

Question 5 (25%)

I. Table 4 is a training dataset for a two-category classification task

Table 4. Data set with 4 attributes

$attr_1$	$attr_2$	$attr_3$	$attr_4$	$class$
a	1	c	-1	c_1
b	0	c	-1	c_1
a	0	c	1	c_1
b	1	c	1	c_1
b	0	c	1	c_2
a	0	a	-1	c_2
a	1	a	-1	c_2
b	1	c	-1	c_2

Given the entropy equations below, derive the information gain and construct a one level decision tree for this data using the information gain as your splitting criterion.

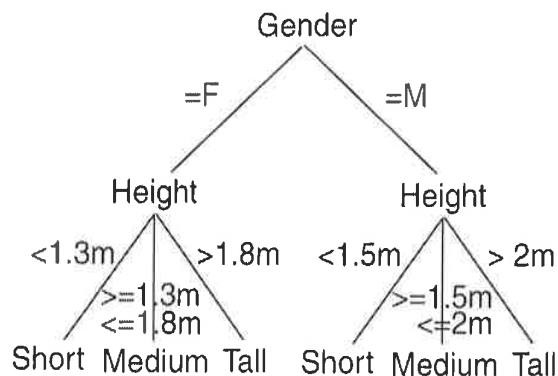
$$entropy(D) = - \sum_{i=1}^n p_i \log_2(p_i)$$

$$entropy(p_1, p_2, \dots, p_n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \dots - p_n \log_2 p_n$$

Please show all calculations and tree construction at the first level.

(15 Mark)

II. The following decision tree can be used to classify students as short, tall or medium. Based on the node information shown in the tree, build a neural network structure that is composed of three layers.



1). Assign labels to each of the neurons in the network.

(1 Mark)

2). Connect all neurons between the layers by arcs.

(1 Mark)

3). Annotate weights for each arc (interconnection).

(2 Marks)

4). Briefly justify why you chose those tree nodes to define neurons in the input and output layers and why you build the neural network in this way.

(6 marks)

Question 4 (25%)

The following table represents loan borrower data that can be used to determine if a loan borrower will default on their payments.

ID	home owner	marital status	annual income	default borrower
1	Yes	Single	[100K, 125K]	No
2	No	Married	[100K, 125K]	No
3	No	Single	[60K, 75K]	No
4	Yes	Married	[100K, 125K]	No
5	No	Divorced	[85K, 95K]	Yes
6	No	Married	[60K, 75K]	No
7	Yes	Divorced	220K	No
8	No	Single	[85K, 95K]	Yes
9	No	Married	[60K, 75K]	No
10	No	Single	[85K, 95K]	Yes

I. Use these data to estimate the following probabilities:

- 1). $P(\text{default borrower} = \text{Yes}) = ?$ (1 Mark)
- 2). $P(\text{default borrower} = \text{No}) = ?$ (1 Mark)
- 3). $P(\text{home owner} = \text{Yes} \mid \text{default borrower} = \text{No}) = ?$ (1 Mark)
- 4). $P(\text{home owner} = \text{No} \mid \text{default borrower} = \text{No}) = ?$ (1 Mark)
- 5). $P(\text{home owner} = \text{Yes} \mid \text{default borrower} = \text{Yes}) = ?$ (1 Mark)
- 6). $P(\text{home owner} = \text{No} \mid \text{default borrower} = \text{Yes}) = ?$ (1 Mark)
- 7). $P(\text{marital status} = \text{Single} \mid \text{default borrower} = \text{No}) = ?$ (1 Mark)
- 8). $P(\text{marital status} = \text{Divorced} \mid \text{default borrower} = \text{No}) = ?$ (1 Mark)
- 9). $P(\text{marital status} = \text{Married} \mid \text{default borrower} = \text{No}) = ?$ (1 Mark)
- 10). $P(\text{marital status} = \text{Single} \mid \text{default borrower} = \text{Yes}) = ?$ (1 Mark)
- 11). $P(\text{marital status} = \text{Divorced} \mid \text{default borrower} = \text{Yes}) = ?$ (1 Mark)
- 12). $P(\text{marital status} = \text{Married} \mid \text{default borrower} = \text{Yes}) = ?$ (1 Mark)

II. Suppose you are given a test instance with the attribute {home owner = No, marital status = Married, annual income = \$120K}. Assume conditional independence between the attributes, then:

- 1). Show the mathematical expression of the Bayes classification rule. (1 mark)
- 2). Use the Bayes classification rule to classify the test instance as either default borrower = Yes or default borrower = No. (12 marks)

Question 5 (25%)

- I. Table 3 describes Cartesian coordinates for three different classes, A, B, and C.

Table 3. Data set describing boats

Class A		Class B		Class C	
A1	(2,12)	B1	(6,5)	C1	(1,5)
A2	(3,8)	B2	(8,1)	C2	(1,3)
A3	(4,9)	B3	(7,7)	C3	(5,1)
A4	(7,10)			C4	(7,8)
				C5	(9,3)

Use the Euclidean distance equation below

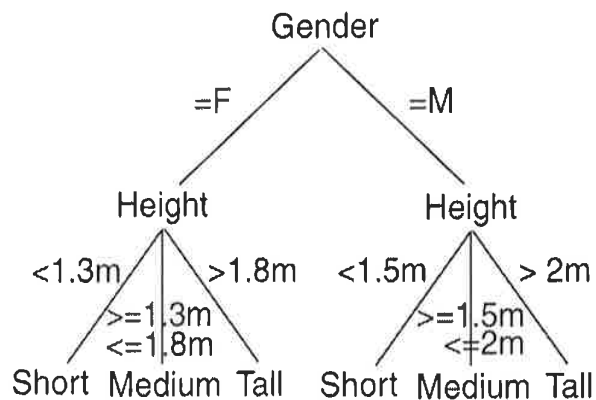
$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

to compute the distances from point (6, 6) to the four instances of Class A, three instances of Class B and five instances of Class C. Then use these distances to classify to which class the point (6, 6) belongs by the k-nearest neighbour technique if:

- 1) k=1;
- 2) k=5.

Note that x and y in the above distance equation are instances being comprised of coordinates. Show all intermediate steps. (13 marks)

- II. The following decision tree can be used to classify students as short, tall or medium. Based on the node information shown in the tree, build a neural network structure that is composed of three layers.



- 1). Assign labels to each of the neurons in the network. (1 Mark)
- 2). Connect all neurons between the layers by arcs. (1 Mark)
- 3). Annotate weights for each arc (interconnection). (2 Marks)
- 4). Briefly justify why you chose those tree nodes to define neurons in the input and output layers, why you build the neural network in this way. (8 marks)