QUESTION 1

a. Consider the following file from a situation where it is required to predict today's stock price from the stock prices of the previous three days.

```
@RELATION Stock_price
@ATTRIBUTE today_3      REAL
@ATTRIBUTE today_2      REAL
@ATTRIBUTE today_1      REAL
@ATTRIBUTE today        REAL
30.22,29.85,29.30,29.38
29.85,29.30,29.38,29.79
29.30,29.38,29.79,30.05
.....
```

1. Could you use a decision tree classifier for this task? Why or why not? (2 marks)
2. Which classifier would you recommend for this task? Why? (2 marks)
3. What would be a suitable error measure for this task? (2 marks)

b. Describe the 10 fold cross validation procedure. Can it be used to prevent overfitting? Why or why not? (5 marks)

c. What is feature selection and what is its purpose? Distinguish the filter and wrapper methods for feature selection? Could a decison tree be used for feature selection? If so how? (5 marks)

d. Describe the basic operation of an ensemble classifier. Show how bagging is implemented within this scheme. (4 marks)

(Total for question 1: 20 marks)

QUESTION 2

a. In the context of data mining and knowledge discovery what is a golden nugget? Give an example. (3 marks)

b. With the aid of a diagram, describe the KDD (Knowledge Discovery in Databases) process. (4 marks)

c. A hospital has accumulated many records over the years. They would like to identify patients who have a high risk of heart attacks.

Describe how you would use a the KDD process to identify these patients. Make any assumptions you think are appropriate about the availability of data. (6 marks)

(Total for question 2: 13 marks)

QUESTION 3

The following data was collected from a computer games shop.

Number	Age	Income	Student	Credit Rating	Class (Buys Minecraft)
1	15	65,000	no	fair	no
2	18	60,000	no	excellent	no
3	48	65,000	no	fair	yes
4	65	43,000	no	fair	yes
5	68	12,000	yes	fair	yes
6	82	14,000	yes	excellent	no
7	45	13,000	yes	excellent	yes
8	19	44,000	no	fair	no
9	18	15,000	yes	fair	yes
10	62	48,000	yes	fair	yes
11	17	46,000	yes	excellent	yes
12	51	47,000	no	excellent	yes
13	51	80,000	yes	fair	yes
14	70	46,000	no	fair	

a. How would ZeroR classify example 14? (1 mark)

b. How would 1-nearest neighbour classify example 14? (2 marks)

c. Derive a OneR classifier from the attribute credit rating. How would example 14 be classified? Show your working. (5 marks)

d. Using only attributes Student and Credit Rating derive a naive Bayes classifier and use it to classify case 14. [A simplified expression is sufficient, it not necessary to do all of the arithmetic.] (5 marks)

e. Show how the numerical attribute Age can be incorporated into your classifier from the previous question. How will case 14 be classified now? [A simplified expression is sufficient, it not necessary to do all of the arithmetic.] (5 marks)

f. In constructing a decision tree, which of the attributes Student and Credit Rating would be chosen for splitting based on information gain? [A simplified expression is sufficient, it not necessary to do all of the arithmetic.] (5 marks)

g. Suppose that examples 1-13 are correctly classified by J48. Give the confusion matrix. (2 marks)

h. Using only the attributes Student, Credit Rating and Class and the apriori algorithm derive two association rules and give their support and confidence. Which rule is better? (5 marks)

(Total for question 3: 30 marks)

QUESTION 4

Show how you would train a neural network classifier for the data in question 3. Your answer should address the following items:

a. Coding of the inputs (2 marks)

b. Coding of the outputs (2 marks)

c. Network architecture, including the hidden layer(s) (2 marks)

d. Network training (5 marks)

e. Computing the classification error rate (2 marks)

(Total for question 4: 13 marks)

QUESTION 5

a. For the data below, write a script that will prepare data patterns for a neural network. The script should

1. Only generate the patterns, not the header.
2. Ignore any records with missing values.
3. Ignore the record number.
4. Extract just the numeric attributes to be the inputs.
5. Replace commas with spaces.
6. Generate the class as a 1-out-of-n coding.

The first record should be
18 60000 1 0

```
Number,Age,Income,Student,CreditRating,Class
1,?,65000,no,fair,no
2,18,60000,no,excellent,no
3,48,65000,no,fair,yes
4,65,43000,no,fair,yes
5,68,12000,yes,fair,yes
6,82,14000,yes,excellent,no
7,45,13000,yes,excellent,yes
8,19,44000,no,fair,no
9,18,15000,yes,fair,yes
10,62,48000,yes,fair,yes
11,17,46000,yes,excellent,yes
12,51,47000,no,excellent,yes
13,51,80000,yes,fair,yes
14,70,46000,no,fair
```

(Total for question 1: 10 marks)

QUESTION 6

a. What is the difference between classification and clustering? Give an example of each task. (4 marks)

b. Using the Age and Income attributes from examples 1-4 of question 3, show the operation of KMeans for two cycles for $K = 2$. (6 marks)

c. What are the two major differences between the KMeans and the EM algorithms? (4 marks)

d. Consider the following output from a run of the EM program:

```
Number of clusters: 3
                        Cluster
Attribute                     0          1          2
                         (0.23)     (0.18)     (0.59)
=====================================================
Age
  mean                  41.2984    23.1719    42.1232
  std. dev.             13.0438     3.9147    12.4025
Occupation
   Adm-clerical        992.8079    445.347   479.8452
   Exec-managerial     500.3791    56.6106  1477.0103
   Handlers-cleaners    34.4042    304.348   338.2478
   Prof-specialty      653.8157   129.4647  1276.7196
   Other-service       577.6233   610.3606    426.016
   Sales                 332.52   467.0152   986.4648
   Craft-repair         18.5034    97.5173  1973.9793
   Transport-moving     15.7716    37.4879   770.7405
   Farming-fishing      11.2909    56.4892   419.2199
   Machine-op-inspct   188.0553   202.4871   598.4575
   Tech-support        119.0342   100.5658      265.4
   ?                   246.8391   328.9673   373.1936
   Protective-serv       19.645    41.9137   268.4413
   Armed-Forces          1.0052      4.155     2.8397
   Priv-house-serv      58.9004    14.0321     1.0674
  [total]             3770.5954  2896.7615  9657.6431
Sex
   Male                 29.1587  1557.3296  9313.5117
   Female             3728.4367  1326.4319   331.1314
  [total]             3757.5954  2883.7615  9644.6431
Clustered Instances

0       3726 ( 23%)
1       3044 ( 19%)
2       9511 ( 58%)
```

1. Sketch the distributions of Age for each cluster. (3 marks)
2. What is the influence of Age on each cluster? (2 marks)
3. What is the influence of Sex on each cluster? (2 marks)
4. Give an English language description of cluster 2. (3 marks)

(Total for question 5: 24 marks)

END OF EXAMINATION (Total Marks 110)

SOME FORMULAS YOU MIGHT FIND USEFUL

n_b	number of examples in branch b
n_t	total number of examples in all branches
n_{bc}	total number of examples in branch b of class c

$$Average\ disorder = \sum_b \left(\frac{n_b}{n_t}\right) \times \left(\sum_c -\frac{n_{bc}}{n_b} log_2 \frac{n_{bc}}{n_b}\right) \tag{1}$$

$$info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} \times log_2 \left(\frac{freq(C_j, T)}{|T|}\right) \quad bits \tag{2}$$

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \tag{3}$$

$$gain(X) = info(T) - info_X(T) \tag{4}$$

$$split\ info(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times log_2 \left(\frac{|T_i|}{|T|}\right) \tag{5}$$

$$gain\ ratio = \frac{gain(X)}{split\ info(X)} \tag{6}$$

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \tag{7}$$

$$P(H|E) = \frac{P(E|H).P(H)}{P(E|H).P(H) + P(E| \sim H).P(\sim H)} \tag{8}$$

$$P(C_i|\vec{x}) = \frac{P(\vec{x}|C_i).P(C_i)}{\sum_j P(\vec{x}|C_j).P(C_j)} \tag{9}$$

$$y_i = \frac{1}{1 + e^{-kx_i}} \tag{10}$$

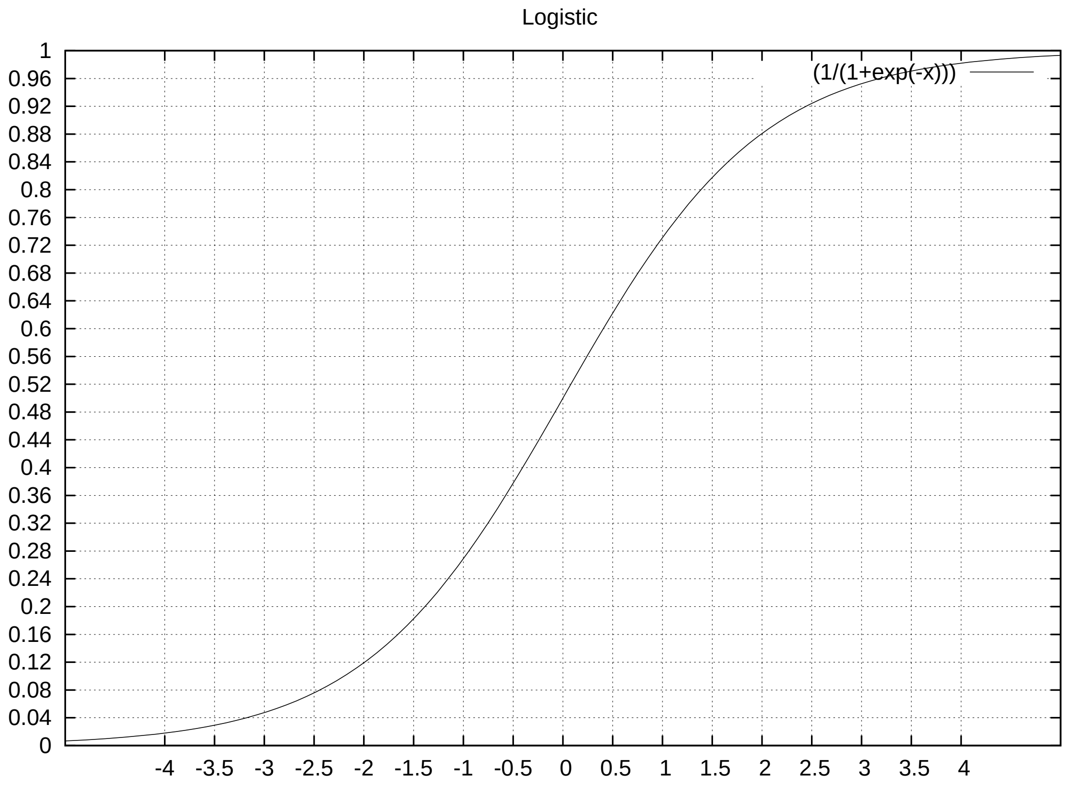
$$\frac{1}{2} \sum_{patterns} \sum_z (d_z - o_z)^2 \tag{11}$$

$$\beta_j = \sum_k w_{i \to j} o_k (1 - o_k) \beta_k \tag{12}$$

$$\Delta w_{i \to j} = \eta o_i o_j (1 - o_j) \beta_j \tag{13}$$

$$\Delta w_{i \to j}(t+1) = \eta o_i o_j (1 - o_j) \beta_j + \alpha \Delta w_{i \to j}(t-1) \tag{14}$$

$$||\vec{x} - \vec{y}|| = \sqrt{\sum_{1}^{n} (x_i - x_j)^2} \tag{15}$$


Logistic
(1/(1+exp(-x)))