UNIVERSITY OF MANITOBA
Final Examination

Winter 2011

COMPUTER SCIENCE

Machine Learning

Paper No.:
Examiners: **Jacky Baltes**
Date: **Monday, 11th April 2011**
Time: **18:00**
Room: **Frank Kennedy, Brown Gym (184-208)**

(Time allowed: 180 Minutes)

NOTE: Attempt all questions.
This is a *closed* book examination.
Use of *non-programmable* 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

Section A: Candidate Elimination

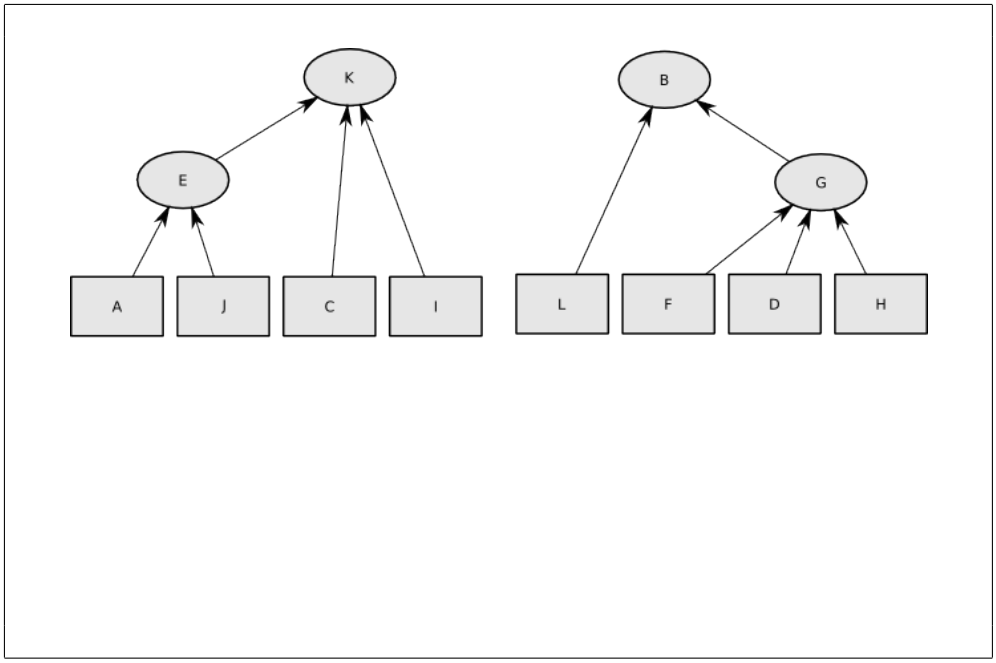
1. You are supposed to implement a simple machine learning system for a domain with two attributes. Unfortunately, the generalization hierarchy was lost in transmission. All you have is a trace of the candidate elimination algorithm as shown below.

   Reconstruct the generalization hierarchy for this domain given the trace below.

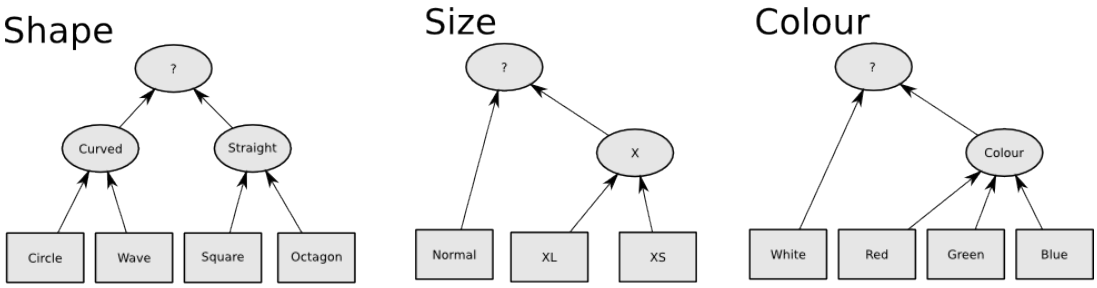
   If it is impossible to reconstruct the hierarchy, then say so in your answer and explain why.

Instance	Classification	S/G-set
<A,D>	+	S-set: <A,D>
		G-set: <K,B>
<J,L>	-	S-set: <A,D>
		G-set: <A,B> <K,G>
<C,L>	-	S-set: <A,D>
		G-set: <A,B> <K,G>
<J,D>	+	S-set: <E,D>
		G-set: <K,G>
<A,H>	+	S-set: <E,G>
		G-set: <K,G>
<A,F>	+	S-set: <E,G>
		G-set: <K,G>
<I,L>	-	S-set: <E,G>
		G-set: <K,G>
<C,F>	+	S-set: <K,G>
		G-set: <K,G>

[8 marks]


K
B
E
G
A
J
C
I
L
F
D
H

2. Given below is the generalization hierarchy H_2 with three attributes (Shape, Size, Colour).



What is the maximum number of possible instances for the generalization hierarchy H_2?

[1 mark]

There are a maximum of 4*4*3=48 different instances in H_2.

3. What is the maximum number of **semantically different concepts** that can be expressed in the generalization hierarchy H_2?

[1 mark]

There are 7*5*6+1=211 semantically different concepts in H_2.

4. Given is the generalization hierarchy H_2 shown above, and the target concept **<Curved,?,White>**. Show the minimum training set such that the candidate elimination algorithm will learn the target concept. For each sample in the training set show the classification, the resulting S-set and G-set. One entry of the training set is already given in the answer box below.

   If it is impossible to determine a minimum training set D such that the candidate elimination algorithm is able to learn the concept `<Curved,?,White>` say so in your answer and explain why.

[6 marks]

The minimum size of the training set is 4 samples.

Instance	Classification	S/G-set
<Circle,XL,Green>	-	S-set: <ni,nil,nil>
		G-set: <?,?,White>, <?,?,Red>, <?,?,Blue>
		<?,Normal,?>,<?,XS,?>
		<Wave,?,?>, <Straight,?,?>
<Circle,XL,White>	+	S-set: <Circle, XL, White>
		G-set: <?,?,White>
<Wave,Normal,White>	+	S-set: <Curved,?,White>
		G-set: <?,?,White>
<Square,Normal,White>	-	S-set: <Curved,?,White>
		G-set: <Curved,?,White>

5. Given the generalization hierarchy H_2 in question 2, assume that after a training sequence D, the candidate elimination algorithm returns the following version space V:

   S-set : <Wave,?,Colour>
   G-set : <?,?,?>

   What is the best next instance, that is the **instance that will maximally reduce the size** of the resulting version space independently of whether this instance will be classified as positive or negative.

   Show one maximally reducing instance as well as the resulting S and G sets if the instance is classified as positive or negative respectively.

   If there are more than one instance that lead to the same reduction in the size of the version space, then select any one of these instances.

   If it is impossible to determine the best next instance, then say so in your answer and explain why.

   [4 marks]

The best instance to classifiy next is: <Wave,Normal,White>
If this instance is classified as positive:
S-Set:<Wave,?,?>
G-Set:<?,?,?>
If this instance is classified as negative:
S-Set:<Wave,?,Colour>
G-Set:<?,?,Colour>

```
Hierarchy:                      Wave,?,Colour
        Curved,?,Colour                                  Wave,?,?
?,?,Colour                      Curved,?,?
                                ?,?,?

One example:
   <Wave,Normal,White>  3:3
```

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.

6. 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
F	F	F	Yes
F	F	T	No
F	T	F	No
T	F	T	No
T	T	T	No

Calculate the information gain (Gain$(S, ?)$) for the attributes A, B, and C. Which attribute would be selected by the standard ID3 algorithm?

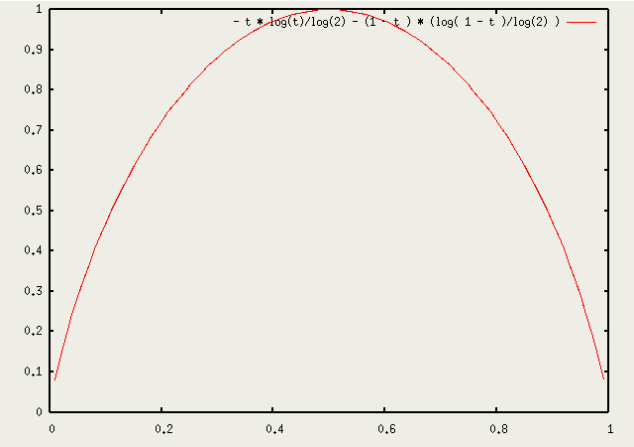


Figure 1: Graph of the Entropy Function

If it is impossible to calculate the information gain from the given information, then specify so in your answer and explain why.

[10 marks]

Gain$(S, \texttt{A}) = 0.17$

Gain$(S, \texttt{B}) = 0.17$

Gain$(S, \texttt{C}) = 0.32$

ID3 would select the attribute: C

```
Information Gain(A)=
  Entropy(1,4)[0.72192809488736231]-
    -3/5*Entropy(1,2)[0.91829583405448956] # A:f
    -2/5*Entropy(0,2)[0.0] # A:t
  =0.17095059445466865
Information Gain(B)=
  Entropy(1,4)[0.72192809488736231]-
    -3/5*Entropy(1,2)[0.91829583405448956] # B:f
    -2/5*Entropy(0,2)[0.0] # B:t
  =0.17095059445466865
Information Gain(C)=
  Entropy(1,4)[0.72192809488736231]-
    -2/5*Entropy(1,1)[1.0] # C:f
    -3/5*Entropy(0,3)[0.0] # C:t
  =0.32192809488736229
```

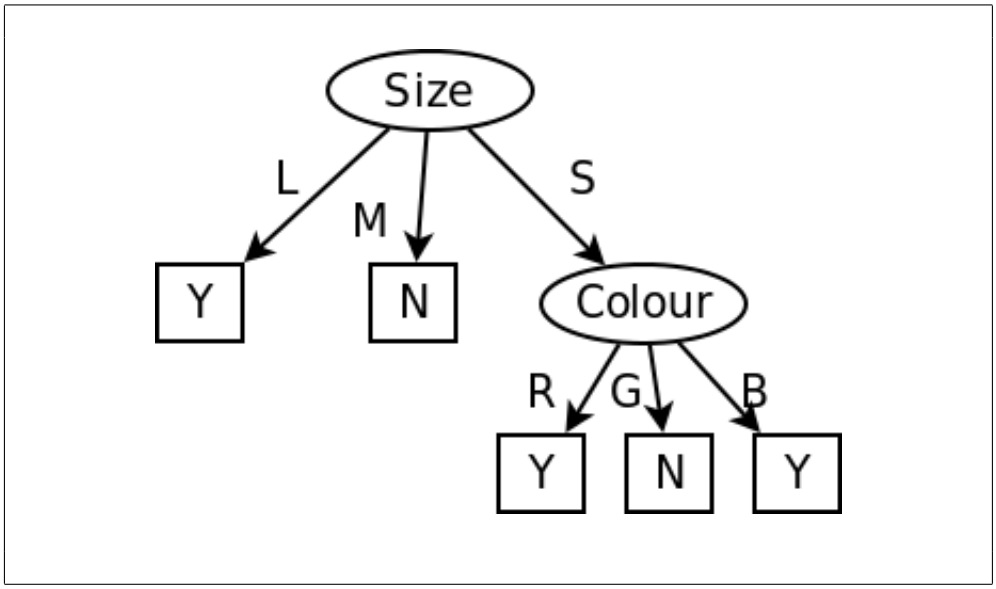
7. Given below is a set of rules. The rule set is **ordered**, that is as soon as all antecedents of a rule are satisfied, the classification is returned and evaluation of rules stops.

Rule 1:	If Shape=X and Size=M	⇒Yes
Rule 2:	If Size=L and Colour=R	⇒Yes
Rule 3:	If Size=M	⇒No
Rule 4:	If Size=S and Colour=R	⇒Yes
Rule 5:	If Size=S and Colour=B	⇒Yes
Rule 6:	If Shape=Z and Size=L	⇒Yes
Rule 7:	If Shape=S	⇒Yes
Rule 8:	If Colour=R or Colour=G or Colour=B	⇒Yes

Show the decision tree with the smallest number of nodes that is equivalent to the rule set shown above.

If it is impossible to convert this rules set into a decision tree, then say so and explain why in your answer.

[10 marks]



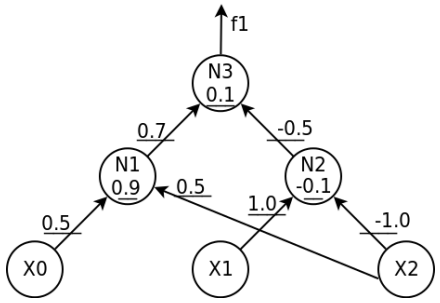
Section C: Neural Nets

8. Given below is an artificial neural network (ANN) with three input nodes (X0,X1,X2), 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 or equal to the threshold, 0 otherwise).

   Show a set of weights and thresholds for all nodes that implement the boolean function f_1 shown in the answer box below.

   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.

   [10 marks]

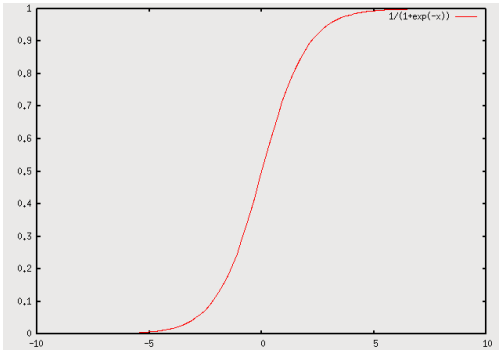


x_2	x_1	x_0	f_1
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

9. The output y of a sigmoid activation unit j is given by

$$a_i = \sum_{j=0}^{n} w_{ij} x_j$$
$$y_j = \frac{1}{1+e^{-a_j}}$$

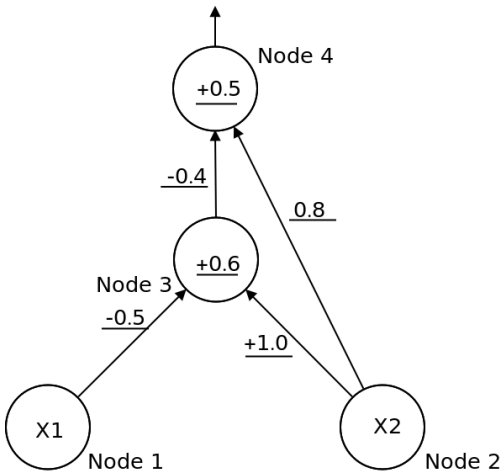
   The following graph shows the activation function for sigmoid activation units.



The backpropagation algorithm shown below is one of the most popular training algorithms for artificial neural networks (ANNs) with sigmoid activation functions.

Backpropagation(Output nodes y,Hidden nodes h, Weights w_{ij}, Training Set D)
$\forall w_{ij}$:= Initialize to small random value
do Until the termination condition has been met
  do $\forall\{< x_1, x_2, \ldots, x_n >, < t_1, t_2, \ldots, t_k >\} \in D$
    Apply $< x_1, x_2, \ldots, x_n >$ to the network and compute outputs $y_1, \ldots, y_k$
    do $\forall y \in$ Output nodes
      $\delta_y = y(1-y)(t-y)$
    od
    do $\forall h \in$ Hidden nodes
      $\delta_h = y_h(1-y_h)\sum_k w_{hk}\delta_k$
    od
    do $\forall w_{ij} \in$ Weights
      $w_{ij} = w_{ij} + \alpha\delta_j x_{ij}$
    od
  od
od

Given below is a small neural network with two sigmoid activation units (one output node and one hidden node), and two input nodes. The current weights and thresholds of the network are shown in the figure.



After applying a training instance $< x, t >$, the backpropagation algorithm is run with a learning rate of α and some of the weights are updated.

Given the information in the table below, compute the new weight between Node3 and Node 4 after applying the training instance $< x, t >$.

Learning rate	α	0.9
Training instance	$< x1, x2 >$	$< 1, 0 >$
Target value t	0.0	

If it is impossible to determine the new weight then say so in your answer and explain why.

[10 marks]

The new value for weight $3 \rightarrow 4$ is -4.035

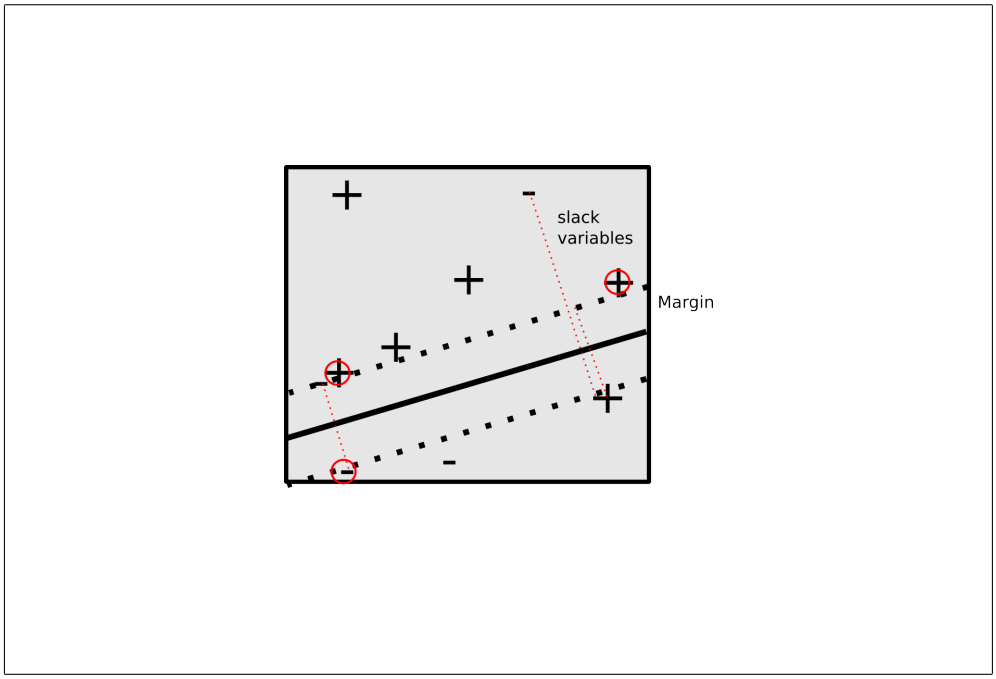
```
Activation of node 3: a_3 = -1.1, y_3=1/(1+exp(-1.1)) = 0.250
Activation of node 4: a_4 = -0.59, y_4=1/(1+exp(-0.59)) = 0.354
d4 = y(1-y)(t-y) = 0.35(1-0.35)(0.0-0.35)=0.147715
dw34=0.035
```

Section D: Miscalleneous

10. Given below is a two dimensional problem for linear soft margin support vector machines. One possible decision line is shown in the answer box below. Show the margin, the support vectors and all slack variables $\zeta_1, ...$ in the figure below.

    It if is impossible to show the margin, the support vectors, or the slack variables, then say so in your answer and explain why.

    [10 marks]



11. Given below is a sequence of random numbers for the random variable X_1. X_1 is distributed according to a normal distribution with unknown mean μ_1 and known variance $\sigma_1^2 = 2.0$.

    The probability density function (pdf) for a Gaussian distribution is given by

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

    The Expectation Maximization (EM) algorithm uses the following two steps

    Expectation:

$$\begin{aligned} E[z_{ij}] &= p(x = x_i|\mu = \mu_j) / \textstyle\sum_{n=1}^{q} p(x = x_i|\mu = \mu_j) \\ &= e^{-\frac{(x_i-\mu_j)^2}{2\sigma^2}} / \textstyle\sum_{n=1}^{q} e^{-\frac{(x_i-\mu_n)^2}{2\sigma^2}} \end{aligned}$$

    Maximization:

$$\mu_j = \sum_{i=1}^{m} E[z_{ij}] x_i / \sum_{i=1}^{m} E[z_{ij}]$$

CONTINUED

Calculate the best estimate for the parameter μ_1, given that the following points were drawn from X and their expected values $E[z_{i1}]$:

$$(x_i, E[z_{i1}]) = [(0.56, 0.87), (2.02, 0.88), (0.30, 0.78), (-0.35, 0.50), (1.15, 0.99), (0.18, 0.73), (0.09, 0.69)]$$

[5 marks]

The best estimate for μ_1 is 0.67

0.56*0.87+2.02*0.88+0.30*0.78-0.35*0.5+1.15*0.99+0.18*0.73+0.09*0.69=3.6558
0.87+0.88+0.78+0.5+0.99+0.73+0.69=5.44 3.6558/5.44=0.67

12. Assume that the current estimate for the mean of a random variable X_2 is $\mu_2 = 1.8$. Calculate the probability that the sample 0.6 was drawn from X_2.

    If it is impossible to calculate the probability, then say so in your answer and explain why.

    [5 marks]

The probability of selecting 0.6 is Unknown
since the variance σ_2 is unknown.

Section E: Bayesian Learning

13. Given below is a data set with some missing classifications. Is it possible to **uniquely** determine the missing classifications, given the fact that the output of a naive Bayesian classifier trained on the data set (shown below the training data) is known?

Training Data

Size	Colour	Target
S	white	Yes
S	green	No
M	white	Yes
M	green	Yes
L	black	?
L	green	No

Naive Bayes Output

Size	Colour	Naive Bayes
L	white	Yes

If it is impossible to calculate the missing classification such that the Naive Bayesian classifier will classify the instances as shown, then say so in your answer and explain why.

[5 marks]

Classification of <L,black> must have been: Unknown

```
Possibility: <L,Black>=Yes
P(Yes|L,White)=4/6*1/4*2/4=8/96
P(No|L,White)=2/6*1/2*0/2=0

Possibility: <L,Black>=No
P(Yes|L,White)=3/6*0/3*0/3=0
P(No|L,White)=3/6*2/3*0/3=0
```

14. You are given the following information about a TV show domain domain. The name of the associated random variables is given in brackets.

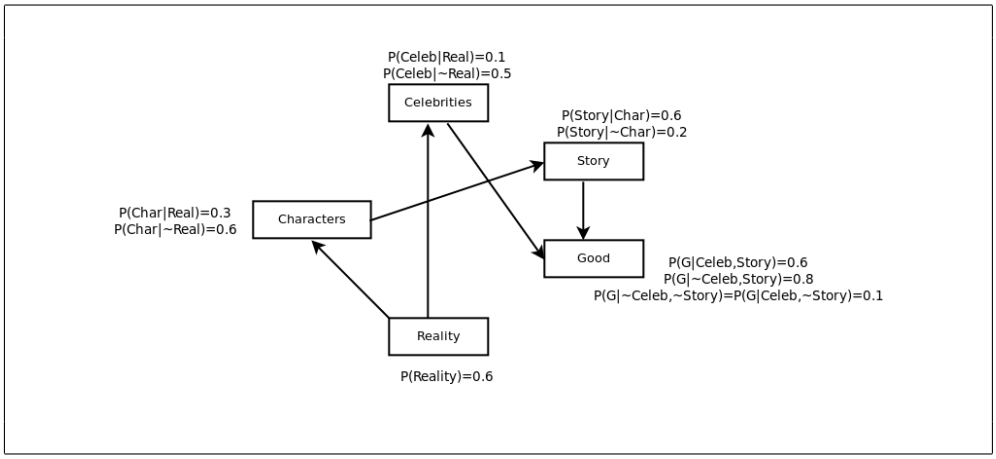
- (Reality) There is a 60% chance that a show on TV is a reality TV show.
- (Characters) If a show is a reality TV show, there is a 30% chance that it has interesting characters. For other shows, there is a 60% chance that it has interesting characters.
- (Story) If a show has interesting characters, there is a 60% chance that the story is good. A story without interesting characters has a 20% chance of being good.

- (Celebrities) Reality TV shows have a 10% chance of having celebrities, and other shows have a 50% chance of having celebrities.
- (Good) If a show has a good story and celebrities, then there is an 60% chance that the show is good. If a show has a good story and no celebrities, then there is an 80% chance that it is good. If a show has a bad story then there is a 10% chance that it is good.

Describe this information in the form of a Bayesian Belief Network (BBN). Show the graph of the BBN network as well as all the conditional probabilities for all random variables.

If it is impossible to derive the Bayesian Belief Network probability given the information above then say so in your answer and explain why.

[5 marks]



15. Without any other information, what is the probability of a TV show being good?

    If it is impossible to calculate this probability given the information above then say so in your answer and explain why.

    [5 marks]

$P(\text{Good}) = 33.8\%$

```
P(Real)=0.6
P(Char)=P(Char|Real)*P(Real)+P(Char|~Real)*P(~Real)=0.3*0.6+0.6*0.4=0.42
P(Celeb)=P(Celeb|Real)*P(Real)+P(Celeb|~Real)*P(~Real)=0.1*0.6+0.5*0.4=0.26
P(Story)=P(Story|Char)*P(Char)+P(Story|~Char)*P(~Char)=0.6*0.42+0.2*0.58=0.3
P(Good) = P(Good|Celeb,Story)*P(Celeb)*P(Story) + //0.6*0.26*0.368=0.057408
   P(Good|Celeb,~Story)*P(Celeb)*P(~Story) + //0.1*0.26*(1-0.368)=0.016432
   P(Good|~Celeb,Story)*P(~Celeb)*P(Story) + //0.8*(1-0.26)*0.368=0.217856
   P(Good|~Celeb,~Story)*P(~Celeb)*P(~Story) + //0.1*(1-0.26)*(1-0.368)=0.04
=0.057408+0.016432+0.217856+0.046768=0.338464
```

16. Would you tell your friend to watch a TV show, if you know that the characters are good. In other words, compare the probabilities of $P(\text{Good}|\text{Characters})$ and $P(\neg\text{Good}|\text{Characters})$.

    If it is impossible to calculate this probability given the information above then say so in your answer and explain why.

    [5 marks]

$P(\text{Good}|\text{Characters}) = 48.8\%$

```
: P(Char|Char)=1.0 # Given
: P(Story|Char)=0.6
: P(Good|Celeb,Story)=P(Good|Celeb,Story)*P(Celeb)*P(Story)
+ P(Good|~Celeb,Story)=0.8*(1-P(Celeb)*P(Story)
+ P(Good|Celeb,~Story)
+ P(Good|~Celeb,~Story)
  =
  0.6*0.26*0.6
+ 0.8*(1-0.26)*0.6
+ 0.1*0.4
```

Additional work pages