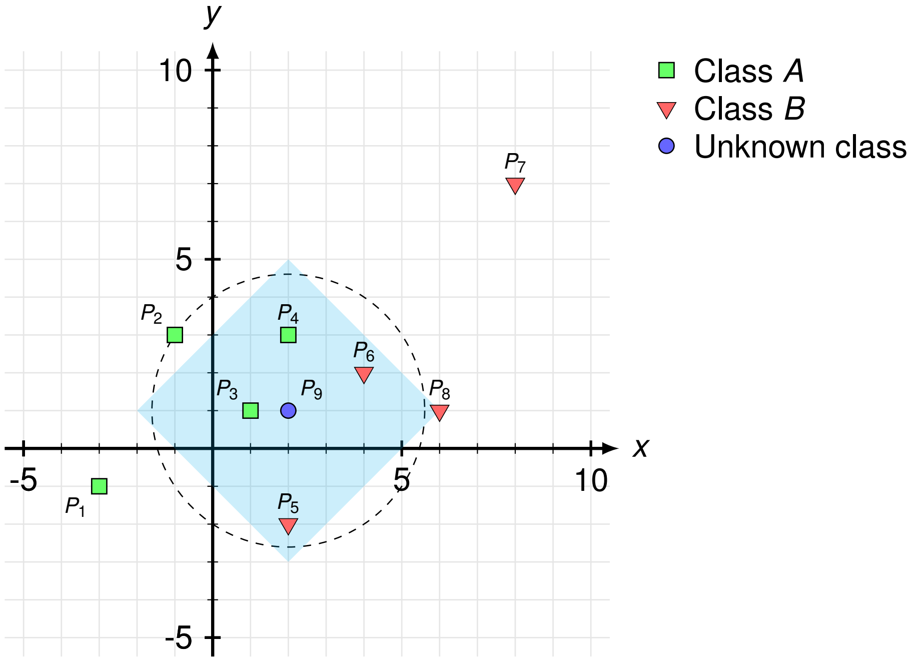


Figure 1: Example instance space for *k*NN with two classes *A* and *B*.

instance	$d_{Euclidian}(P_9, P_i)$	class	rank
P_1	5.385	A	7
P_2	3.606	A	5
P_3	1.000	A	1
P_4	2.000	A	2
P_5	3.000	B	4
P_6	2.236	B	3
P_7	8.485	B	8
P_8	4.000	B	6

With $k = 5$, the 5 closest neighbours gives 3 instances of class A and 2 instances of class B. The majority is therefore class A, hence P_9 is assigned class A. The dashed circle in Figure 1, with radius $r = 3.606$ shows the minimum Euclidian radius that encloses the 5 closest neighbours to P_9.

Now use the Manhattan distance measure

$$d_{Manhattan}(p, q) = |p_x - q_x| + |p_y - q_y|$$

to do the same calculation (read up on the Hamming and the Cityblock distance measures).

instance	$d_{Euclidian}(P_9, P_i)$	class	rank
P_1	7	A	7
P_2	5	A	6
P_3	1	A	1
P_4	2	A	2
P_5	3	B	3
P_6	3	B	4
P_7	12	B	8
P_8	4	B	5

Now, the 5 closest neighbours gives a different result, with 2 instances of class A and 3 instances of class B. The majority is therefore class B, hence P_9 is assigned class B. The cyan diamond in Figure 1, with Manhattan radius $r = 4$ shows the minimum Manhattan radius that encloses the 5 closest neighbours to P_9. This illustrates the importance of choosing the correct distance measure for the data set. If x and y are simply coordinates, the Euclidian distance measure is appropriate, but if x and y represent natural numbers (say x are the number of petals on a flower, and y is the number of sees lobes), then the Manhattan distance may be a better choice.

It is often a good idea to normalise the data, so that all attributes fall within the same range, i.e. have the same scale so that distance measures compares the attributes equally.

Here are some resources you should consult on this topic:

- `http://www.saedsayad.com/k_nearest_neighbors.htm`
  Discusses the basic algorithm and some distance measures and the normalisation process.
- `%http://www.statsoft.com/textbook/k-nearest-neighborshttps://stats.libretexts.org/Bookshelves/Computing_and_Modeling/RTG%3A_Classification_Methods/3%3A_K-Nearest_Neighbors_(KNN)`
  This document is part of series of open-access text for tertiary education. Keep in mind that the article uses both a probabilistic and a distance approach to classifiying new data points. Don't get confused between the two. Also, read the paragraph about the effect of k carefully.
- `http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/`
  Gives a good overview of the curse of dimensionality and the problem of overfitting, which are problems that can occur with classification methods. It also discusses the usefulness of cross-validation.

Do not use the Wikipedia entry on kNN. It ignores the basic algorithm and focuses on the more complex variants of the algorithm, and may be confusing.

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

Question 4

Let **X** be an instance space consisting of points in the Euclidian plane with *integer* coordinates (x, y), with positive and negative instances as shown in Figure 2.

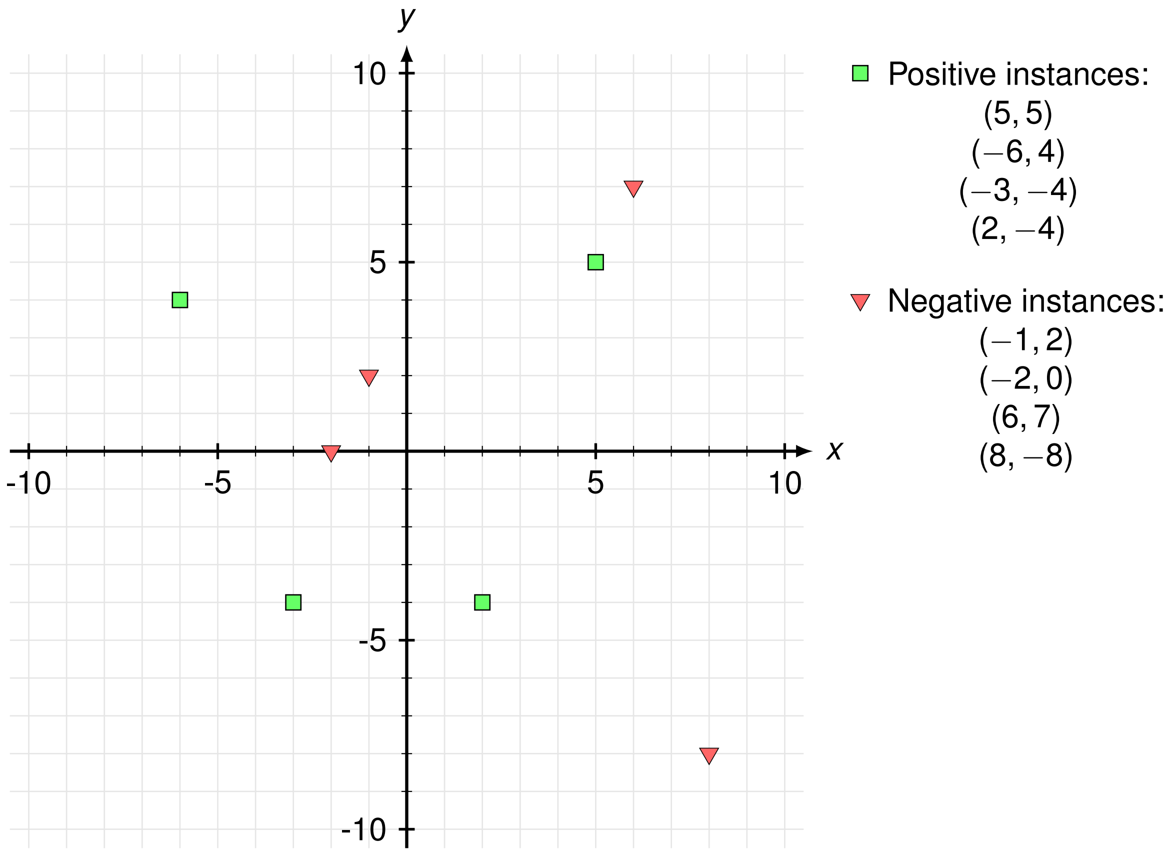


Figure 2: Instance space with positive and negative instances.

Let **H** be the set of hypotheses consisting of origin-centered *donuts*. Formally, the *donut* hypothesis has the form $h \leftarrow \langle a < \sqrt{x^2 + y^2} < b \rangle$, where $a < b$ and $a, b \in \mathbb{Z}$ ($\mathbb{Z}$ is the set of non-negative integers, $\{0, 1, 2, 3, ...\}$). This can be shortened to $h \leftarrow \langle a, b \rangle$.

An example of a *donut* hypothesis is $h \leftarrow \langle 2, 5 \rangle$ and is shown in Figure 3. Notice that this hypothesis does *not* explain the data correctly, since there are both positive and negative instances inside the *donut* and neither does the *donut* contain *all* the positive or *all* the negative instances, exclusively.

(a) What is the **S**-boundary set of the given version space? Write out the hypotheses in the form given above and draw them.

(b) What is the **G**-boundary set of the given version space? Write out the hypotheses in the form given above and draw them.

(c) Suppose that the learner now suggests a new (x, y) instance and asks the trainer for its classification. Suggest a query guaranteed to reduce the size of the version space, regardless of how the trainer classifies it. Suggest one that will not reduce the size of the version space, regardless of how the trainer classifies is. Explain why in each case.

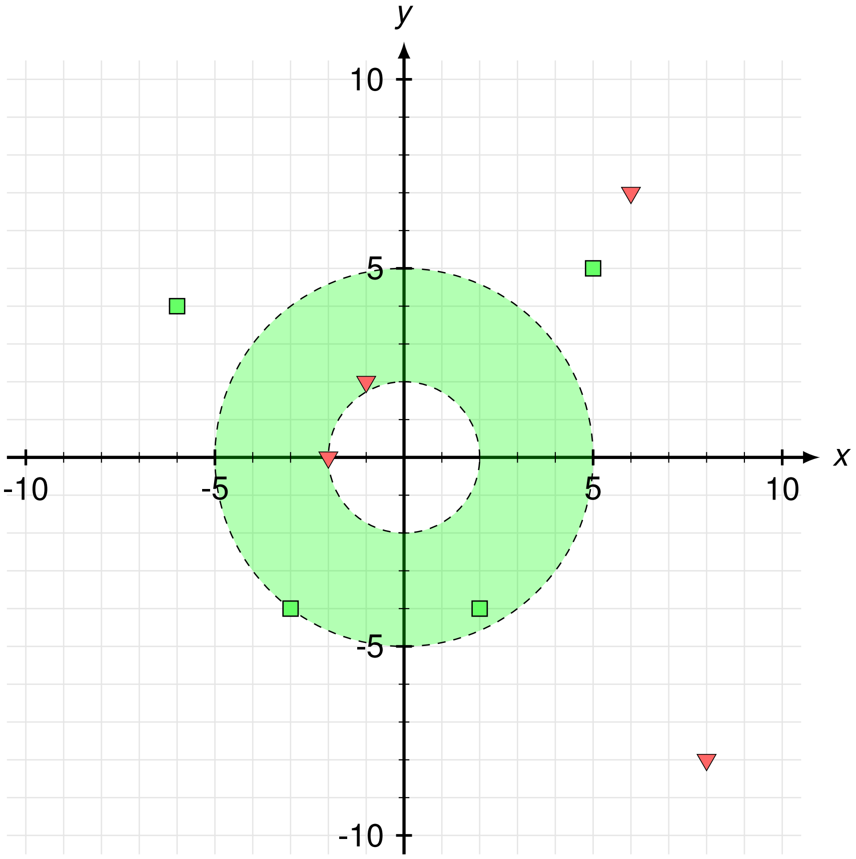


Figure 3: Instance space with a *donut* hypothesis $h \leftarrow \langle 2, 5 \rangle$.

(d) The *donuts* are one of many possible hypothesis spaces that could explain this data set. Propose one alternative hypothesis space and explicitly define its parameters as was done using *a* and *b* for the *donuts*. Choose an instance from your hypothesis space that separates the given data. Write out this hypothesis and sketch it.

Here are some resources you could consult on this topic:

- `http://cse-wiki.unl.edu/wiki/index.php/Concept_Learning_and_the_General-to-Specific_Ordering`
- `http://www.cs.northwestern.edu/~pardo/courses/mmml/lectures/NU%20EECS%20349%20Fall%2009%20topic%201%20-%20version%20spaces.pdf`
- `http://www.ccs.neu.edu/home/rjw/csg220/lectures/version-spaces.pdf`

Discussion on (a) and (b):

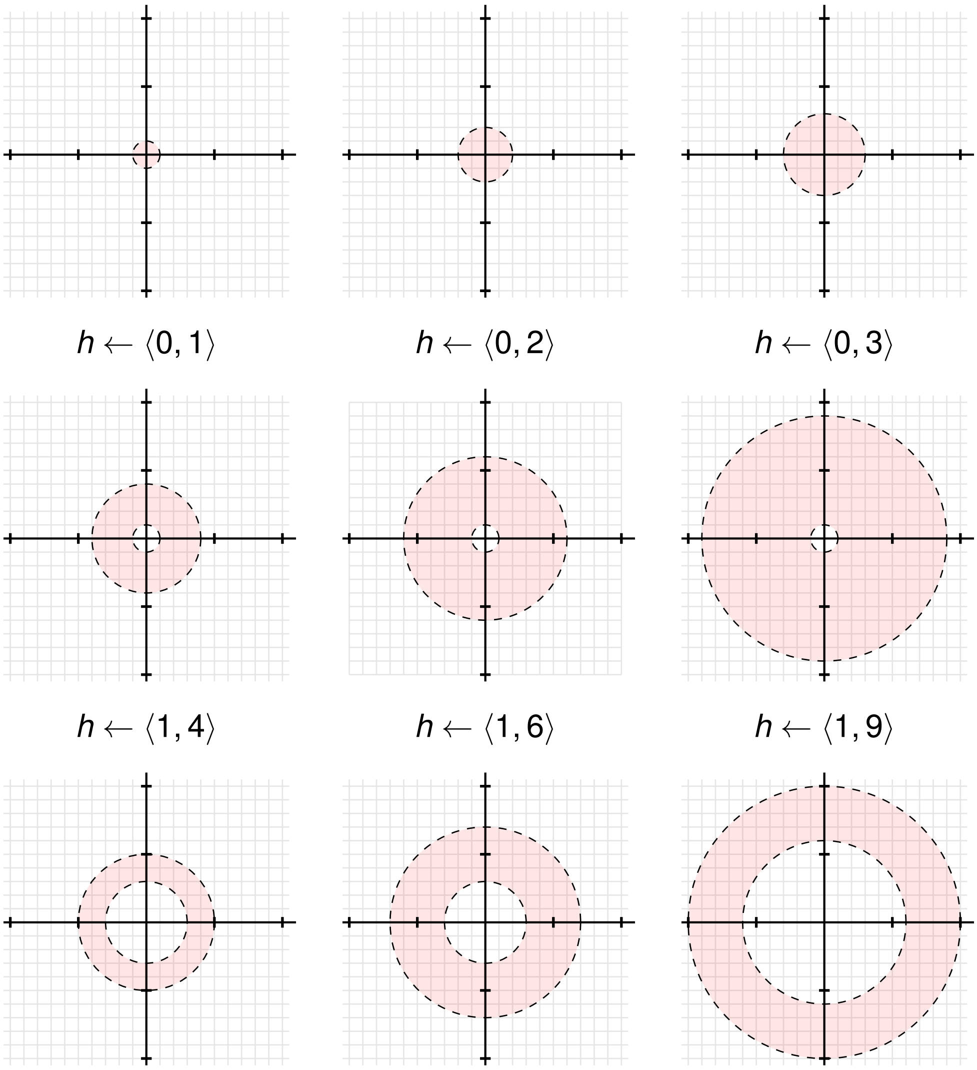
Assume (for purposes of explaining the answer) that the instance space is limited to $-10 \leq x, y \leq 10$. The hypothesis space will then be all the *donuts* that can be drawn with $0 \leq a \leq 10$ and $0 \leq b \leq 10$, with $a < b$ (remember that $a, b \in \mathbb{Z}$). A quick calculation will show that there are 55 possible hypotheses given this limited instance space. Figure 4 shows 9 examples of the 55 possible hypotheses.

If the instance space is not limited there are an infinite number of hypotheses. The final answer shows that the assumed limitation makes no difference. Most real-world problems have infinite instance- and search spaces. In general care need to be taken on the assumptions so that the models will still be valid.

To help understand some of the concepts, the complete set H_{55} of all 55 possible hypotheses is:

$$\begin{aligned} H_{55} = \{ \quad & \langle 0,1\rangle, \langle 0,2\rangle, \langle 0,3\rangle, \langle 0,4\rangle, \langle 0,5\rangle, \langle 0,6\rangle, \langle 0,7\rangle, \langle 0,8\rangle, \langle 0,9\rangle, \langle 0,10\rangle, \\ & \langle 1,2\rangle, \langle 1,3\rangle, \langle 1,4\rangle, \langle 1,5\rangle, \langle 1,6\rangle, \langle 1,7\rangle, \langle 1,8\rangle, \langle 1,9\rangle, \langle 1,10\rangle, \\ & \langle 2,3\rangle, \langle 2,4\rangle, \langle 2,5\rangle, \langle 2,6\rangle, \langle 2,7\rangle, \langle 2,8\rangle, \langle 2,9\rangle, \langle 2,10\rangle, \\ & \langle 3,4\rangle, \langle 3,5\rangle, \langle 3,6\rangle, \langle 3,7\rangle, \langle 3,8\rangle, \langle 3,9\rangle, \langle 3,10\rangle, \\ & \langle 4,5\rangle, \langle 4,6\rangle, \langle 4,7\rangle, \langle 4,8\rangle, \langle 4,9\rangle, \langle 4,10\rangle, \\ & \langle 5,6\rangle, \langle 5,7\rangle, \langle 5,8\rangle, \langle 5,9\rangle, \langle 5,10\rangle, \\ & \langle 6,7\rangle, \langle 6,8\rangle, \langle 6,9\rangle, \langle 6,10\rangle, \\ & \langle 7,8\rangle, \langle 7,9\rangle, \langle 7,10\rangle, \\ & \langle 8,9\rangle, \langle 8,10\rangle, \\ & \langle 9,10\rangle \} \end{aligned}$$

General-to-specific ordering of hypotheses In order to sequence the hypotheses from 'most specific' to 'most general' decide what is meant by 'more specific' and 'more general' (and 'less general' and 'less specific'). In other words decide how to define the *more_general_than_or_equal_to* and *more_general_than* relations.



$h \leftarrow \langle 0, 1 \rangle$ $h \leftarrow \langle 0, 2 \rangle$ $h \leftarrow \langle 0, 3 \rangle$

$h \leftarrow \langle 1, 4 \rangle$ $h \leftarrow \langle 1, 6 \rangle$ $h \leftarrow \langle 1, 9 \rangle$

$h \leftarrow \langle 3, 5 \rangle$ $h \leftarrow \langle 3, 7 \rangle$ $h \leftarrow \langle 6, 10 \rangle$

Figure 4: 9 examples of the 55 possible *donut* hypotheses.

Depending on the definition of these relations different answers will result. Assume that larger donuts are more general than smaller donuts. In this simple case this is a fair assumption to make since no more information about the data is available. Smaller donuts is also a valid choice. Look at three possible ways of defining the size of donuts:

1. The value of *a* - if *b* stays constant, smaller *a*-values produce larger donuts.
2. The value of *b* - if *a* stays constant, larger *b*-values produce larger donuts.
3. The surface area *A* of the donut - this is the obvious way of measuring one donut to be larger than another, but is slightly more complex to calculate.

The last two measures are simplified mechanisms to give an approximate measure of the relative sizes of donuts and is less complex to calculate.

Pick four hypotheses to show how these three measures work:

$$\begin{array}{llll}
h_{\langle 0,1\rangle} \leftarrow \langle 0,1\rangle & A=\pi & a=0 & b=1\\
h_{\langle 3,9\rangle} \leftarrow \langle 3,9\rangle & A=72\pi & a=3 & b=9\\
h_{\langle 4,9\rangle} \leftarrow \langle 4,9\rangle & A=65\pi & a=4 & b=9\\
h_{\langle 4,10\rangle} \leftarrow \langle 4,10\rangle & A=86\pi & a=4 & b=10
\end{array}$$

Let us look at the last two measures, since they are slightly simpler to calculate.

Smaller *a*-values Using *a* to rank the four chosen hypotheses from most specific to most general, the result is:

$$h_{\langle 0,1\rangle} <_g h_{\langle 3,9\rangle} <_g h_{\langle 4,9\rangle} =_g h_{\langle 4,10\rangle}$$

In other words, by this measure, the last two hypotheses are equally general.

Larger *b*-values Using *b* to rank the four chosen hypotheses from most specific to most general, gives:

$$h_{\langle 0,1\rangle} <_g h_{\langle 3,9\rangle} =_g h_{\langle 4,9\rangle} <_g h_{\langle 4,10\rangle}$$

In other words, by this measure, the middle two hypotheses are equally general.

Surface area Using the surface area of the donut is the intuitive choice (but is slightly more complex to calculate) and gives us the following order:

$$h_{\langle 0,1\rangle} <_g h_{\langle 3,9\rangle} <_g h_{\langle 4,9\rangle} <_g h_{\langle 4,10\rangle}$$

The obvious lesson here is that how ‘more specific’ and ‘more general’ is **measured** has an effect on the order of hypotheses. In other words, when comparing two hypotheses with each other, which one is ‘more general’ than the other one is determined by how specificity is measured. Three different measures were applied: smaller *a*-values, smaller *b*-values and the surface area of the donuts.

Using surface area as a criterion to determine whether a specific hypothesis is more general than another and apply it to the 55 hypotheses for this problem (and limited instance space) gives the sequence as in Table 1.

$$\begin{array}{rlll}
h_{\langle 0,1\rangle} : & A = \pi 1^2 - \pi 0^2 & = 1\pi \\
h_{\langle 1,2\rangle} : & A = \pi 2^2 - \pi 1^2 & = 3\pi \\
h_{\langle 0,2\rangle} : & A = \pi 2^2 - \pi 0^2 & = 4\pi \\
h_{\langle 2,3\rangle} : & A = \pi 3^2 - \pi 2^2 & = 5\pi \\
h_{\langle 3,4\rangle} : & A = \pi 4^2 - \pi 3^2 & = 7\pi \\
h_{\langle 1,3\rangle} : & A = \pi 3^2 - \pi 1^2 & = 8\pi \\
h_{\langle 0,3\rangle} : & A = \pi 3^2 - \pi 0^2 & = 9\pi \\
h_{\langle 4,5\rangle} : & A = \pi 5^2 - \pi 4^2 & = 9\pi \\
h_{\langle 2,4\rangle} : & A = \pi 4^2 - \pi 2^2 & = 11\pi \\
h_{\langle 5,6\rangle} : & A = \pi 6^2 - \pi 5^2 & = 11\pi \\
h_{\langle 6,7\rangle} : & A = \pi 7^2 - \pi 6^2 & = 13\pi \\
h_{\langle 1,4\rangle} : & A = \pi 4^2 - \pi 1^2 & = 15\pi \\
h_{\langle 7,8\rangle} : & A = \pi 8^2 - \pi 7^2 & = 15\pi \\
h_{\langle 0,4\rangle} : & A = \pi 4^2 - \pi 0^2 & = 16\pi \\
h_{\langle 3,5\rangle} : & A = \pi 5^2 - \pi 3^2 & = 16\pi \\
h_{\langle 8,9\rangle} : & A = \pi 9^2 - \pi 8^2 & = 17\pi \\
h_{\langle 9,10\rangle} : & A = \pi 10^2 - \pi 9^2 & = 19\pi \\
h_{\langle 4,6\rangle} : & A = \pi 6^2 - \pi 4^2 & = 20\pi \\
h_{\langle 2,5\rangle} : & A = \pi 5^2 - \pi 2^2 & = 21\pi \\
h_{\langle 1,5\rangle} : & A = \pi 5^2 - \pi 1^2 & = 24\pi \\
h_{\langle 5,7\rangle} : & A = \pi 7^2 - \pi 5^2 & = 24\pi \\
h_{\langle 0,5\rangle} : & A = \pi 5^2 - \pi 0^2 & = 25\pi \\
h_{\langle 3,6\rangle} : & A = \pi 6^2 - \pi 3^2 & = 25\pi \\
h_{\langle 6,8\rangle} : & A = \pi 8^2 - \pi 6^2 & = 28\pi \\
h_{\langle 7,9\rangle} : & A = \pi 9^2 - \pi 7^2 & = 32\pi \\
h_{\langle 2,6\rangle} : & A = \pi 6^2 - \pi 2^2 & = 32\pi \\
h_{\langle 4,7\rangle} : & A = \pi 7^2 - \pi 4^2 & = 33\pi \\
h_{\langle 1,6\rangle} : & A = \pi 6^2 - \pi 1^2 & = 35\pi \\
h_{\langle 0,6\rangle} : & A = \pi 6^2 - \pi 0^2 & = 36\pi \\
h_{\langle 8,10\rangle} : & A = \pi 10^2 - \pi 8^2 & = 36\pi \\
h_{\langle 3,7\rangle} : & A = \pi 7^2 - \pi 3^2 & = 38\pi \\
h_{\langle 5,8\rangle} : & A = \pi 8^2 - \pi 5^2 & = 39\pi \\
h_{\langle 6,9\rangle} : & A = \pi 9^2 - \pi 6^2 & = 45\pi
\end{array}$$

$h_{\langle 2,7\rangle}$:	$A = \pi 7^2 - \pi 2^2$	$= 45\pi$
$h_{\langle 1,7\rangle}$:	$A = \pi 7^2 - \pi 1^2$	$= 48\pi$
$h_{\langle 4,8\rangle}$:	$A = \pi 8^2 - \pi 4^2$	$= 48\pi$
$h_{\langle 0,7\rangle}$:	$A = \pi 7^2 - \pi 0^2$	$= 49\pi$
$h_{\langle 7,10\rangle}$:	$A = \pi 10^2 - \pi 7^2$	$= 51\pi$
$h_{\langle 3,8\rangle}$:	$A = \pi 8^2 - \pi 3^2$	$= 55\pi$
$h_{\langle 5,9\rangle}$:	$A = \pi 9^2 - \pi 5^2$	$= 56\pi$
$h_{\langle 2,8\rangle}$:	$A = \pi 8^2 - \pi 2^2$	$= 60\pi$
$h_{\langle 1,8\rangle}$:	$A = \pi 8^2 - \pi 1^2$	$= 63\pi$
$h_{\langle 0,8\rangle}$:	$A = \pi 8^2 - \pi 0^2$	$= 64\pi$
$h_{\langle 6,10\rangle}$:	$A = \pi 10^2 - \pi 6^2$	$= 64\pi$
$h_{\langle 4,9\rangle}$:	$A = \pi 9^2 - \pi 4^2$	$= 65\pi$
$h_{\langle 3,9\rangle}$:	$A = \pi 9^2 - \pi 3^2$	$= 72\pi$
$h_{\langle 5,10\rangle}$:	$A = \pi 10^2 - \pi 5^2$	$= 75\pi$
$h_{\langle 2,9\rangle}$:	$A = \pi 9^2 - \pi 2^2$	$= 77\pi$
$h_{\langle 1,9\rangle}$:	$A = \pi 9^2 - \pi 1^2$	$= 80\pi$
$h_{\langle 0,9\rangle}$:	$A = \pi 9^2 - \pi 0^2$	$= 81\pi$
$h_{\langle 4,10\rangle}$:	$A = \pi 10^2 - \pi 4^2$	$= 86\pi$
$h_{\langle 3,10\rangle}$:	$A = \pi 10^2 - \pi 3^2$	$= 91\pi$
$h_{\langle 2,10\rangle}$:	$A = \pi 10^2 - \pi 2^2$	$= 96\pi$
$h_{\langle 1,10\rangle}$:	$A = \pi 10^2 - \pi 1^2$	$= 99\pi$
$h_{\langle 0,10\rangle}$:	$A = \pi 10^2 - \pi 0^2$	$= 100\pi$

Table 1: The 55 hypotheses, ordered from most specific (top) to most general (bottom).

Notice that there are some hypotheses that are equally general to one or more other hypotheses, but that most are strictly more specific or more general than any other hypotheses. For almost all problems that are encountered (even textbook examples) it is impossible to enumerate the hypotheses like this.

Keep in mind that none of the hypotheses have been tested to determine if it explains the data. This is a simple ordering of hypotheses without considering the instances. Figure 5 shows an incomplete diagram of the hypothesis space using surface area a criteria for ordering the hypotheses from most general to most specific. In the discussion that follows the surface area measure is used.

FIND-S There is now enough information to attempt to find the maximally specific hypothesis. FIND-S ignores all negative instances so only the four positive instances need to be considered.

The sequence of instances should not effect the outcome of the algorithm if there are no errors in the data. After observing the first positive instance $(5, 5)$ pick the smallest donut that will contain the instance (remember that smaller donuts are more specific), which turns out to be (and is shown in Figure 6)

$$h \leftarrow \langle 7, 8\rangle$$

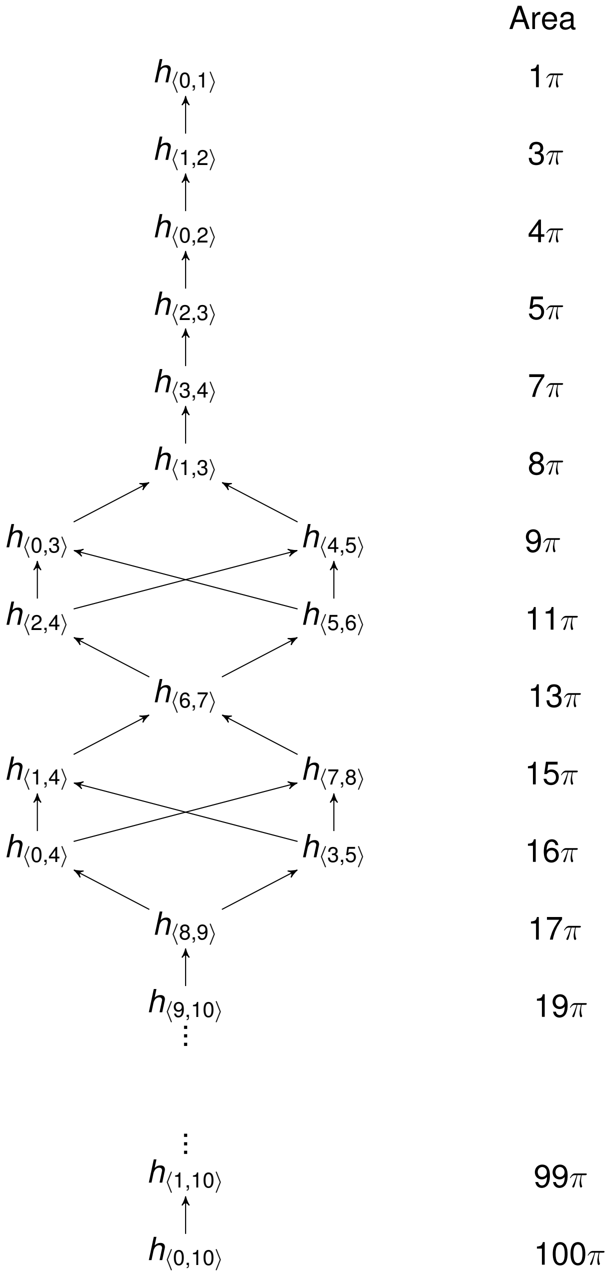

Area
$h_{\langle 0,1\rangle}$ 1π
$h_{\langle 1,2\rangle}$ 3π
$h_{\langle 0,2\rangle}$ 4π
$h_{\langle 2,3\rangle}$ 5π
$h_{\langle 3,4\rangle}$ 7π
$h_{\langle 1,3\rangle}$ 8π
$h_{\langle 0,3\rangle}$ $h_{\langle 4,5\rangle}$ 9π
$h_{\langle 2,4\rangle}$ $h_{\langle 5,6\rangle}$ 11π
$h_{\langle 6,7\rangle}$ 13π
$h_{\langle 1,4\rangle}$ $h_{\langle 7,8\rangle}$ 15π
$h_{\langle 0,4\rangle}$ $h_{\langle 3,5\rangle}$ 16π
$h_{\langle 8,9\rangle}$ 17π
$h_{\langle 9,10\rangle}$ 19π
$h_{\langle 1,10\rangle}$ 99π
$h_{\langle 0,10\rangle}$ 100π

Figure 5: Partial hypothesis space using surface area as criterion. Area shown on the right.

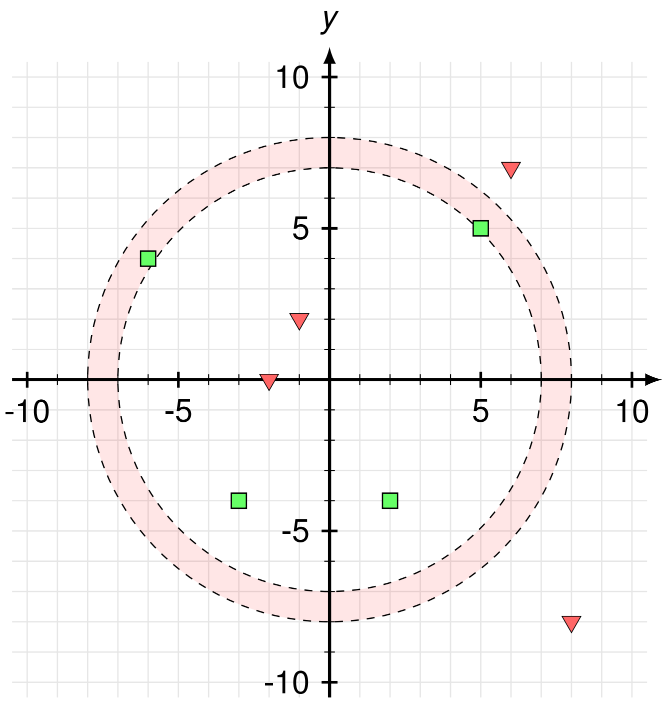


Figure 6: FIND-S after the first and second instances $(5, 5)$ and $(-6, 4)$ produce $h \leftarrow \langle 7, 8\rangle$

After observing the second positive instance $(-6, 4)$ the current hypotheses still explains the data and the hypothesis remains the same.

After observing the third positive instance $(-3, -4)$ the current hypothesis is not general enough and needs to be expanded to a larger donut. Keep in mind that in the chosen definition of the donut the edge of the donut is not included in the hypotheses. This gives:

$$h \leftarrow \langle 4, 8\rangle$$

The fourth positive instance does not require any further expansion of the hypotheses. The maximally specific hypothesis for this instance space is therefore:

$$h \leftarrow \langle 4, 8\rangle$$

with $A = 48\pi$.

FIND-S does not find all the hypotheses that are consistent with the data. In the example, FIND-S only finds the smallest possible donut that is consistent with the data (i.e. includes all the positive instances). There may be more donuts that explain the data.

List-Then-Eliminate Since all the hypotheses are known for this example this algorithm can be applied to find all the consistent hypotheses for this instance space. The result are the following four hypotheses (shown in Figure 8):

$$h_1 \leftarrow \langle 3, 8\rangle$$

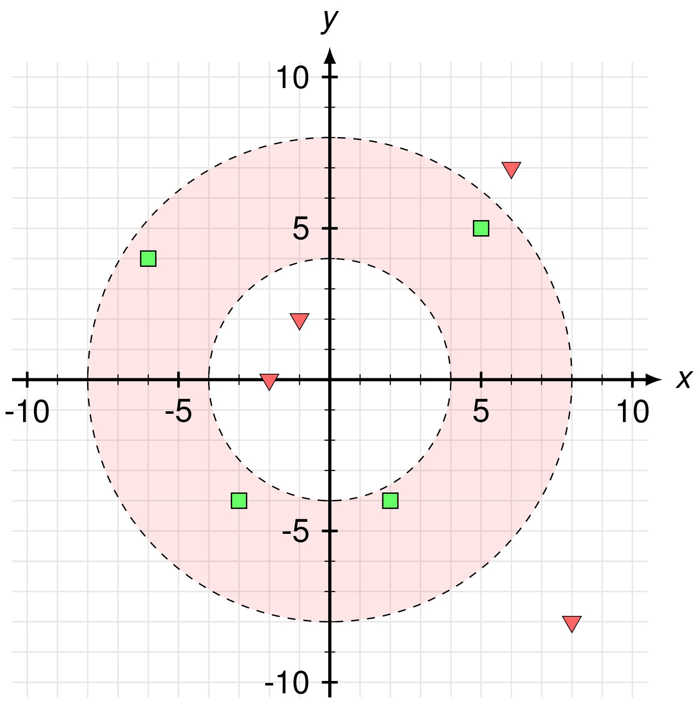


Figure 7: FIND-S after the third and fourth instances $(-3,-4)$ and $(2,-4)$ produce $h \leftarrow \langle 4, 8\rangle$

$$h_2 \leftarrow \langle 4, 8\rangle$$
$$h_3 \leftarrow \langle 3, 9\rangle$$
$$h_4 \leftarrow \langle 4, 9\rangle$$

FIND-G In a similar fashion to the FIND-S algorithm calculate the maximally general hypothesis.

The algorithm is initiated with the most general hypothesis and after observing each *negative* instance the hypothesis is made more specific until the algorithm terminates with the maximally general hypothesis. Note that positive instances have no effect on the hypotheses in FIND-G, just as negative instances play no role in FIND-S.

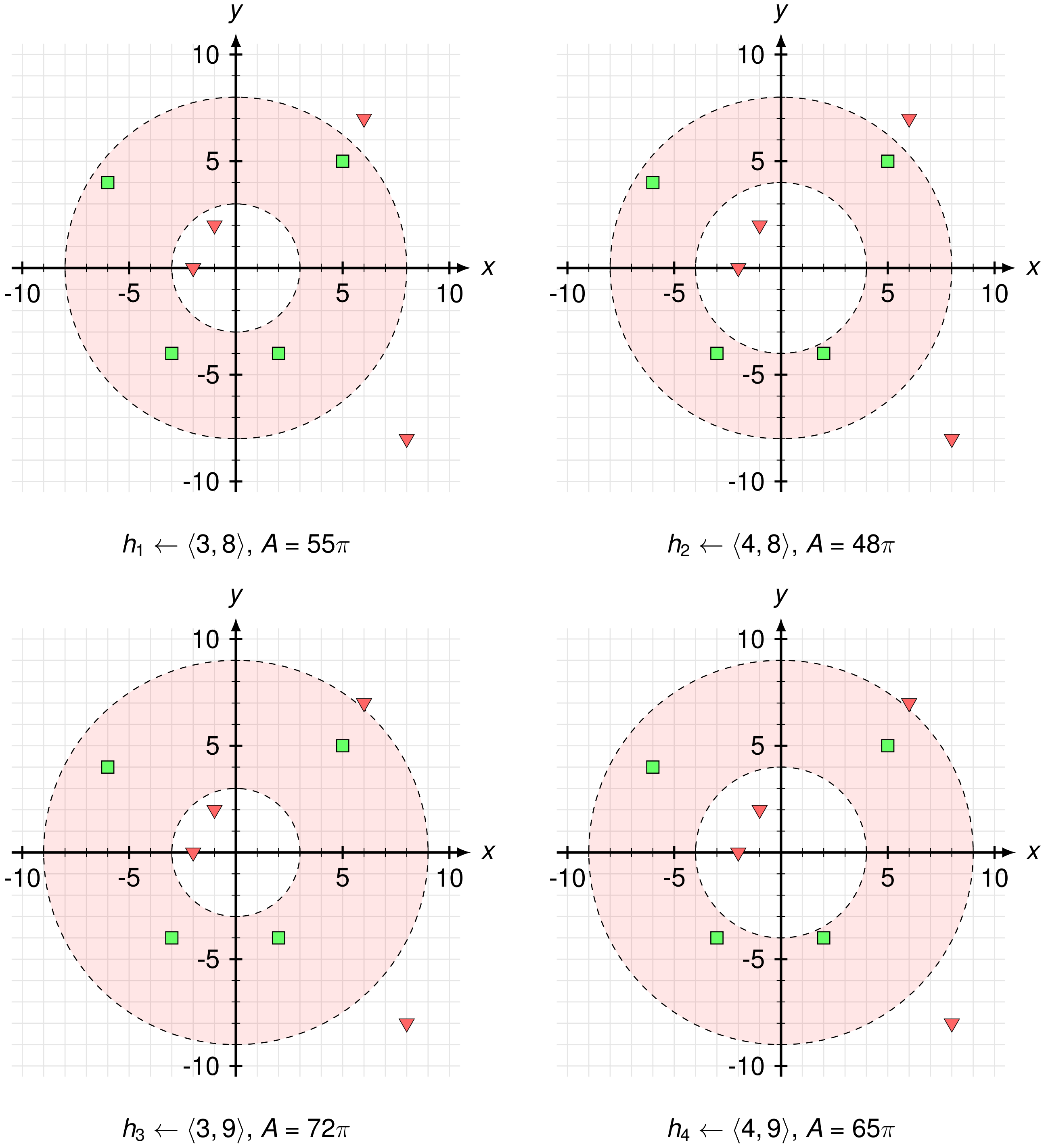
Start with the most general hypothesis, in this case $h \leftarrow \langle 0, 10\rangle$.

After observing the first negative instance $(-1, 2)$ pick the largest donut that will contain the instance (remember that larger donuts are more general), which turns out to be

$$h \leftarrow \langle 3, 10\rangle$$

and is shown in Figure 9.

After observing the second negative instance $(-2, 0)$ the current hypotheses still explains the data and the hypothesis remains the same.



$h_1 \leftarrow \langle 3,8 \rangle$, $A = 55\pi$

$h_2 \leftarrow \langle 4,8 \rangle$, $A = 48\pi$

$h_3 \leftarrow \langle 3,9 \rangle$, $A = 72\pi$

$h_4 \leftarrow \langle 4,9 \rangle$, $A = 65\pi$

Figure 8: Find-Then-Eliminate results in four consistent hypotheses.

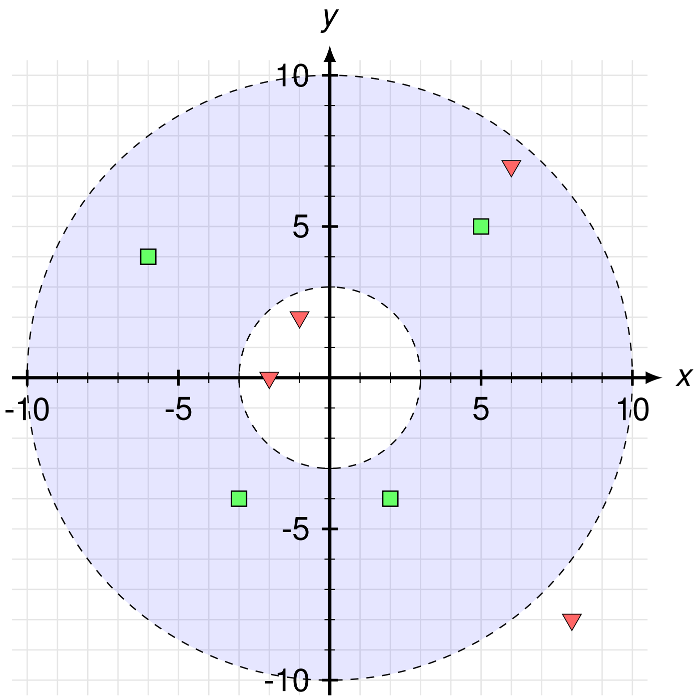


Figure 9: FIND-G after the first and second instances $(-1, 2)$ and $(-2, 0)$ produce $h \leftarrow \langle 3, 10 \rangle$

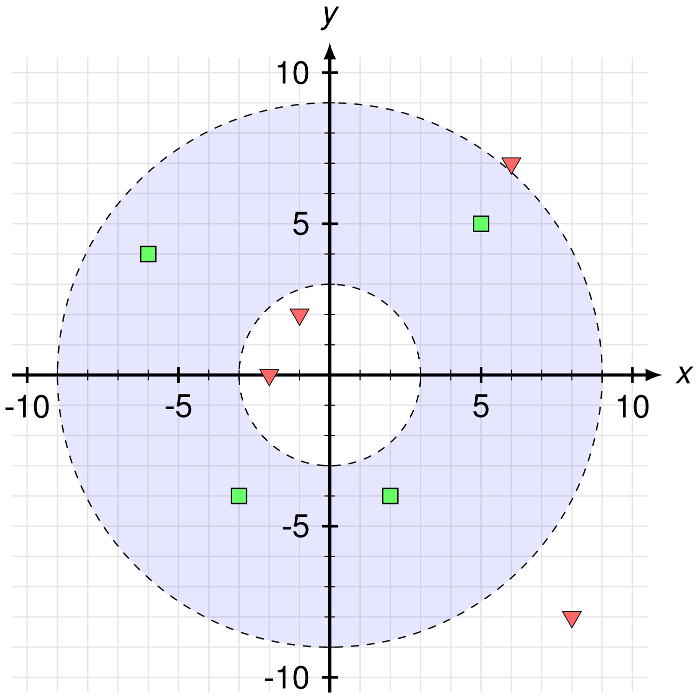


Figure 10: FIND-G after the third and fourth instances $(6, 7)$ and $(8, -8)$ produce $h \leftarrow \langle 3, 9 \rangle$

After observing the third negative instance $(6, 7)$ the current hypothesis is not specific enough and needs to be contracted to a smaller donut. The result is:

$$h \leftarrow \langle 3, 9 \rangle$$

The fourth negative instance does not require any further contraction of the hypotheses. The maximally general hypothesis for this instance space is therefore:

$$h \leftarrow \langle 3, 9 \rangle$$

with $A = 72\pi$. This means that this is also the *maximally general hypothesis* for this instance space.

***S*- and *G*-boundary sets** Since there are only one maximally specific hypothesis and one maximally general hypothesis for this instance space this means that the *S*- and *G*-boundary sets each contain one element, namely:

$$S = \{\langle 4, 8 \rangle\}$$
$$G = \{\langle 3, 9 \rangle\}$$

The *S*- and *G*-boundary sets are shown in Figure 11.

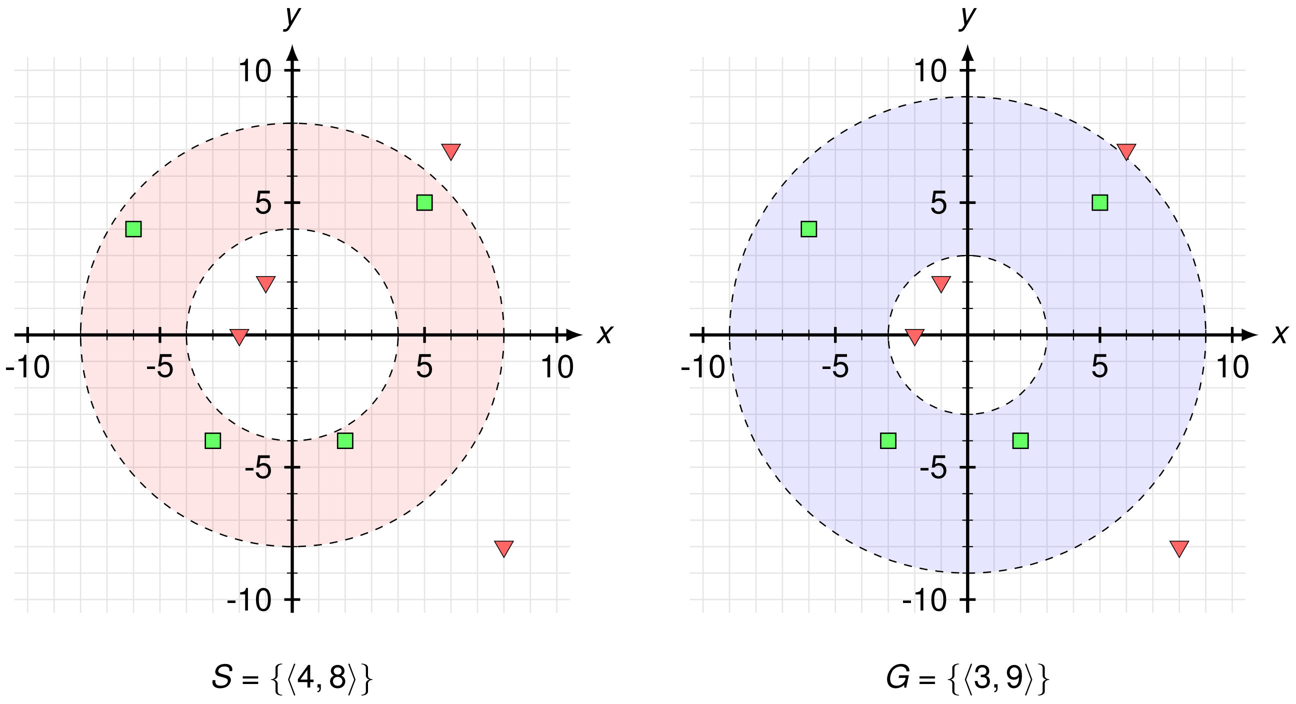


Figure 11: *S*- and *G*-boundary sets.

Discussion on (c)

Any new instance that would not fit into any current hypothesis that explains the data will force a change in the S- or G-boundary sets. For example, a new positive instance at $(-3, -2)$ will force the S-boundary to expand to $S = \{\langle 3, 8\rangle\}$. Similarly a new negative instance at $(5, 7)$ will shrink the G-boundary set to $G = \{\langle 3, 8\rangle\}$.

Discussion on (d)

There are many possible alternative hypotheses (actually an infinite number). The following discussion shows some possibilities.

Origin centered rectangular 'donuts' Define an origin centered rectangular donut as a rectangular area with a smaller rectangular area inside subtracted from the first, where the center of the rectangles coincide with the origin of the coordinate system. A hypothesis would then take the form:

$$h \leftarrow (a < |x| < b) \wedge (c < |y| < d)$$

where $a, b, c, d \in \mathbb{Z}$.

One (there are more) specific hypothesis that will explain the data is:

$$h \leftarrow (4 < |x| < 7) \wedge (3 < |y| < 6)$$

which can be shortened to

$$h \leftarrow \langle 4, 7, 3, 6\rangle$$

This hypothesis is illustrated in Figure 12. This form of hypothesis uses four parameters instead of the two for the normal donuts, hence having some effect on the computational complexity of finding a solution.

Figures 13 to 15 show a few more examples of possible hypotheses. Try to define these hypotheses formally as was done with the donuts and the rectangular 'donuts'.

Further notes

Note that all the hypotheses discussed so far had the edges excluded. This was simply a choice of how the hypothesis was defined and the edges may as well have been included. *Donuts* with the edges included are defined as:

$$h \leftarrow \langle a \leq \sqrt{x^2 + y^2} \leq b\rangle$$

where $a < b$ and $a, b \in \mathbb{Z}$. Figure 16 shows such a hypothesis with the instance $(-3, -4)$ falling on the edge of the hypothesis. Using solid or dashed lines is a convention to indicate whether or not the edge is included in the hypothesis or not.

This choice of hypothesis will have an effect in the FIND-S algorithm. Considering the third negative instance, the hypotheses becomes

$$h \leftarrow \langle 5, 8\rangle$$

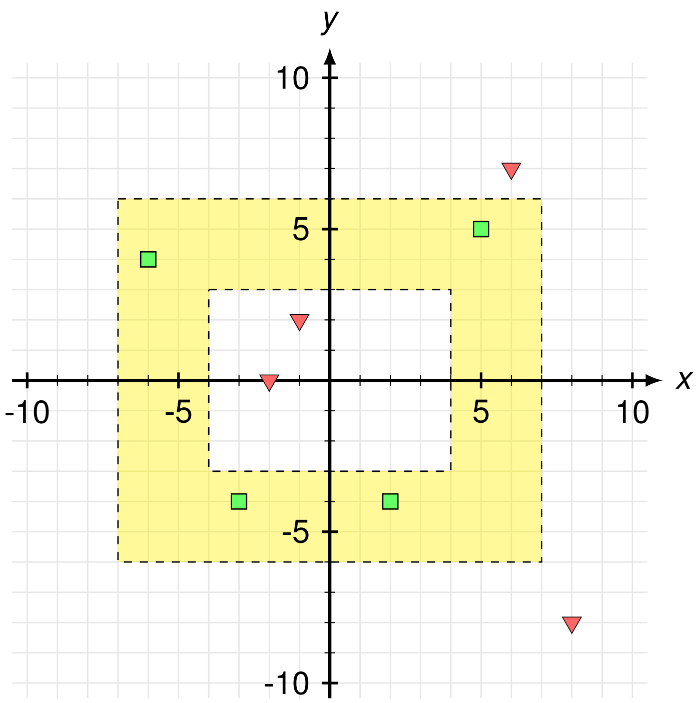


Figure 12: First alternative hypothesis: rectangular *donuts*.

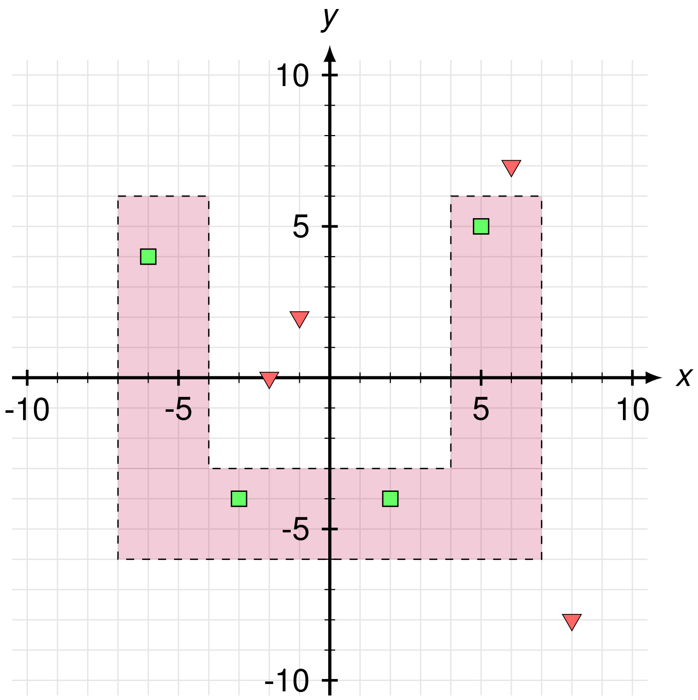


Figure 13: Second alternative hypothesis: U-shaped areas, with edges parallel to the axes.

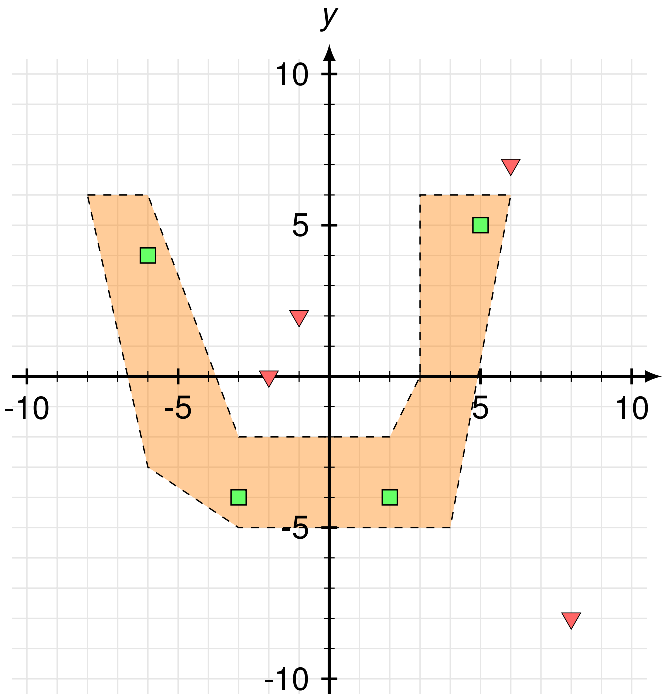


Figure 14: Third alternative hypothesis: arbitrarily-shaped areas.

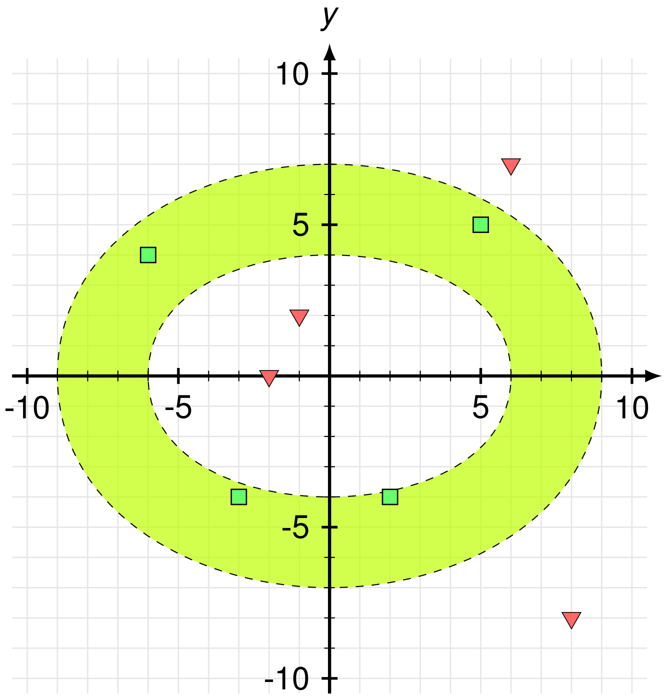


Figure 15: Fourth alternative hypothesis. Origin-centered oval donuts. This requires four parameters.

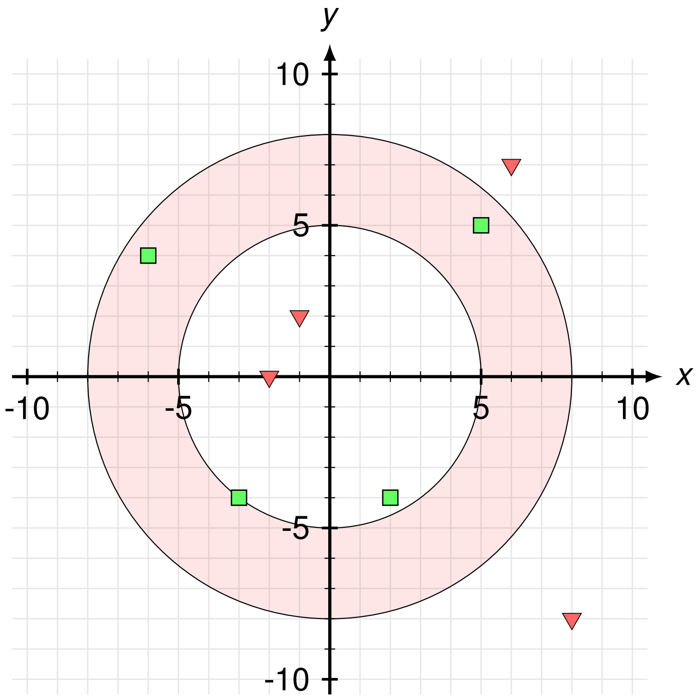


Figure 16: Donut hypothesis with edges included.

However, the fourth negative instance expands the hypothesis and the maximally specific hypothesis remains the same as before.

Further exercises

Here are some exercises you could try to make sure that you understood the concepts discussed and are able to do the calculations.

- Complete the version space diagram (Figure 5). Find the *S*- and *G*-boundary sets of the hypotheses in this diagram.
- Find the *S*- and *G*-boundaries using the 'smaller *a*' criterion for origin-centered donuts.
- Calculate and draw the version space diagram for the 'smaller *a*' and the 'larger *b*'criteria.
- Find the *S*- and *G*-boundaries using the 'larger *b*' criterion for origin-centered donuts.
- Repeat the exercises above using one of the alternative hypotheses.
- What are the pros and cons of the different measures by which specificity is measured?
- How are the first and second alternative hypotheses equivalent?

Mark out of 100.
40 or less for clear indication that student does not understand the topic or evidence of plagiarism, or answers are correct, but have not shown complete workings
50 correct and sufficient workings
60-70 correct and complete workings
80+ indicating thorough understanding of the work

Question 5

Give binary decision trees to represent the following Boolean functions:

(a) $f_1(A, B) = \neg A \wedge B$

(b) $f_2(A, B, C) = [A \wedge B] \vee C]$

(c) $f_3(A, B) = A \veebar B$

(d) $f_4(A, B, C, D) = [A \vee B] \wedge [C \vee D]$

Remember that there is a difference between a graph and a tree.

Read: `https://www.geeksforgeeks.org/difference-between-graph-and-tree/`

The symbol $\veebar$ represents the Boolean operator for XOR (exclusive-or). For this exercise you do not need to do the Gain or Entropy calculations. There is a direct mapping between a Boolean function and its corresponding binary decision tree. The binary decision tree can usually by simplified as well to produce a simpler, more compact tree. Do not just write down the final, simplified tree. Show how you do the simplification.

Here are resources you could consult on this topic - they are also a good introduction to material for the next question:

- `https://www.cs.cmu.edu/~fp/courses/15122-f10/lectures/19-bdds.pdf`
- `http://cs.nyu.edu/~dsontag/courses/ml12/slides/lecture11.pdf`

When constructing a decision tree (binary decision trees included) the choice of variable to use as the root node (and any subsequent nodes) effect the structure, accuracy and simplicity of the tree. This is the most important feature of decision tree learning algorithms, such as ID3. They choose the best variable at each node. In this question DO NOT use any algorithms. The aim of here is to learn that there is a direct correlation between a Boolean function and a binary decision tree.

Discussion on (a)

Boolean function given:

$$f_1(A, B) = \neg A \wedge B$$

The truth table for this Boolean function is given in Table 2.

Start by choosing A as the root node. This gives us the binary decision tree as in Figure 17. On the diagram you can see the mapping between specific parts of the truth table and the binary decision tree. Each leaf node corresponds to one row in the truth table, while the level above the leaf nodes correspond to two rows in the truth table, etc. By merging leaf nodes with the same value the tree can be simplified, as in Figure 18.

Using B as the start node results in a different binary decision tree. In this particular case the tree turns out to be as simple as the first. This is not the case for all decision trees.

The binary decision tree starting with B is shown in Figure 19.

A	B	$\neg A$	f_1
0	0	1	0
0	1	1	1
1	0	0	0
1	1	0	0

Table 2: Truth table for f_1.

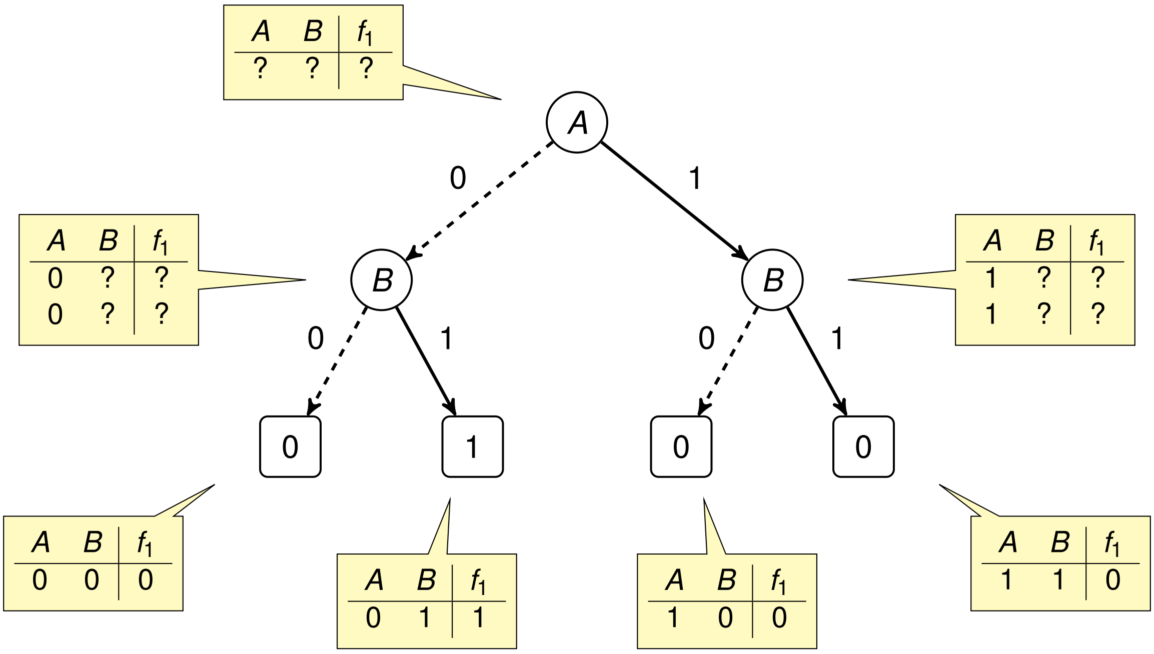


Figure 17: A binary decision tree for f_1 starting with A.

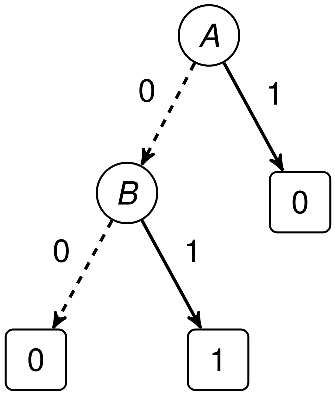


Figure 18: A simplified binary decision tree for $f_1(A, B) = \neg A \land B$ starting with A.

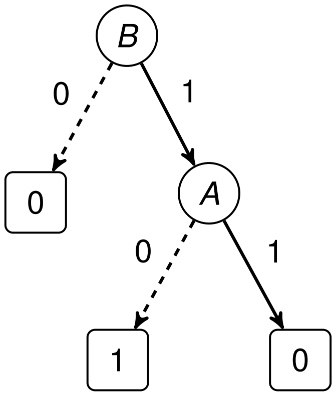


Figure 19: A binary decision tree for f_1 starting with B.

Discussion on (b)

Boolean function given:

$$f_2(A, B, C) = [A \wedge B] \vee C]$$

The truth table for this Boolean function is given in Table 3.

A	B	C	$A \wedge B$	f_2
0	0	0	0	0
0	0	1	0	1
0	1	0	0	0
0	1	1	0	1
1	0	0	0	0
1	0	1	0	1
1	1	0	1	1
1	1	1	1	1

Table 3: Truth table for f_2.

Choose A for the root node. This produces the binary decision tree as in Figure 20.

A simplified binary decision tree for f_2 is shown in Figure 21. Work out the different combinations of binary decision trees that are possible and see which result in the most compact tree when they are simplified. See if you agree that the simplified tree shown for f_2 is the simplest possible tree or not.

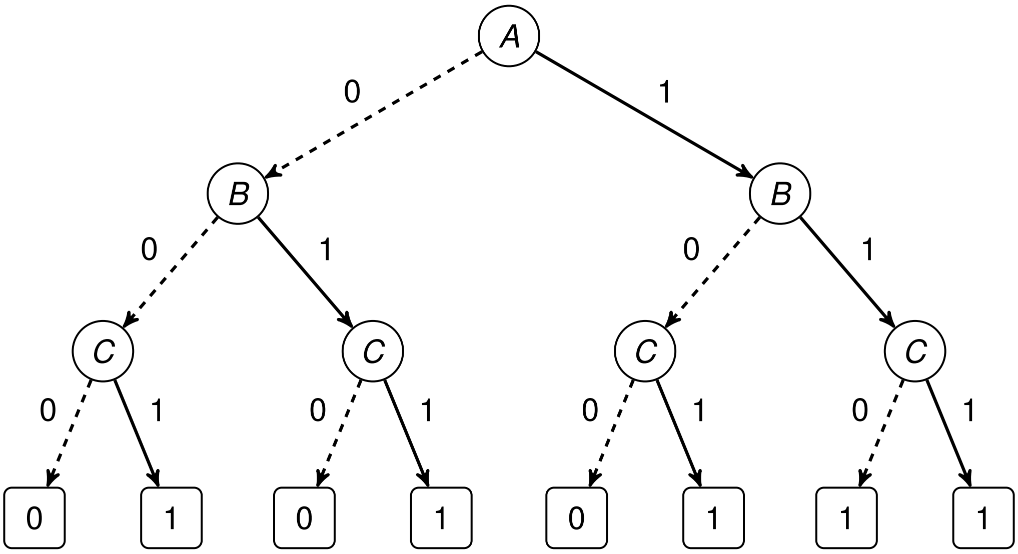


Figure 20: A binary decision tree for f_2 starting with *A*.

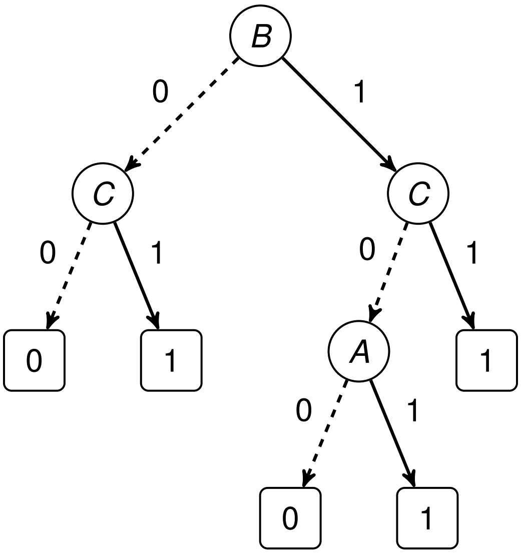


Figure 21: A simplified binary decision tree for f_2, using node sequence *B*, *C*, *A*.

Discussion on (c)

Boolean function given:

$$f_3(A, B) = A \veebar B$$

The truth table for this Boolean function is given in Table 4.

A	B	f_3
0	0	0
0	1	1
1	0	1
1	1	0

Table 4: Truth table for f_3.

Choosing A for the root node, produces the binary decision tree as in Figure 22.

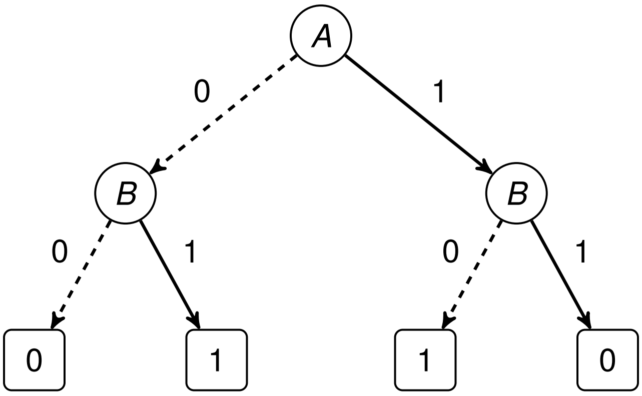


Figure 22: A binary decision tree for f_3 starting with A.

Similarly, the binary decision tree for f_3 starting with B produces the tree as in Figure 23.

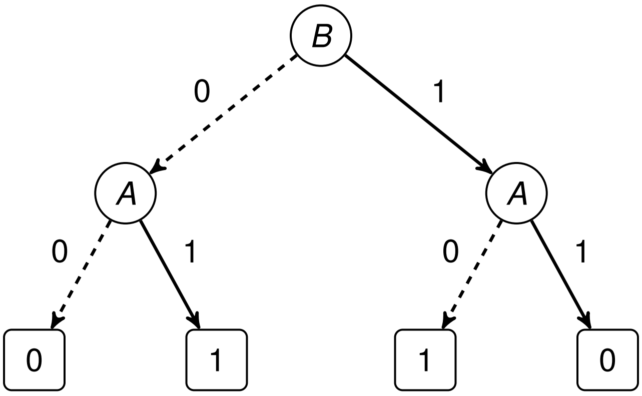


Figure 23: A binary decision tree for f_3 starting with B.

What is immediately apparent from these trees are that they cannot be simplified further.

Discussion on (d)

Boolean function given:

$$f_4(A, B, C, D) = [A \vee B] \wedge [C \vee D]$$

The truth table for this Boolean function is given in Table 5.

A	B	C	D	$[A \vee B]$	$[C \vee D]$	f_4
0	0	0	0	0	0	0
0	0	0	1	0	1	0
0	0	1	0	0	1	0
0	0	1	1	0	1	0
0	1	0	0	1	0	0
0	1	0	1	1	1	1
0	1	1	0	1	1	1
0	1	1	1	1	1	1
1	0	0	0	1	0	0
1	0	0	1	1	1	1
1	0	1	0	1	1	1
1	0	1	1	1	1	1
1	1	0	0	1	0	0
1	1	0	1	1	1	1
1	1	1	0	1	1	1
1	1	1	1	1	1	1

Table 5: Truth table for f_4.

Start by choosing A for the root node. This produces the binary decision tree as in Figure 24. This tree has not been simplified. Figures 25 to 27 show the steps in simplifying this specific decision tree. Notice that the last tree has two identical sub-trees, but it cannot be further simplified.

Look at the results of the next question, and compare the two trees. What do you see?

Mark out of 100.
40 or less for clear indication that student does not understand the topic or evidence of plagiarism, or answers are correct, but have not shown complete workings
50 correct and sufficient workings
60-70 correct and complete workings
80+ indicating thorough understanding of the work

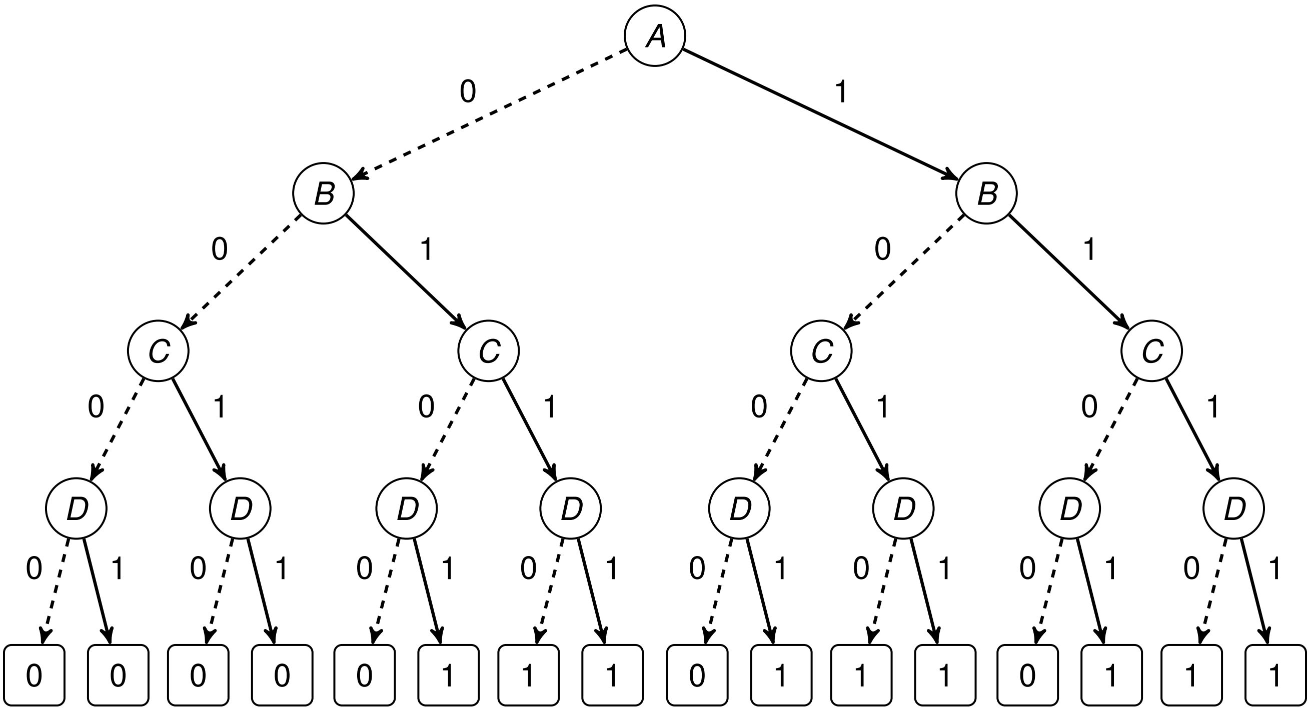


Figure 24: A binary decision tree for f_4 starting with *A*.

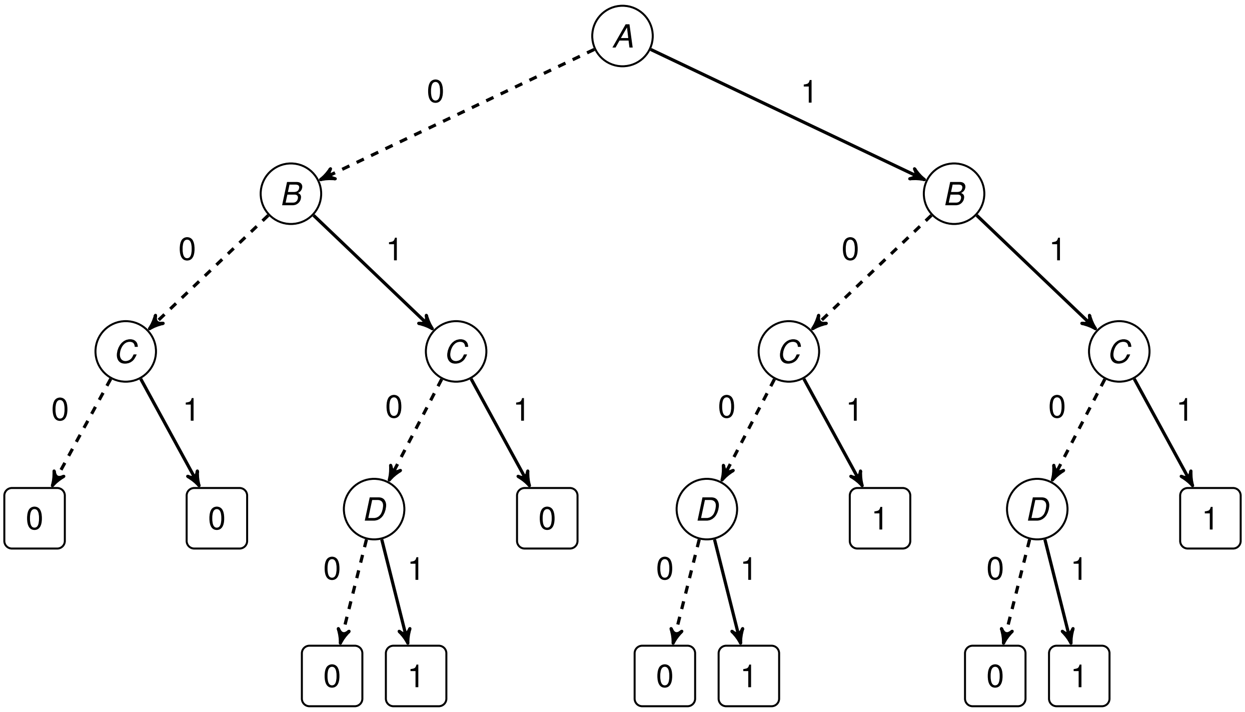


Figure 25: Step 1 in simplifying the binary decision tree for f_4.

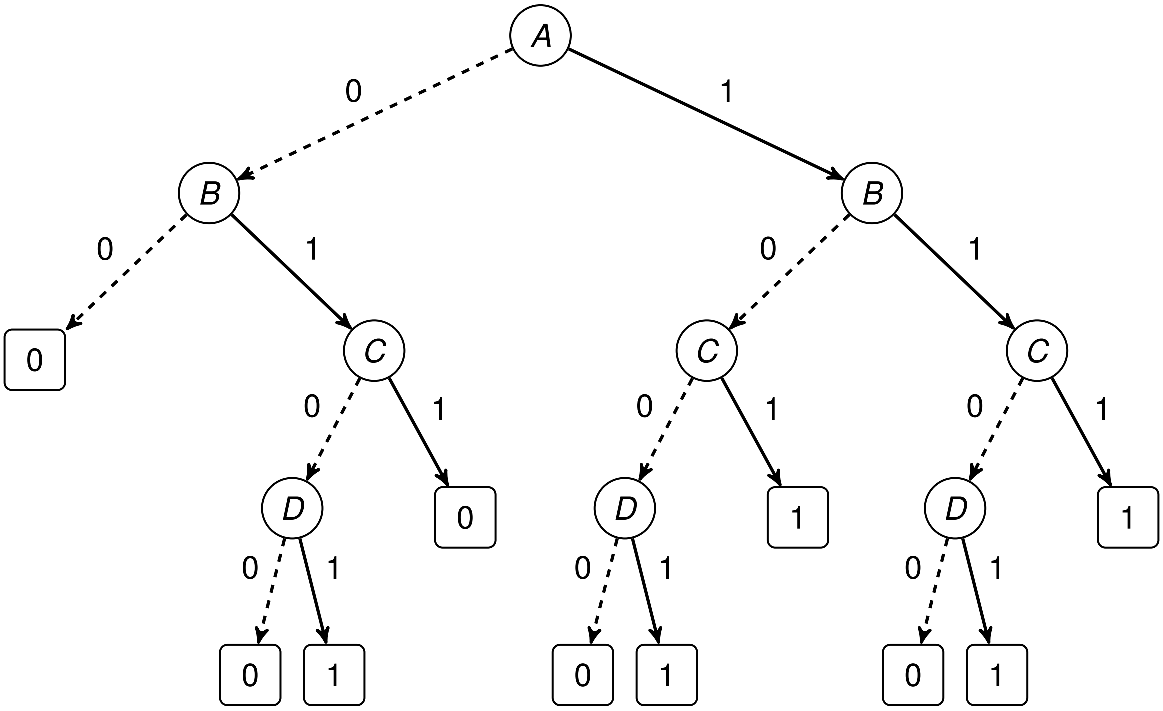


Figure 26: Step 2 in simplifying the binary decision tree for f_4.

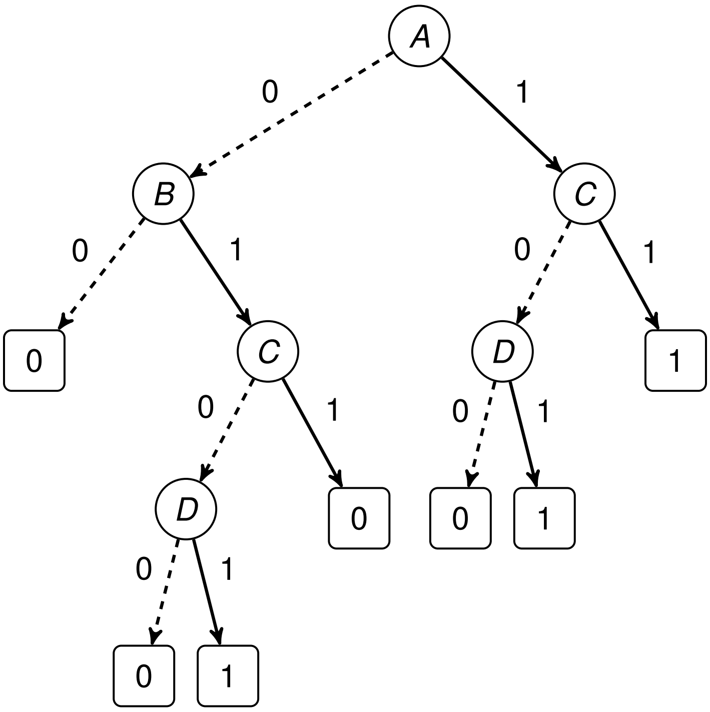


Figure 27: Step 3 in simplifying the binary decision tree for f_4.

Question 6

Use the ID3 algorithm to construct a decision tree for the data in Table 6. Show all your calculations, including all the steps of the Gain and Entropy calculations. Show the formulas that you use. Clearly explain your choices.

A	B	C	D	f_5
F	F	F	F	*no*
F	F	F	T	*no*
F	F	T	F	*no*
F	F	T	T	*no*
F	T	F	F	*no*
F	T	F	T	*yes*
F	T	T	F	*yes*
F	T	T	T	*yes*
T	F	F	F	*no*
T	F	F	T	*yes*
T	F	T	F	*yes*
T	F	T	T	*yes*
T	T	F	F	*no*
T	T	F	T	*yes*
T	T	T	F	*yes*
T	T	T	T	*yes*

Table 6: Truth table for f_5.

Here are some resources you could consult on this topic (focus on the ID3 algorithm):

- `https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-`
- `https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm`
- `http://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf`
- `https://cis.temple.edu/~ingargio/cis587/readings/id3-c45.html`

Discussion

Table 6 shows that the target function, f_5 can take on two possible values, *yes* or *no*. There are 16 examples, of which 7 result in $f_5 = no$ and 9 gives $f_5 = yes$, in other words: $S \equiv [7_{no}, 9_{yes}]$ There are 4 attributes (A, B, C, D) whose combination of values determine the value of the target attribute, f_5.

Calculate the *entropy* of the data set:

$$\begin{aligned}
Entropy(S) &\equiv \sum_{i=1}^{c} -p_i log_2(p_i) \\
&= -p_{no}\ log_2(p_{no}) - p_{yes}\ log_2(p_{yes}) \\
&= -{}^{7}/_{16}\ log_2({}^{7}/_{16}) - {}^{9}/_{16}\ log_2({}^{9}/_{16}) \\
&= 0.522 + 0.467 \\
&= 0.989
\end{aligned}$$

An Entropy value close to 1 means that the data is very well structured. Most real-world data sets are sparse and not always as well structured, which will result in lower Entropy values. Completely random data should have an Entropy value of 0.

Next, calculate the *Information Gain* of the subsets of the data, as it is sorted by the values each attribute can take.

The attribute *A* can take on two values: *F* or *T*.

$$\begin{aligned} Values(A) &= F, T \\ S_A &= [7_{no}, 9_{yes}] \\ S_{A=F} &\leftarrow [5_{no}, 3_{yes}] \\ S_{A=T} &\leftarrow [2_{no}, 6_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute *A*:

$$\begin{aligned} Entropy(S_{no}) &= -{}^5\!/\!_8\, log_2({}^5\!/\!_8) - {}^3\!/\!_8\, log_2({}^3\!/\!_8) \\ &= 0.954 \\ Entropy(S_{yes}) &= -{}^2\!/\!_8\, log_2({}^2\!/\!_8) - {}^6\!/\!_8\, log_2({}^6\!/\!_8) \\ &= 0.811 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of *A*:

$$\begin{aligned} Gain(S, A) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S) - {}^7\!/\!_{16}\, Entropy(S_{no}) - {}^9\!/\!_{16}\, Entropy(S_{yes}) \\ &= 0.989 - {}^7\!/\!_{16} \times 0.954 - {}^9\!/\!_{16} \times 0.811 \\ &= 0.115 \end{aligned}$$

When we look at the divisions of the dataset using the other three attributes, *B*, *C* and *D*, we see that the divisions are exactly the same.

$$\begin{aligned} Values(B) &= F, T \\ S_B &= [7_{no}, 9_{yes}] \\ S_{B=F} &\leftarrow [5_{no}, 3_{yes}] \\ S_{B=T} &\leftarrow [2_{no}, 6_{yes}] \end{aligned}$$

$$\begin{aligned} Values(C) &= F, T \\ S_C &= [7_{no}, 9_{yes}] \\ S_{C=F} &\leftarrow [5_{no}, 3_{yes}] \\ S_{C=T} &\leftarrow [2_{no}, 6_{yes}] \end{aligned}$$

$$\begin{aligned}
Values(D) &= F, T \\
S_D &= [7_{no}, 9_{yes}] \\
S_{D=F} &\leftarrow [5_{no}, 3_{yes}] \\
S_{D=T} &\leftarrow [2_{no}, 6_{yes}]
\end{aligned}$$

This means that the *Information Gain* values will also be the same for these attributes:

$$\begin{aligned}
Gain(S, B) &= 0.115 \\
Gain(S, C) &= 0.115 \\
Gain(S, D) &= 0.115
\end{aligned}$$

Since there is no single attribute which has the highest gain value, we can choose any one as the root node of the tree. Choose *A* as the root node.

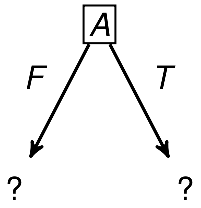


Figure 28: Leaf node of the decision tree for f_5.

The dataset is now divided into two subsets, using the values of attribute A, as in Tables 7 and 8.

A	B	C	D	f_5
F	F	F	F	*no*
F	F	F	T	*no*
F	F	T	F	*no*
F	F	T	T	*no*
F	T	F	F	*no*
F	T	F	T	*yes*
F	T	T	F	*yes*
F	T	T	T	*yes*

Table 7: Truth table for $f_{5,A=F}$

A	B	C	D	f_5
T	F	F	F	*no*
T	F	F	T	*yes*
T	F	T	F	*yes*
T	F	T	T	*yes*
T	T	F	F	*no*
T	T	F	T	*yes*
T	T	T	F	*yes*
T	T	T	T	*yes*

Table 8: Truth table for $f_{5,A=T}$

Therefore:

$$S_{A=F} \equiv [5_{no}, 3_{yes}]$$

and

$$S_{A=T} \equiv [2_{no}, 6_{yes}]$$

Calculate the *entropy* of the data set in Table 7, where $A = F$:

$$\begin{aligned} Entropy(S_{A=F}) &\equiv \sum_{i=1}^{c} -p_i log_2(p_i) \\ &= -p_{no}\ log_2(p_{no}) - p_{yes}\ log_2(p_{yes}) \\ &= -{}^5\!/\!_8\ log_2({}^5\!/\!_8) - {}^3\!/\!_8\ log_2({}^3\!/\!_8) \\ &= 0.954 \end{aligned}$$

The attribute B can take on two values: F or T.

$$\begin{aligned} Values(B) &= F, T \\ S_{A=F} &= [5_{no}, 3_{yes}] \\ S_{A=F,B=F} &\leftarrow [4_{no}, 0_{yes}] \\ S_{A=F,B=T} &\leftarrow [1_{no}, 3_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute B, when $A = F$:

$$\begin{aligned}
Entropy(S_{A=F,no}) &= -{}^{4}\!/_{4}\ log_2({}^{4}\!/_{4}) - {}^{0}\!/_{4}\ log_2({}^{0}\!/_{4}) \\
&= 0 \\
Entropy(S_{A=F,yes}) &= -{}^{1}\!/_{3}\ log_2({}^{1}\!/_{3}) - {}^{2}\!/_{3}\ log_2({}^{2}\!/_{3}) \\
&= 0.918
\end{aligned}$$

Calculate the information gained when dividing the data by using the values of B, when $A = F$:

$$\begin{aligned}
Gain(S_{A=F,B}) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S_{A=F}) - {}^{5}\!/_{8}\ Entropy(S_{A=F,no}) - {}^{3}\!/_{8}\ Entropy(S_{A=F,yes}) \\
&= 0.954 - {}^{5}\!/_{8} \times 0 - {}^{3}\!/_{8} \times 0.918 \\
&= 0.610
\end{aligned}$$

The attribute C can take on two values: F or T.

$$\begin{aligned}
Values(C) &= F, T \\
S_{A=F} &= [5_{no}, 3_{yes}] \\
S_{A=F,C=F} &\leftarrow [3_{no}, 1_{yes}] \\
S_{A=F,C=T} &\leftarrow [2_{no}, 2_{yes}]
\end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute C, when $A = F$:

$$\begin{aligned}
Entropy(S_{A=F,no}) &= -{}^{3}\!/_{4}\ log_2({}^{3}\!/_{4}) - {}^{1}\!/_{4}\ log_2({}^{1}\!/_{4}) \\
&= 0.722 \\
Entropy(S_{A=F,yes}) &= -{}^{2}\!/_{4}\ log_2({}^{2}\!/_{4}) - {}^{2}\!/_{4}\ log_2({}^{2}\!/_{4}) \\
&= 1
\end{aligned}$$

Calculate the information gained when dividing the data by using the values of C, when $A = F$:

$$\begin{aligned}
Gain(S_{A=F}, C) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S_{A=F}) - {}^{5}\!/_{8}\ Entropy(S_{A=F,no}) - {}^{3}\!/_{8}\ Entropy(S_{A=F,yes}) \\
&= 0.954 - {}^{5}\!/_{8} \times 0.722 - {}^{3}\!/_{8} \times 1 \\
&= 0.128
\end{aligned}$$

The attribute D can take on two values: F or T.

$$\begin{aligned} Values(D) &= F, T \\ S_{A=F} &= [5_{no}, 3_{yes}] \\ S_{A=F,D=F} &\leftarrow [3_{no}, 1_{yes}] \\ S_{A=F,D=T} &\leftarrow [2_{no}, 2_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute D, when $A = F$:

$$\begin{aligned} Entropy(S_{A=F,no}) &= -{}^3\!/_4\ log_2({}^3\!/_4) - {}^1\!/_4\ log_2({}^1\!/_4) \\ &= 0.722 \\ Entropy(S_{A=F,yes}) &= -{}^2\!/_4\ log_2({}^2\!/_4) - {}^2\!/_4\ log_2({}^2\!/_4) \\ &= 1 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of C, when $A = F$:

$$\begin{aligned} Gain(S_{A=F}, D) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S_{A=F}) - {}^5\!/_8\ Entropy(S_{A=F,no}) - {}^3\!/_8\ Entropy(S_{A=F,yes}) \\ &= 0.954 - {}^5\!/_8 \times 0.722 - {}^3\!/_8 \times 1 \\ &= 0.128 \end{aligned}$$

The highest gain value for the subset of the data where $A = F$, is attribute B, where:

$$Gain(S_{A=F,B}) = 0.610$$

The decision tree grows a new node B as in Figure 29.

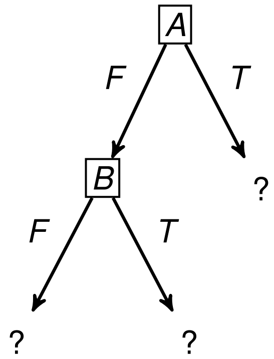


Figure 29: Partial decision tree for f_5.

Calculate the *entropy* of the data set in Table 8, where $A = T$, and

$$S_{A=T} \equiv [2_{no}, 6_{yes}]$$

$$\begin{aligned} Entropy(S_{A=T}) &\equiv \sum_{i=1}^{c} -p_i log_2(p_i) \\ &= -p_{no}\ log_2(p_{no}) - p_{yes}\ log_2(p_{yes}) \\ &= -{}^2\!/_8\ log_2({}^2\!/_8) - {}^6\!/_8\ log_2({}^6\!/_8) \\ &= 0.811 \end{aligned}$$

The attribute B can take on two values: F or T.

$$\begin{aligned} Values(B) &= F, T \\ S_{A=T} &= [2_{no}, 6_{yes}] \\ S_{A=T,B=F} &\leftarrow [1_{no}, 3_{yes}] \\ S_{A=T,B=T} &\leftarrow [1_{no}, 3_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute B, when $A = F$:

$$\begin{aligned} Entropy(S_{A=T,no}) &= -{}^1\!/_4\ log_2({}^1\!/_4) - {}^3\!/_4\ log_2({}^3\!/_4) \\ &= 0.811 \\ Entropy(S_{A=T,yes}) &= -{}^1\!/_4\ log_2({}^1\!/_4) - {}^3\!/_4\ log_2({}^3\!/_4) \\ &= 0.811 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of B, when $A = T$:

$$\begin{aligned} Gain(S_{A=T,B}) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S_{A=T}) - {}^2\!/_8\ Entropy(S_{A=T,no}) - {}^6\!/_8\ Entropy(S_{A=T,yes}) \\ &= 0.954 - {}^2\!/_8 \times 0.811 - {}^6\!/_8 \times 0.811 \\ &= 0.143 \end{aligned}$$

The attribute C can take on two values: F or T.

$$\begin{aligned} Values(C) &= F, T \\ S_{A=T} &= [2_{no}, 6_{yes}] \\ S_{A=T,C=F} &\leftarrow [2_{no}, 2_{yes}] \\ S_{A=T,C=T} &\leftarrow [0_{no}, 4_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute C, when $A = T$:

$$\begin{aligned} Entropy(S_{A=T,no}) &= -{}^{2}\!/_{4}\ log_2({}^{2}\!/_{4}) - {}^{2}\!/_{4}\ log_2({}^{2}\!/_{4}) \\ &= 1 \\ Entropy(S_{A=T,yes}) &= -{}^{0}\!/_{4}\ log_2({}^{0}\!/_{4}) - {}^{4}\!/_{4}\ log_2({}^{4}\!/_{4}) \\ &= 0 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of C, when $A = F$:

$$\begin{aligned} Gain(S_{A=T}, C) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S_{A=T}) - {}^{2}\!/_{8}\ Entropy(S_{A=T,no}) - {}^{6}\!/_{8}\ Entropy(S_{A=T,yes}) \\ &= 0.954 - {}^{2}\!/_{8} \times 1 - {}^{6}\!/_{8} \times 0 \\ &= 0.704 \end{aligned}$$

The attribute D has the same ratios of values as attribute C, and will therefore have the same Gain value:

$$Gain(S_{A=T}, D) = 0.704$$

The highest gain value for the subset of the data where $A = F$, are attributes C and D. We can choose either for the node. Choose attribute C.

$$Gain(S_{A=F,C}) = Gain(S_{A=F,D}) = 0.704$$

The decision tree grows a new node C as in Figure 30.

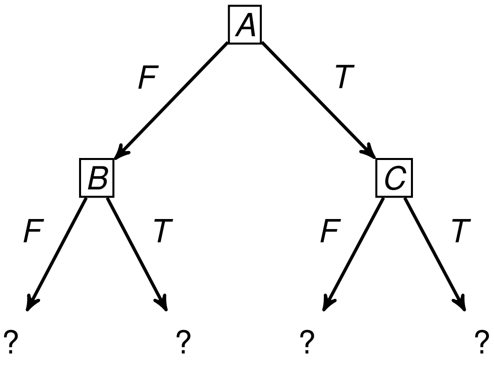


Figure 30: Partial decision tree for f_5.

The subset for $A = F$ is now divided into two further subsets, using the values of attribute B to make the division, as in Tables 9 to 10.

A	B	C	D	f_5
F	F	F	F	*no*
F	F	F	T	*no*
F	F	T	F	*no*
F	F	T	T	*no*

Table 9: Truth table for $f_{5,A=F,B=F}$

A	B	C	D	f_5
F	T	F	F	*no*
F	T	F	T	*yes*
F	T	T	F	*yes*
F	T	T	T	*yes*

Table 10: Truth table for $f_{5,A=F,B=T}$

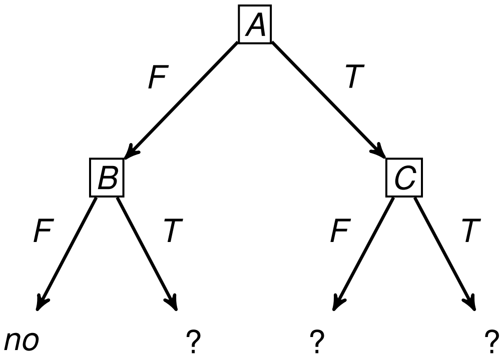


Figure 31: Partial decision tree for f_5.

In Table 9 we see that $f_5 = no$ for all rows. This means we can add a leaf node to the decision tree at this node, as in Figure 31.

Consider the data in Table 10.

Calculate the *entropy* of the data set in Table 10, where $A = F$, $B = T$ and

$$S_{A=F,B=T} \equiv [1_{no}, 3_{yes}]$$

$$\begin{aligned}
Entropy(S_{A=F,B=T}) &\equiv \sum_{i=1}^{c} -p_i log_2(p_i) \\
&= -p_{no}\ log_2(p_{no}) - p_{yes}\ log_2(p_{yes}) \\
&= -{}^1\!/\!_4\ log_2({}^1\!/\!_4) - {}^3\!/\!_4\ log_2({}^3\!/\!_4) \\
&= 0.811
\end{aligned}$$

The attribute C can take on two values: F or T.

$$\begin{aligned}
Values(C) &= F, T \\
S_{A=F,B=T} &= [1_{no}, 3_{yes}] \\
S_{A=F,B=T,C=F} &\leftarrow [1_{no}, 1_{yes}] \\
S_{A=F,B=T,C=T} &\leftarrow [0_{no}, 2_{yes}]
\end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute C, when $A = F$, $B = T$:

$$\begin{aligned}
Entropy(S_{A=F,B=T,no}) &= -{}^1\!/\!_2\ log_2({}^1\!/\!_2) - {}^1\!/\!_2\ log_2({}^1\!/\!_2) \\
&= 1 \\
Entropy(S_{A=F,B=T,yes}) &= -{}^0\!/\!_2\ log_2({}^0\!/\!_2) - {}^2\!/\!_2\ log_2({}^2\!/\!_2) \\
&= 0
\end{aligned}$$

Calculate the information gained when dividing the data by using the values of C, when $A = F, B = T$:

$$\begin{aligned}
Gain(S_{A=F,B=T,C}) &= Entropy(S) - \sum_{v\in\{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S_{A=F,B=T}) - {}^1\!/\!_4\ Entropy(S_{A=F,B=T,no}) - {}^3\!/\!_4\ Entropy(S_{A=F,B=T,yes}) \\
&= 0.811 - {}^1\!/\!_4 \times 1 - {}^3\!/\!_4 \times 0 \\
&= 0.561
\end{aligned}$$

The attribute D can take on two values: F or T.

$$\begin{aligned} Values(D) &= F, T \\ S_{A=F,B=T} &= [1_{no}, 3_{yes}] \\ S_{A=F,B=T,D=F} &\leftarrow [1_{no}, 1_{yes}] \\ S_{A=F,B=T,D=T} &\leftarrow [0_{no}, 2_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute D, when $A = F$, $B = T$:

$$\begin{aligned} Entropy(S_{A=F,B=T,no}) &= -{}^{1}\!/_{2}\ log_2({}^{1}\!/_{2}) - {}^{1}\!/_{2}\ log_2({}^{1}\!/_{2}) \\ &= 1 \\ Entropy(S_{A=F,B=T,yes}) &= -{}^{0}\!/_{2}\ log_2({}^{0}\!/_{2}) - {}^{2}\!/_{2}\ log_2({}^{2}\!/_{2}) \\ &= 0 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of D, when $A = F, B = T$:

$$\begin{aligned} Gain(S_{A=F,B=T,D}) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S_{A=F,B=T}) - {}^{1}\!/_{4}\ Entropy(S_{A=F,B=T,no}) - {}^{3}\!/_{4}\ Entropy(S_{A=F,B=T,yes}) \\ &= 0.811 - {}^{1}\!/_{4} \times 1 - {}^{3}\!/_{4} \times 0 \\ &= 0.561 \end{aligned}$$

We see that:

$$Gain(S_{A=F,B=T,C}) = Gain(S_{A=F,B=T,D}) = 0.561$$

This means we can choose either C or D for the next decision tree node. Choose C to add to the decision tree, as in Figure 32.

The subset for $A = T$ is now divided into two further subsets, using C to make the division, as in Tables 11 and 12.

A	B	C	D	f_5
T	F	F	F	*no*
T	F	F	T	*yes*
T	T	F	F	*no*
T	T	F	T	*yes*

Table 11: Truth table for $f_{5,A=T,C=F}$

Calculate the *entropy* of the data set in Table 11, where $A = T$, $C = F$ and

$$S_{A=T,C=F} \equiv [2_{no}, 2_{yes}]$$

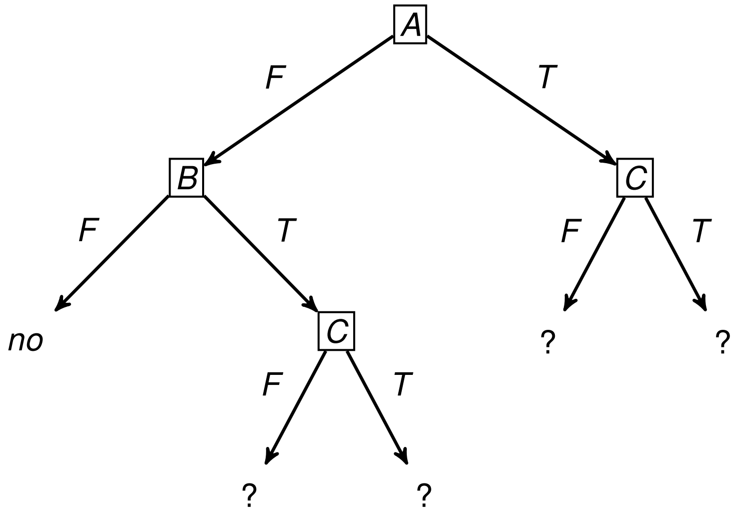


Figure 32: Partial decision tree for f_5.

A	B	C	D	f_5
T	F	T	F	*yes*
T	F	T	T	*yes*
T	T	T	F	*yes*
T	T	T	T	*yes*

Table 12: Truth table for $f_{5,A=T,C=T}$

$$\begin{aligned} Entropy(S_{A=T,C=F}) &\equiv \sum_{i=1}^{c} -p_i log_2(p_i) \\ &= -p_{no}\ log_2(p_{no}) - p_{yes}\ log_2(p_{yes}) \\ &= -{}^2\!/_4\ log_2({}^2\!/_4) - {}^2\!/_4\ log_2({}^2\!/_4) \\ &= 1 \end{aligned}$$

The attribute B can take on two values: F or T.

$$\begin{aligned} Values(B) &= F, T \\ S_{A=T,C=F} &= [2_{no}, 2_{yes}] \\ S_{A=T,C=F,B=F} &\leftarrow [1_{no}, 1_{yes}] \\ S_{A=T,C=F,B=T} &\leftarrow [1_{no}, 1_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute B, when $A = T$, $C = F$:

$$\begin{aligned} Entropy(S_{A=T,C=F,no}) &= -{}^1\!/_2\ log_2({}^1\!/_2) - {}^1\!/_2\ log_2({}^1\!/_2) \\ &= 1 \\ Entropy(S_{A=T,C=F,yes}) &= -{}^1\!/_2\ log_2({}^1\!/_2) - {}^1\!/_2\ log_2({}^1\!/_2) \\ &= 1 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of C, when $A = T$, $C = F$:

$$\begin{aligned}
Gain(S_{A=T,C=F,B}) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S_{A=T,C=F}) - {}^2\!/_4\, Entropy(S_{A=T,C=F,no}) - {}^2\!/_4\, Entropy(S_{A=T,C=F,yes}) \\
&= 1 - {}^2\!/_4 \times 1 - {}^2\!/_4 \times 1 \\
&= 0
\end{aligned}$$

The attribute D can take on two values: F or T.

$$\begin{aligned} Values(D) &= F, T \\ S_{A=T,C=F} &= [2_{no}, 2_{yes}] \\ S_{A=T,C=F,D=F} &\leftarrow [2_{no}, 0_{yes}] \\ S_{A=T,C=F,D=T} &\leftarrow [0_{no}, 2_{yes}] \end{aligned}$$

Calculate the *entropy* values of the subsets of the data, when it is divided using the values of the attribute B, when $A = T$, $C = F$:

$$\begin{aligned} Entropy(S_{A=T,C=F,no}) &= -{}^{2}\!/_{2}\, log_2({}^{2}\!/_{2}) - {}^{0}\!/_{2}\, log_2({}^{0}\!/_{2}) \\ &= 0 \\ Entropy(S_{A=T,C=F,yes}) &= -{}^{0}\!/_{2}\, log_2({}^{0}\!/_{2}) - {}^{2}\!/_{2}\, log_2({}^{2}\!/_{2}) \\ &= 0 \end{aligned}$$

Calculate the information gained when dividing the data by using the values of C, when $A = T$, $C = F$:

$$\begin{aligned} Gain(S_{A=T,C=F,D}) &= Entropy(S) - \sum_{v \in \{F,T\}} \frac{|S_v|}{|S|} Entropy(S_v) \\ &= Entropy(S_{A=T,C=F}) - {}^{2}\!/_{4}\, Entropy(S_{A=T,C=F,no}) - {}^{2}\!/_{4}\, Entropy(S_{A=T,C=F,yes}) \\ &= 1 - {}^{2}\!/_{4} \times 0 - {}^{2}\!/_{4} \times 0 \\ &= 1 \end{aligned}$$

A gain value of 1 means that D describes this subtree completely. The decision is now adjusted to become as in Figure 33.

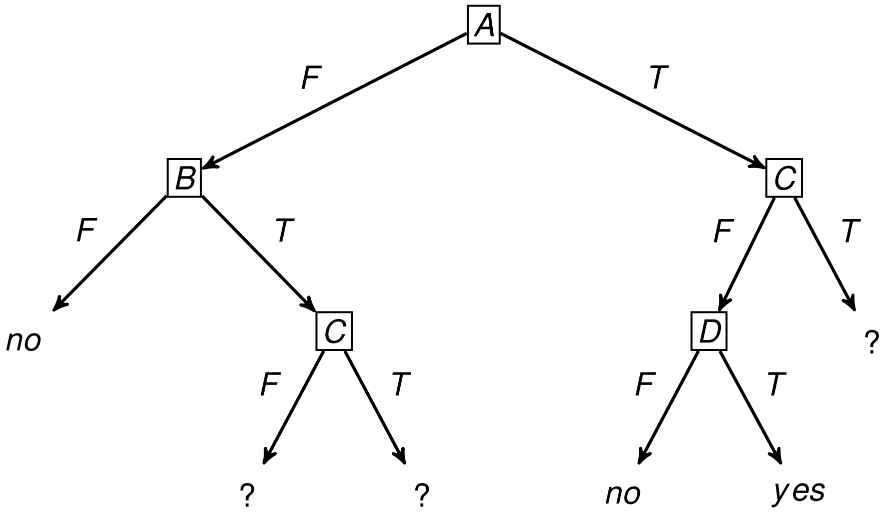


Figure 33: Partial decision tree for f_5.

Looking at the data in Table 12 we see that f_5 only has *yes* values, which means Entropy for this will be zero, and no further subtrees will result. The decision tree then becomes as in Figure 34.

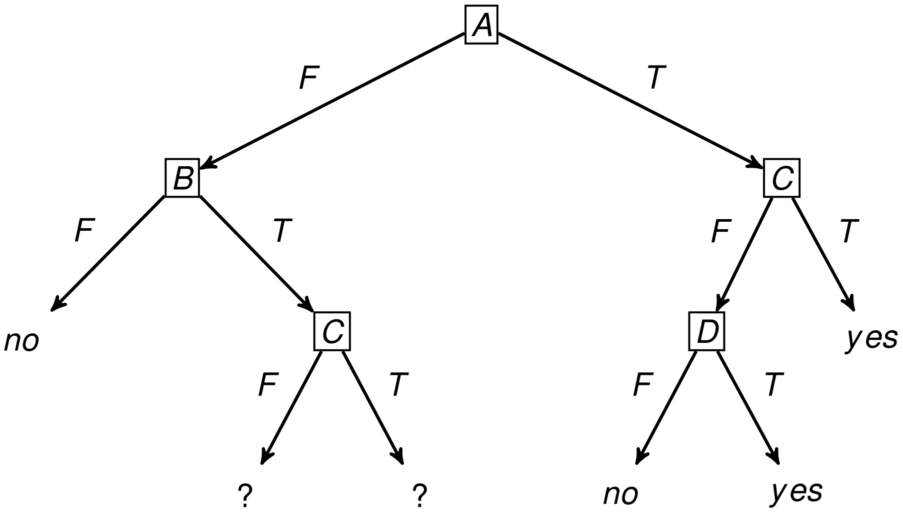


Figure 34: Partial decision tree for f_5.

This leaves us with two subsets of the data, as in Tables 13 and 14.

A	B	C	D	f_5
F	T	F	F	*no*
F	T	F	T	*yes*

Table 13: Truth table for $f_{5,A=F,B=T,C=F}$

A	B	C	D	f_5
F	T	T	F	*yes*
F	T	T	T	*yes*

Table 14: Truth table for $f_{5,A=F,B=T,C=T}$

The ID3 algorithm will complete the calculations here (as you should in the examination to get marks), but by inspection of Table 13 we can see that for the subtree $A = F, B = T, C = F$, the attribute D describes the rest of the data perfectly, and becomes our subtree.

Similarly, in Table 14 f_5 = *yes* for all rows, and becomes a leaf node in the decision tree.

This completes the decision tree, as in Figure 35. Compare this tree with the last binary decision diagram in the previous question.

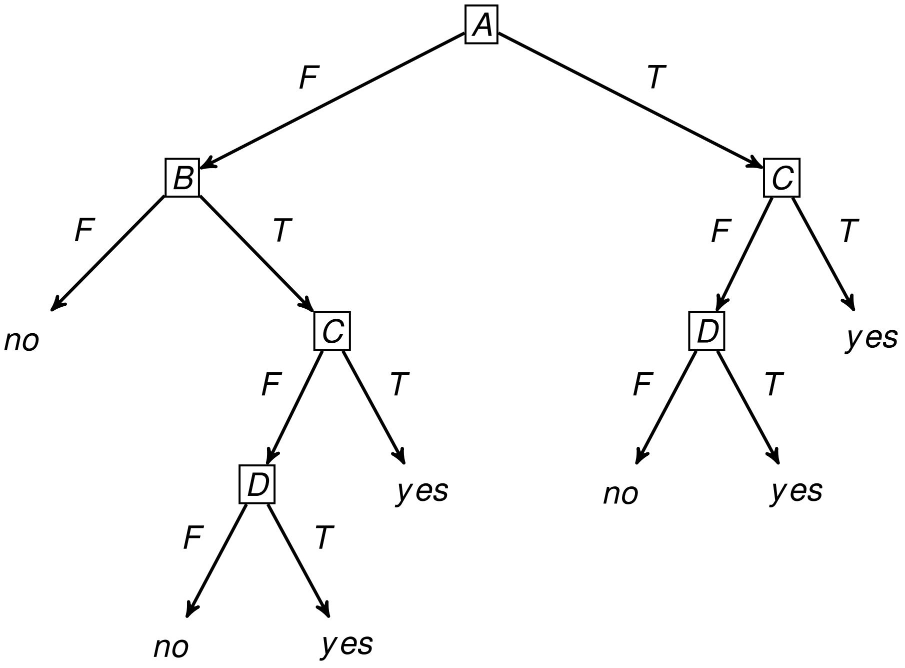


Figure 35: Complete decision tree for f_5.

Mark out of 100.
40 or less for clear indication that student does not understand the topic or evidence of plagiarism, or answers are correct, but have not shown complete workings
50 correct and sufficient workings
60-70 correct and complete workings
80+ indicating thorough understanding of the work

Total marks out of 510.