PRIFYSGOL ABERYSTWYTH UNIVERSITY

SEMESTER ONE 2026

COMPUTER SCIENCE

CS36110 : FUNDAMENTALS OF MACHINE LEARNING

Time allowed: **2 HOURS**

Answer **THREE** from **FOUR** questions.

All questions carry equal marks.

YOU MAY BRING THE FOLLOWING INTO THE EXAMINATION:
UP TO 3 SHEETS OF NOTES ON A4 PAPER (WHICH MAY BE DOUBLE-SIDED). THERE ARE NO ADDITIONAL RESTRICTIONS ON CONTENT OR FORMATTING OF THESE NOTES.
NO ELECTRONIC DEVICES.

CALCULATORS ARE NOT PERMITTED.

STUDENTS MAY SUBMIT ANSWERS TO THIS PAPER IN EITHER WELSH OR ENGLISH

Once seated, students may complete the front cover of the answer book(s) and the attendance slip ONLY.

DO NOT TURN THE EXAMINATION QUESTION PAPER OVER UNTIL INSTRUCTED TO DO SO.

1. This question is about Support Vector Machines.

a) Briefly define the following in the context of Support Vector Machines:

- The primary goal of a Support Vector Machine (SVM) classifier.
- The concept of the "Maximum Margin Hyperplane."
- Support Vectors

[8 marks]

b) Consider the following training dataset in Table 1 with six data objects and two conditional attributes: a_1 and a_2, along with the class attribute *class* which takes the binary values x and y.

obj.	a_1	a_2	$class$
1	2.0	2.0	y
2	1.5	1.5	y
3	0.0	1.0	x
4	2.0	0.0	y
5	1.0	0.0	x
6	0.0	0.0	x

Table 1: Dataset

(i) Manually plot these six training data objects and determine if they are linearly separable. [4 marks]

(ii) Construct the maximum margin hyperplane by visual inspection and identify the corresponding support vectors by providing their coordinates, object numbers, and the classes they represent. [7 marks]

(iii) Suppose you were to a) remove each one of the support vectors individually and b) remove all support vectors, how would the optimal margin change in each case: decrease, increase or remain the same? Demonstrate this by sketching the changes in your answer book and discussing the outcome with reference to the position of the SVM and the margin. [6 marks]

(iv) Suppose we were to add another data object to the training dataset: data obj 7: with the values: $a_1 = 0.5$, $a_2 = 1.0$, $class = y$, would this affect the hyperplane that we have already established? If it does, explain how

can we tackle this scenario. Give details of two different methods for dealing with it. [8 marks]

2. This question is about Artificial Neural Networks.

a) Consider the artificial neural network in in Fig. 1.

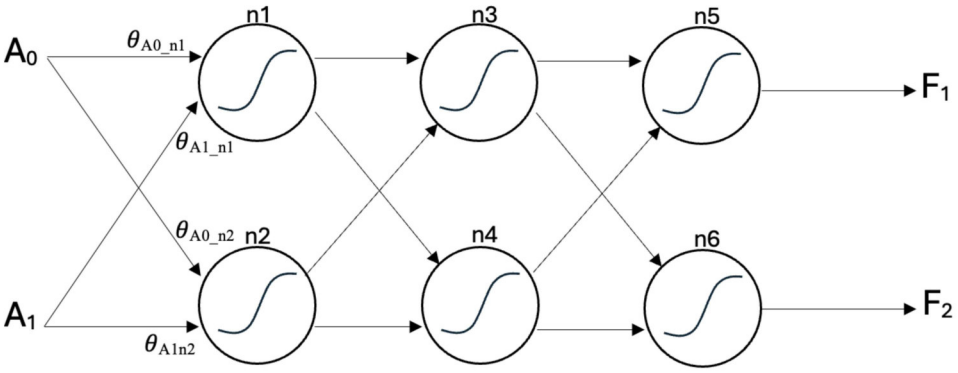


Figure 1: An Artificial Neural Network

(i) How many parameters does the network in Fig. 1 have? [2 marks]

(ii) Briefly (in 1-2 sentences) describe the purpose of the activation function in artificial neural networks, and state which type of function that is implemented in the network in Fig. 1. [4 marks]

(iii) Provide two advantages for the use of the type of activation function such as that described in part (ii) (as opposed to, e.g., a step function), for artificial neural networks. [6 marks]

(iv) If we assume that the activation function for each neuron in Fig. 1 is given by the equation:

$$\frac{1}{1+e^{(x)}}$$

Write out the formulation (no calculation required) for the output for the single neuron: $n1$ in Fig. 1. Note that you are not expected to perform any calculation. [6 marks]

b) Without any learning mechanism, neural networks offer little in the way of utility. The equation: $\Delta w_i = \eta(t-o)x_i$ describes a weight back-propagation approximation or learning method for ANNs known as the *Perceptron Training Rule*.

Using the values shown below, demonstrate the effect on the output of increasing the value of the **learning rate** (η) using a single input: $x_a = 2.0$ which has an

associated weight: $w_a = 0.25$, and a target output $\mathbf{t} = 1.0$ with an initial actual output $\mathbf{o} = 0.5$.

Show the clear and complete workings for two separate cases:

- Initial Learning Rate: $\eta = 0.5$
- Increased Learning Rate: $\eta = 0.9$

Demonstrate the effect on the **change in weight** (Δw_a) and the **new weight** (w_a^{new}) for both cases. [10 marks]

c) What would a good value for η be in the above problem in part b)? Explain your answer clearly and provide a reasoning. [5 marks]

3. This question is about Bayesian Learning.

a) The definition of Bayes Theorem is shown below. Give the names of each of the terms in the equation.

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

Figure 2: Bayes Theorem

[4 marks]

b) Naïve Bayes is one of the more popular Bayes-based approaches in Machine Learning.

(i) Describe <u>two</u> advantages that the Naïve Bayes classifier learner has over other Machine Learning methods. [2 marks]

(ii) A common issue that one may encounter when attempting to build a Naïve Bayes classifier on different types of data is the so-called *'zero-frequency problem'*. There is also another shortcoming that is related to the characteristic of a certain type of data distribution. Describe this second problem and show using a diagram an example of this and explain how a Naïve Bayes classifier would fail in this scenario. [9 marks]

c) A water company is testing the water quality of one of its treatment plants. It suspects that the water is contaminated with harmful bacteria. The test result for the water sample returns positive. Suppose that 8.9% of all water treatment plants across Wales have water that is actually contaminated with this bacteria and that 94.6% of those treatment plants show positive test results; 90.8% of those that are not contaminated with the bacteria show a negative result.

(i) What are the prior and conditional probabilities? [3 marks]

(ii) Using Bayes theorem, formulate (no calculation required) the posterior probability that:

1) The water is contaminated with bacteria, given a positive test result;

2) The water is not contaminated, given a positive test result;

3) The water is not contaminated given a negative test result. [10 marks]

(iii) Is it possible to use the Maximum a-posteriori probability (h_{MAP}) or Maximum Likelihood (h_{ML}), in the problem above? How would the answer to this question differ if we did? Explain your answer. [5 marks]

4. This question is about Decision Tree Learning.

a) Decision Tree Learners often employ information entropy as a metric in order to decide which features to split upon. What does entropy measure? [2 marks]

b) Consider the following dataset of apartments for sale. The decision concept is 'Buy' (yes/no).

data obj.	Furnished	Rooms	Cost	Buy
1	no	2	low	yes
2	yes	2	high	no
3	no	3	high	yes
4	no	2	high	no
5	yes	3	high	yes

Table 2: Apartment Dataset

Note: If you need to compute the Information Gain measure, the following can be used. You also do not need to compute the final values:

* $log_2 3 = 1.58,\ log_2 5 = 2.32,\ log_2 1 = 0,\ log_2 2 = 1$
* $log_2(a/b) = log_2(a) - log_2(b)$

* $Entropy([a, b]) = Entropy(a, b) = -a\ log_2(a) - b\ log_2(b)$

(i) Formulate (no calculation required) the entropy of the decision feature. [2 marks]

(ii) Show the process along with the workings (and an illustration) that a greedy top-down algorithm would utilise to build a decision tree from the above training data in Table 1, in order to predict the value of *Buy* based on the other attributes of the apartment. Explain your choice of attributes placed at each of the nodes. Note that no calculation of the IG values is necessary; only formulation. [13 marks]

(iii) In the above problem of classifying whether we might buy an apartment, why would we prefer to use a decision tree induction approach rather than a classifier based on Artificial Neural Networks or Support Vector Machines? [2 marks]

c) We have been given a household energy usage dataset for different days of the month with a decision feature $F = \{low, medium, high\}$. When attempting to learn a classifier using an ID3 decision tree learner on this data, we notice that one feature in particular has a higher *Information Gain* score than all of the others. It turns out that this feature is related to the `day-of-the-month` and it perfectly classifies the decision.

(i) Is the feature `day-of-the-month` sufficient to classify all of the data? Explain why. [4 marks]

(ii) If a different metric such as *Information Gain Ratio* were to be used, would that make any difference to the feature/s that are split upon? [4 marks]

(iii) Decision tree learners tend to be susceptible to overfitting. Give two reasons why this is the case and also say if the decision tree discussed above may suffer from overfitting, and how. [6 marks]