Tutorial Letter A1/0/2024

Machine Learning

COS4852

Year module

Department of Computer Science

School of Computing

CONTENTS

This document contains the questions for Assignment 1 for COS4852 for 2024.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

This document discusses the questions in Assignment 1 for COS4852 for 2024.

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 $\times$ 100)) = 510.

When we mark the question we want to see that YOU understand the work. Simply copying or regurgitating other people's 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.

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.

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.

Write a short research report (2-4 pages) that describes the k-Nearest Neighbours (kNN) algorithm and its variants. Your report must show a complete kNN algorithm as well as a detailed, worked example with all the steps and calculations. Use the following data set in your worked example:

Positive instances:

$$P_1 = (5, 5, P)$$
$$P_2 = (-6, 4, P)$$
$$P_3 = (-3, -4, P)$$
$$P_4 = (2, -4, P)$$

Negative instances:

$$N_1 = (-1, 2, N)$$
$$N_2 = (-2, 0, N)$$
$$N_3 = (6, 7, N)$$
$$N_4 = (8, -8, N)$$

A new data point of unknown class is at $Q_1 = (2, 1, c)$. Use the kNN algorithm to determine its class. Use at least two different distance metrics.

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 Euclidean plane with *integer* coordinates (x, y), with positive and negative instances as shown in Figure 1.

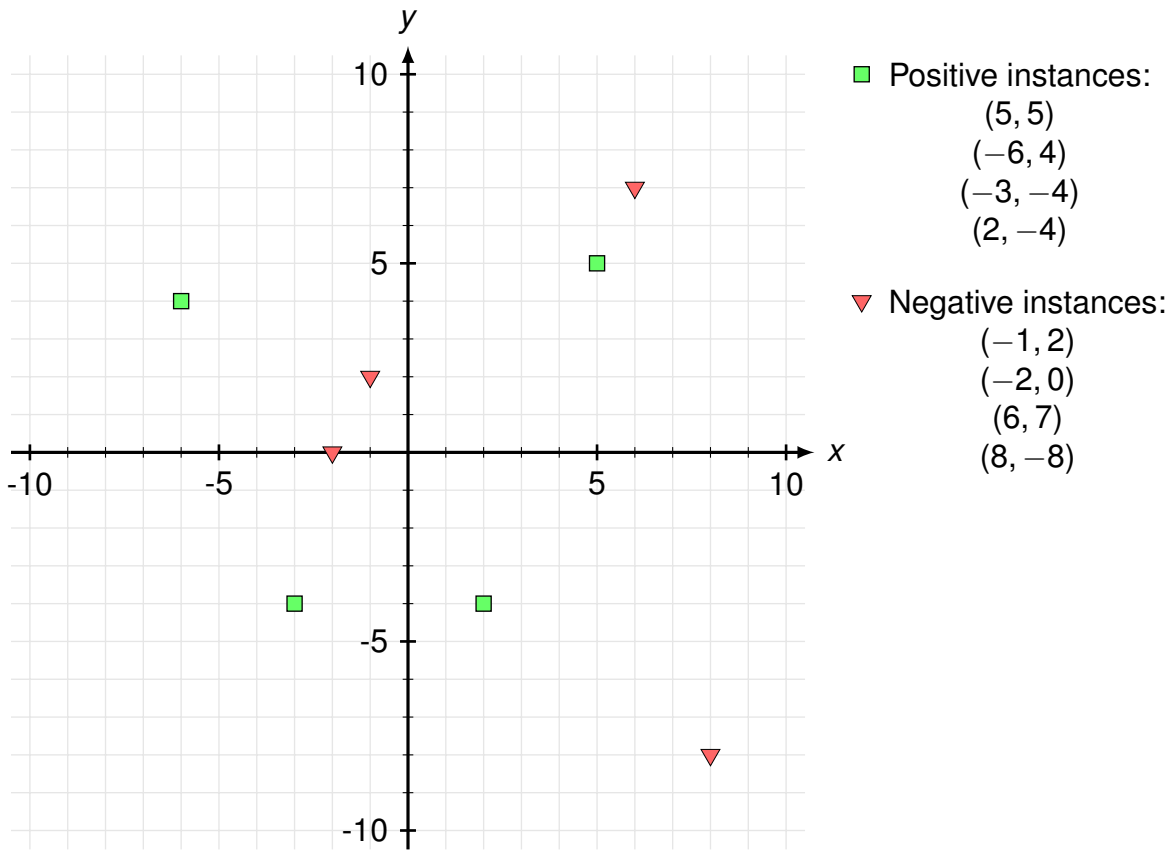

Figure 1: Instance space with positive and negative instances.

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 2. 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.

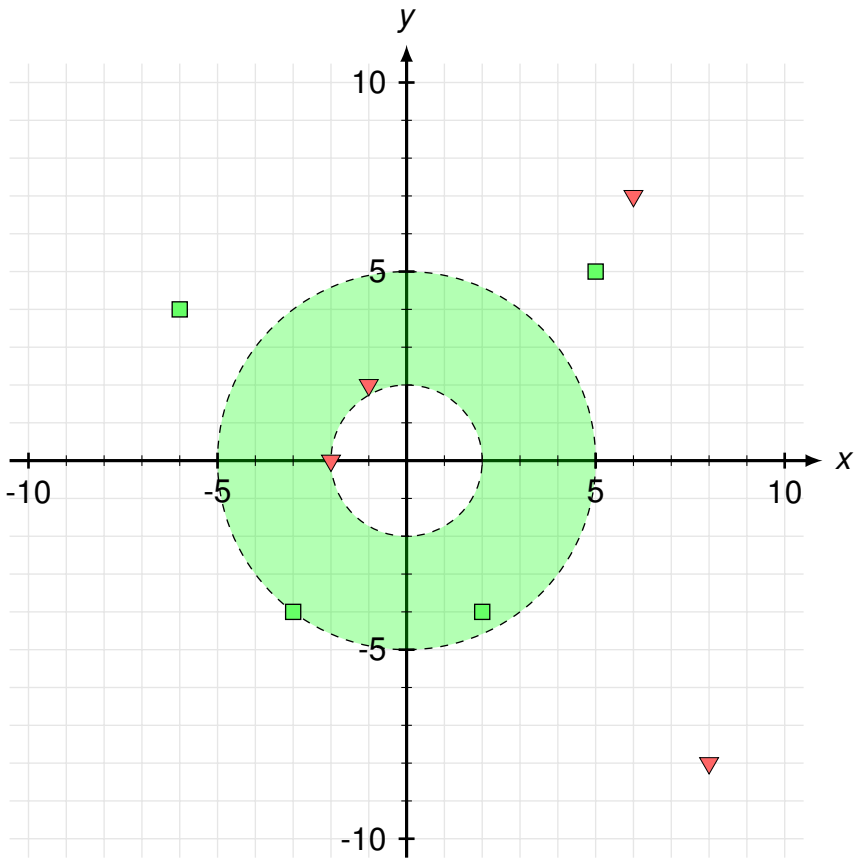

Figure 2: Instance space with a *donut* hypothesis $h \leftarrow \langle 2, 5 \rangle$.

(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.ccs.neu.edu/home/rjw/csg220/lectures/version-spaces.pdf`

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

Show the equivalent binary decision trees that represent the following Boolean functions:

(a) $f_1(A, B, C) = C \vee (A \wedge B)$

(b) $f_2(A, B, C) = (A \vee B) \Rightarrow C$

(c) $f_3(A, B, C) = B \wedge (A \Leftrightarrow C)$

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

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), $\Rightarrow$ is the logical implication, and *Leftrightarrow* is the logical equivalence (look at your undergrad CS notes again).

For this exercise you do **not** need to do the Gain or Entropy calculations. The purpose of the exercise is to understand that there is a direct mapping between a Boolean function and its corresponding binary decision tree. The binary decision tree can often be consolidated to produce a simpler, more compact tree. When you study the ID3 algorithm later, you will see that this is one of the aims of the algorithm – it tries to build the decision tree to be as compact as possible.

Do not just write down the final, simplified tree. Show how you do the simplification.

Here is a resource you could consult on this topic:

- `https://www.cs.cmu.edu/~wlovas/15122-r11/lectures/25-bdds.pdf`

Keep in mind that the discussion there refers to diagrams, which are in effect graphs, and not trees. From your CS knowledge you should understand the difference between a tree and a graph. The question asks you to contruct binary decision *trees*, **not** graphs or diagrams.

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 6

The data in Table 1 correlates a number of variables with whether a student will attend a study session. Use the ID3 algorithm to construct a decision tree for the data in the table. Show all your calculations, including all the steps of the Gain and Entropy calculations. Show the formulas that you use. Clearly explain your choices.

#	TimeOfDay	EnergyLevel	PreviousAttendance	AssignmentDue	Attend?
1	Morning	High	Yes	Yes	Yes
2	Evening	Low	No	Yes	No
3	Evening	High	Yes	No	Yes
4	Afternoon	Medium	Yes	Yes	Yes
5	Morning	Low	No	No	No
6	Afternoon	Low	No	Yes	No
7	Evening	Medium	No	No	Yes
8	Afternoon	Low	Yes	Yes	No
9	Morning	High	Yes	No	Yes
10	Evening	Low	Yes	Yes	Yes
11	Afternoon	High	No	No	Yes
12	Morning	Medium	Yes	Yes	No
13	Afternoon	High	Yes	No	Yes
14	Evening	Medium	No	Yes	Yes

Table 1: Truth table for the Study Session data.

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.geeksforgeeks.org/sklearn-iterative-dichotomiser-3-id3-algorithms/`
- `https://cis.temple.edu/~ingargio/cis587/readings/id3-c45.html`

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.
