UNIVERSITY OF MANITOBA
Final Examination

Winter 2004

COMPUTER SCIENCE

Machine Learning

Paper No.: **482**
Examiners: **Jacky Baltes**
Date: **21 April 2004**
Time: **18:00**
Room: **Fr. Kennedy, Brown Gym (360 - 385)**

(Time allowed: 180 Minutes)

NOTE: Attempt all questions.
This is a *closed* book examination.
Use of calculators is *permitted*.
Show your work to receive full marks.

SURNAME:

FORENAME(S):

STUDENT ID:

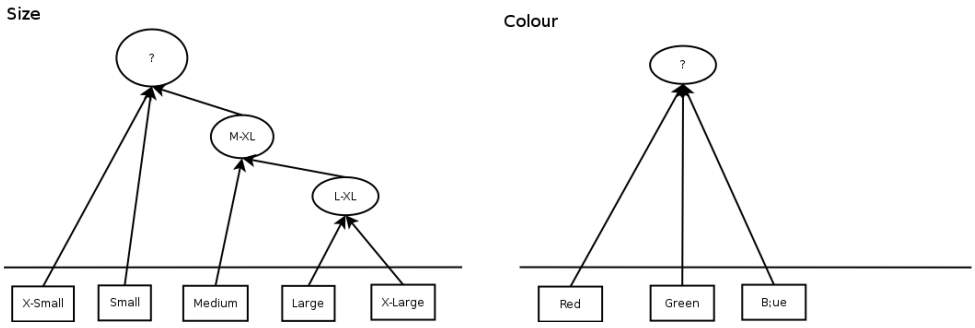
A	B	C	D	E	Total
20	20	20	20	20	100

Surname: _____

Forename(s): _____

Section A: Candidate Elimination

1. Given the hypothesis space H below, the candidate elimination algorithm is trained on a sequence of instances (Training data D).



Given the generalization hierarchy H in question 1, show a trace (i.e., S-set and G-set) of the candidate elimination algorithm for the following five instances. If it is impossible to trace the execution of the candidate elimination algorithm based on the information above, then say so in your answer and explain why.

[5 marks]

<S, R> -
S-Set: 0
G-Set: <XS,?>,<M-XL,?>,<?,G>,<?,B>
<L, G> +
S-Set: <L,G>
G-Set: <M-XL,?>,<?,G>
<XS, B> -
S-Set: <L,G>
G-Set: <M-XL,?>,<?,G>
<XL, B> +
S-Set: <L-XL,?>
G-Set: <M-XL,?>
<S, B> -
S-Set: <L-XL,?>
G-Set: <M-XL,?>

2. Given the generalization hierarchy H in question 1, assume that after a training sequence D, the candidate elimination algorithm returns the following version space V:

S-set : <L-XL,G>
G-set : <M-XL,?>,<?,G>

Surname: _____

Forename(s): _____

How many concepts consistent with version space V classify the instance <M,B> as positive and how many classify it as negative? If it is impossible to determine the classification from the information above, then say so in your answer and explain why it is impossible.

[5 marks]

<M,B> is classified as positive by 1 concept and negative by 4 concepts.

3. Given the generalization hierarchy H in question 1, the candidate elimination algorithm is trained on the training sequence D' and returns the version space V':

   S-set : <L-XL,G>
   G-set : <?,?>

   Assuming an active learner (i.e., a learner that can request the classification of instances), what is the best instance to query next. (i.e., the instance that maximally reduces the size of the remaining version space independently of whether that instance is classified as positive or negative).

   Show the instance as well as the resulting version spaces given that this version space is classified as positive or negative respectively.

   [5 marks]

   The best instance to query next is <L,B>
   If this instance is **positive** the resulting version space is
   S-Set: <L-XL,?>
   G-Set: <?,?>
   If this instance is **negative** the resulting version space is
   S-Set: <L-XL,G>
   G-Set: <?,G>

4. What is the maximum number of semantically different concepts that can be expressed given the generalization hierarchy H in question 1? If it is impossible to determine the maximum number of semantically different concepts in H, then say so in your answer and explain why.

   [5 marks]

   The maximum number of semantically different concepts in H is: 1 + 8 * 4 = 33

Surname: _____

Forename(s): _____

Section B: Decision Trees

The information gain Gain(S,A) of an attribute A for a sample set S is defined as

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{\|S_v\|}{\|S\|} \text{Entropy}(S_v)$$

A graph of the entropy function is shown in Fig. 1 below. You can use this graph when answering the following questions.

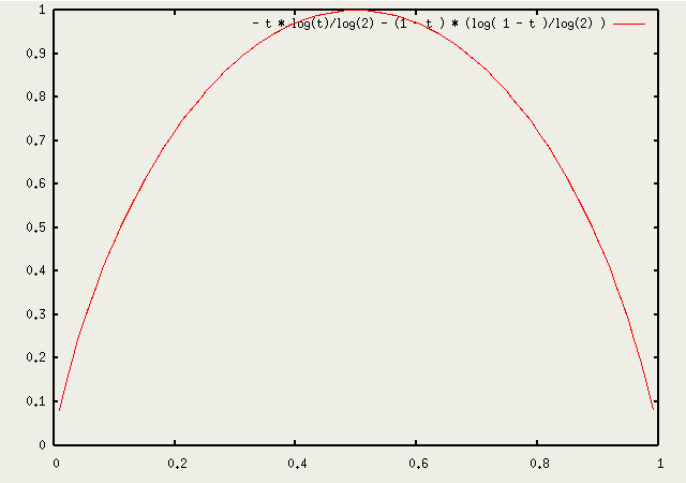


Figure 1: Graph of the Entropy Function

5. Assume a domain with three attributes A, B, and C. Each attribute has two possible values T and F. Given below is a set of instances.

A	B	C	Target
T	T	T	Yes
T	F	T	No
T	F	F	Yes
F	T	F	No
F	F	F	Yes

Calculate the information gain (Gain$(S, ?)$) for the attributes A, B, and C. Which attribute would be selected by the standard ID3 algorithm? If it is impossible to calculate the information gain from the given information, then specify so in your answer and explain why.

[7 marks]

Surname: _____

Forename(s): _____

Gain$(S, \mathsf{A}) = 0.0199$

Gain$(S, \mathsf{B}) = 0.0199$

Gain$(S, \mathsf{C}) = 0.0199$

ID3 would select any attribute
ent(3,2)=0.97
IG(A)=ent(3,2) - 3/5 * ent(2,1) - 2/5 * ent(1,1) = 0.0199
IG(B)=ent(3,2) - 2/5 * ent(1,1) - 3/5 * ent(1,2) = 0.0199
IG(C)=ent(3,2) - 2/5 * ent(1,1) - 3/5 * ent(2,1) = 0.0199

6. Given is a sample set S' with unknown classification. However, you are told that the entropy of the whole sample set S' is 0.918 (Entropy$(S') = 0.918$). Furthermore, the information gain of attribute A and attribute B are exactly the same. (Gain$(S', \mathsf{A}) =$ Gain(S', B)).

   Assign classifications (True/False) to S' that satisfy these two constraints. If it is impossible to make such an assignment, then say so in your answer and explain why.

   [8 marks]

A	B	Target
T	T	Yes/No
T	F	No/Yes
F	F	Yes/No

Two solutions shown above.

7. Assume a domain with three attributes (X,Y,Z). Each attribute can have the value True or False. You are given a training set S_3 with four instances and their classification (Yes/No).

   How many nodes are in the largest decision tree (i.e., the tree with the maximum number of nodes) that can be generated from this training data using the ID3 algorithm. Explain your answer.

   If it is impossible to determine the maximum number of nodes in the decision tree without further information, then say so in your answer and explain why.

CONTINUED

Surname: _____

Forename(s): _____

[5 marks]

The maximum number of nodes (including terminal nodes) in the decision tree generated by ID3 is 7
Both types of trees lead to the

Section C: Neural Nets

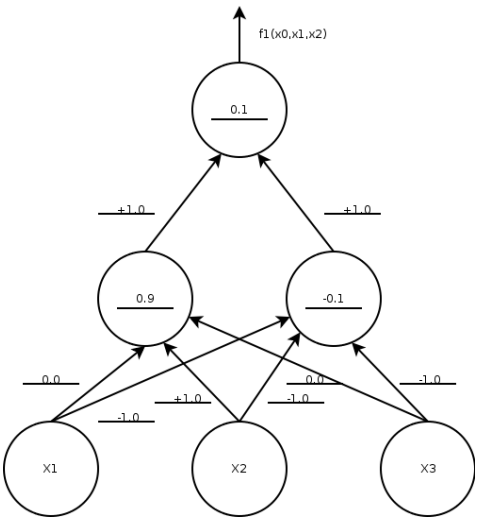
8. Given below is an artificial neural network (ANN) with three input nodes (X1,X2,X3), two hidden nodes, and one output node. The network uses simple threshold nodes (i.e., the node will output 1.0 if the sum of the weighted inputs is greater than the threshold, 0 otherwise).

   Show a set of weights and thresholds for all nodes that implement the boolean function f_1. If it is impossible to represent the boolean function f_1 with the given neural network, then state this in your answer and explain why this is impossible.

   [10 marks]

Surname: _____

Forename(s): _____

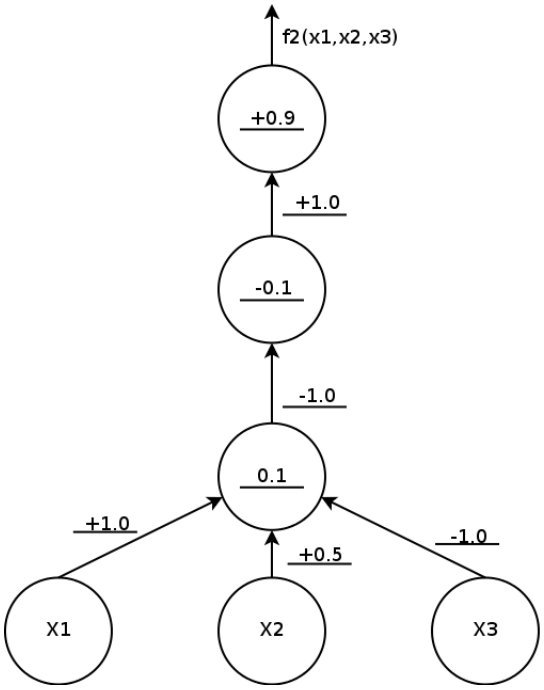


x_1	x_2	x_3	f_1
0	0	0	1
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

9. Given below is a network N_2 with two hidden layers. This network computes the function f_2. The weights of the network N_2 are shown in the figure.

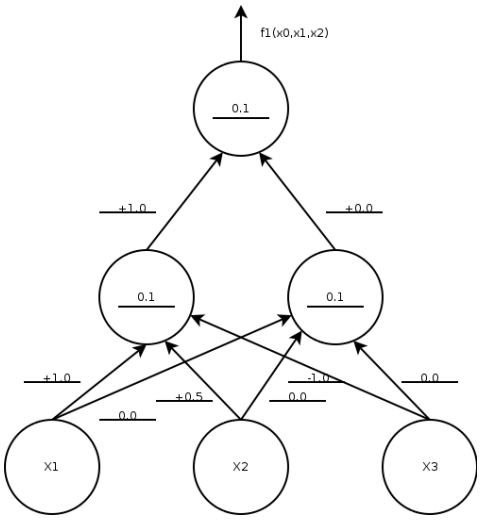
Surname: _____

Forename(s): _____



Convert this two level network N_2 into a new network N_3 with a single hidden layer that computes the same output function f_2. The thresholds of the nodes in the hidden layer are fixed and can not be changed. Show the missing weights in the new network N_3.

[10 marks]



CONTINUED

Surname: _____

Forename(s): _____

Section D: Bayesian Learning

10. Given the following data set, what is the naive Bayesian classification of the new instance <L,green>. Show your work for full marks.

Size	Color	Target
S	black	Yes
M	white	No
L	green	Yes
S	black	Yes
M	white	No
L	green	No
S	black	Yes
L	red	Yes
XS	green	No

Table 1: Training set for the Size, Color domain

[5 marks]

```
P(Yes) = 5/9
P(No) = 4/9
P(L|Yes) = 2/5
P(L|No) = 1/4
P(green|Yes) = 1/5
P(green|No) = 2/4
P(yes|L ^ green) = P(Yes) * P(L|Yes) * P(green|Yes) = 5/9*2/5*1/5=2/45
P(no|L ^ green) = P(No) * P(L|No) * P(green|No) = 4/9*1/4*2/4=2/36=1/18
-> No
```

11. You are given the following information about the hollywood movie domain. The name of the associated random variables is given in brackets.

- (Type) The new movie is either a comedy (25%) or an action movie (75%).
- (Star) With 90% probability, the star in an action movie is famous. With 50% probability, the star in a comedy is famous.
- (Script) The probability that the movie has a good script is 10%.
- (Explosions) If an action movie has a good script, the probability of explosions in the movie are 60%. A comedy with a good script has a 40% probability of explosions. Independently of the type of movie, a movie with a bad script has a 80% probability of explosions.

CONTINUED

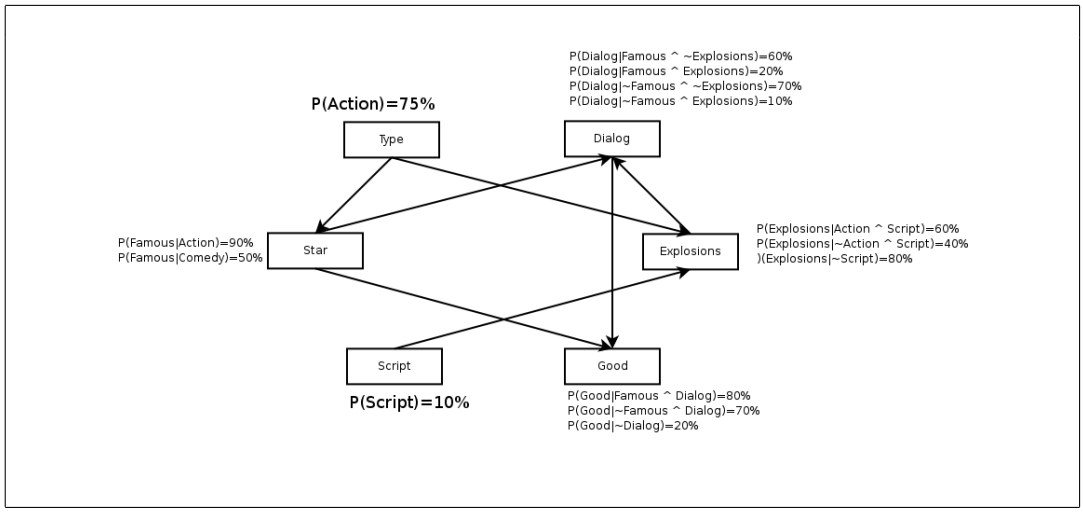
Surname: _____

Forename(s): _____

- (Dialog) Any movie with a famous star and no explosions has a 60% probability of having good dialog. Any movie with a famous star and explosions has a 20% probability of having good dialog. Any movie without a famous star and no explosions has a 70% probability of having good dialog. Any movie without a famous star and explosions has a 10% probability of having good dialog.
- (Good) A movie with a famous star and good dialog is a good movie with 80% probability. A movie with an unknown star and good dialog is a good movie with 70% probability. Any movie with bad dialog has a 20% change of being a good movie.

Describe this information in the form of a bayesian belief network. Show the graph of the Bayesian belief network as well as all the conditional probabilities for all random variables.

[5 marks]



12. What is the probability of a new movie being good, given that it has a good script. In other words, calculate $P(\text{Good}|\text{Script})$.

[5 marks]

Surname: _____

Forename(s): _____

$P(\text{Good}|\text{Script}) = 42\%$

```
P(G) given that the script is good P(S):42\%
P(F) = P(F|A)*P(A) + P(F|~A)*P(~A)
     = 0.9*0.75 + 0.5*0.25=0.8
P(E) = P(E|A^S)*P(A)*1.0 + P(E|~A^S)*P(~A)*1.0     S is given,P(S)=1.0
     = 0.6*0.75 + 0.4*0.25=0.55
P(D) =    P(D|F^~E)*P(F)*P(~E)         0.6*0.8*0.45
       +  P(D|F^E)*P(F)*P(E)          + 0.2*0.8*0.55
       +  P(D|~F^~E)*P(~F)*P(~E)      + 0.7*0.2*0.45
       +  P(D|~F^E)*P(~F)*P(E)        + 0.1*0.2*0.55=0.378
P(G) =    P(G|D^F)*P(D)*P(F)           0.8*0.387*0.8
       +  P(G|D^~F)*P(D)*P(~F)        + 0.7*0.387*0.2
       +  P(G|~D)*P(~D)               + 0.2*0.622 = 0.42
```

13. Without any other information, what is the probability of a new movie having explosions. In other words, calculate $P(\text{Explosions})$.

[5 marks]

$P(\text{Explosions}) = 0.775$

```
P(E) =     P(E|A^S)*P(A)*P(S)          0.6*0.75*0.1
         + P(E|~A^S)*P(~A)*P(S)      + 0.4*0.25*0.1
         + P(E|~S)*P(~S)             + 0.8*0.9 = 0.775
```

14. You watched a bad movie. What is the probability that the star was famous. In other words, calculate $P(\text{Star}|\tilde{}\text{Good})$.

[5 marks]

Surname: _____

Forename(s): _____

$P(\text{Star}|\tilde{}\text{Good})=0.78$

```
P(F|~G) = P(F^~G)/P(~G)
P(~G):
P(A)=0.75
P(S)=0.1
P(F)=P(F|A)*P(A)+P(F|~A)*P(~A)
     0.9*0.75+0.5*0.25=0.8
P(E)=   P(E|A^S)*P(A)*P(S)           0.6*0.75*0.1
      + P(E|~A^S)*P(~A)*P(S)       + 0.4*0.25*0.1
      + P(E|~S)*P(~S)              + 0.8*0.9 = 0.775
P(D)=   P(D|F^~E)*P(F)*P(~E)         0.6*0.8*0.225
      + P(D|F^E)*P(F)*P(E)         + 0.2*0.8*0.775
      + P(D|~F^~E)*P(~F)*P(~E)     + 0.7*0.2*0.225
      + P(D|~F^E)*P(~F)*P(E)       + 0.1*0.2*0.775 = 0.279
P(G)=   P(G|D^F)*P(D)*P(F)           0.8*0.279*0.8
      + P(G|D^~F)*P(D)*P(~F)       + 0.7*0.279*0.2
      + P(G|~D)*P(~D)              + 0.2*0.721 = 0.36182
P(~G) = 1 - P(G) = 1 - 0.36182 = 0.63818

P(F^~G):
Given Famous
P(A)=0.75
P(S)=0.1
P(F)=P(F|A)*P(A)+P(F|~A)*P(~A)
     0.9*0.75+0.5*0.25=0.8
P(E)=   P(E|A^S)*P(A)*P(S)           0.6*0.75*0.1
      + P(E|~A^S)*P(~A)*P(S)       + 0.4*0.25*0.1
      + P(E|~S)*P(~S)              + 0.8*0.9 = 0.775
P(D)=   P(D|F^~E)*P(F)*P(~E)         0.6*0.225        # Given Famous!!!
      + P(D|F^E)*P(F)*P(E)         + 0.2*0.775 = 0.29
P(G)=   P(G|D^F)*P(D)*P(F)           0.8*0.29        # Given Famous
      + P(G|~D)*P(~D)              + 0.2*0.71 = 0.374
P(~G|F)= 1-P(G) = 1 - 0.374 = 0.626
P(F^~G) = 0.8*0.626 = 0.5008
P(F|~G) = P(F^~G)/P(~G) = 0.5008/0.63818 = 0.7847 = 78%
```

Section E: Instance-Based Learning

15. Given a domain with a single attribute `M`. The domain of `M` are integers in the interval 0 to 10.

Surname: _____

Forename(s): _____

Show the estimate of the target function of the `distance weighted Nearest Neighbor` algorithm. The target function estimate $f'(x_q)$ of the distance weighted Nearest Neighbor algorithm with cases C_1,C_2,...,C_n is given by:

The distance metric used is the absolute difference between instances:

$$d(I_1, I_2) = abs(I_1 - I_2)$$

$$f'(x_q) = \begin{cases} \dfrac{\sum_{i=1}^{n} f(C_i)K(d(x_q,C_i))}{\sum_{i=1}^{n} K(d(x_q,C_i))} & \text{if } \sum_{i=1}^{n} K(d(x_q,C_i)) \; != 0 \\ 0 & \text{else} \end{cases}$$

Use the kernel function K

$$K(d) = \begin{cases} 2 & \text{if } d = 0 \\ 1 & \text{if } d = 1 \\ 0.5 & \text{if } d = 2 \\ 0.25 & \text{if } d = 3 \\ 0 & \text{else} \end{cases}$$

The training data consists of the instances $C_1 : f(2) = 4$, $C_2 : f(4) = 1$, $C_3 : f(7) = 7$, and $C_4 : f(10) = 3$.

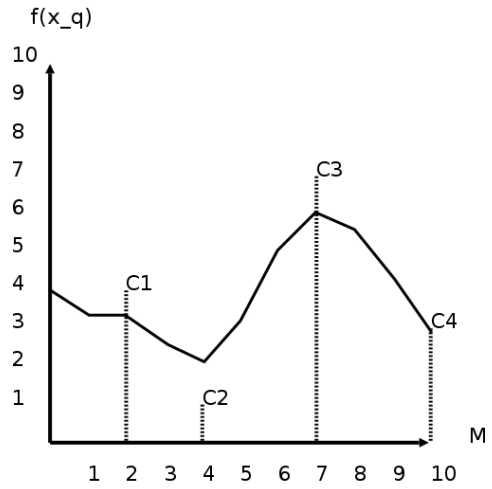
Calculate the target function estimate $f'(x_q)$ for the query points 0..10 using the distance weighted nearest neighbor algorithm and show the approximation in the figure below.

[10 marks]

Surname: _____

Forename(s): _____

x	0	1	2	3	4	5	6	7	8	9	10
f'(x)	4.0	3.39	3.39	2.5	2.09	3.14	5.0	6.0	5.56	4.33	3.44



16. Given below is the target function for an instance based learning algorithm. The classification uses the 1-nearest neighbor algorithm with the Manhattan distance as distance metric.

    The Manhattan distance between two instances $I_1 =$<n_1,m_1> and $I_2 =$<n_2,m_2> is defined as

$$d_{Man}(I_1, I_2) = abs(n_1 - n_2) + abs(m_1 - m_2)$$

    For example, the Manhattan distance between <F,3> and <D,4> is 3.

    If a square is equi-distant (i.e., has the same distance) to two or more cases, then a classification is calculated as the average of all equi-distant cases.

Surname: _____

Forename(s): _____

	1	2	3	4	5	6	7	8
A	7	7	7	7	6	4	2	2
B	7	7	7	7	6	4	2	2
C	7	7	7	7	6	4	2	2
D	7	7	7	7	6	4	2	2
E	6	6	6	6	5	3	2	2
F	6	6	6	6	5	3	2	3
G	6	6	6	6	5	3	3	3
H	6	6	6	6	5	3	3	3

What is the minimum number of cases necessary to represent the target function? Show the row, column and value of a set of cases that can represent the target function correctly.

If it is impossible to find a set of cases to represent the target function, then say so and explain why in your answer.

[10 marks]

The minimum number of cases to represent the target function correctly is 8.

Cases: C_1 =<D,4>=7, C_2 =<D,5>=6, C_3 =<D,6>=4, C_4 =<D,7>=2 C_5 =<E,5>=5 C_6 =<E,6>=3 C_7 =<E,7>=2 C_8 =<G,8>=3

Surname: _____

Forename(s): _____

Additional work pages

Surname: _____

Forename(s): _____

Additional work pages

Surname: _____

Forename(s): _____

Additional work pages