R228/406

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc (Hons) in Computer Science

Stage 4

SUPPLEMENTAL EXAMINATIONS 2010

*** *SOLUTIONS* ***

ARTIFICIAL INTELLIGENCE 2

Dr. John Kelleher
Dr. D. Lillis
Dr. I. Arana

Duration: 2 Hours

Answer Question 1 (40 marks) **and**

any 2 Other Questions (30 marks each).

*** *SOLUTIONS* ***

*** *SOLUTIONS* ***

1. (a) Given the full joint distribution shown in Table 1, calculate the following:

Table 1: Full joint distribution for a dentist visit

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

(i) $P(toothache)$

(5 marks)

This asks for the probability that $Toothache$ is true. $P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

(ii) $\mathbf{P}(Cavity)$

(5 marks)

This asks for the vector of probability values for the random variable $Cavity$. It has two values, which we list in the order $\langle true, false \rangle$. First add up $P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$. Then we have $textbf P(Cavity) = \langle 0.2, 0.8 \rangle$.

(iii) $\mathbf{P}(Toothache|cavity)$

(5 marks)

This asks for the vector of probability values for $Toothache$, given that $Cavity$ is true. $textbf P(Toothache|cavity) = \langle \frac{0.108+.012}{0.2}, \frac{0.072+0.008}{0.2} \rangle = \langle 0.6, 0.4 \rangle$

(iv) $\mathbf{P}(Cavity|toothache \vee catch)$

(5 marks)

This asks for the vector of probability values for $Cavity$, given that either $Toothache$ or $Catch$ is true.
Recall $P(a|b) = \frac{P(a \wedge b)}{P(b)} \rightarrow$
$\mathbf{P}(Cavity|toothache \vee catch) =$
$\langle \frac{P(cavity \wedge (toothache \vee cavity))}{P(toothache \vee catch)}, \frac{P(\neg cavity \wedge (toothache \vee cavity))}{P(toothache \vee catch)} \rangle$
First compute $P(toothache \vee catch) = 0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.144 = 0.416$.
Then $\mathbf{P}(Cavity|toothache \vee catch) =$
$\langle \frac{0.108+0.012+0.072}{0.416}, \frac{0.016+0.064+0.144}{0.416} \rangle = \langle 0.4615, 0.5384 \rangle$

(b) Describe the problems associated with measuring **classifier performance** using a single accuracy figure **and** describe a more appropriate alternative.

(20 marks)

This is a discursive question so giving a precise answer is not appropriate. However, an answer to this question should describe how a single accuracy figure can hide a classifier's real performance. An example should be provided such as the following:

- Text cases: 1000
- Positive examples: 900
- Negative examples: 100

Assuming the classifier always classifies positively then its accuracy on the given text set would be 90% which is not an accurate reflection of the classifiers performance.

The most obvious alternative would be to describe the use of **specificity**, **sensitivity** and **precision** along with a **confusion matrix**. Students should explain how a confusion matrix can be use as follows:

Classifier Results			
Class A (yes)	Class B (no)		
Correct	F_n	Class A (yes)	Expected Results
F_p	Correct	Class B (no)	

And finally it is expected that the students would describe specificity, sensitivity and precision as follows:

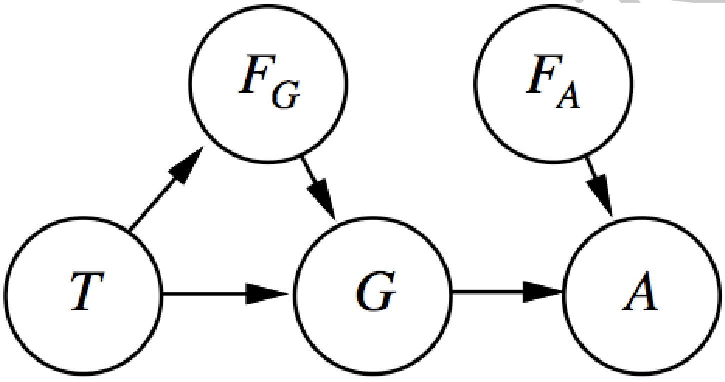
$sensitivity = \frac{t_{pos}}{pos}$ $specificity = \frac{t_{neg}}{neg}$ $precision = \frac{t_{pos}}{t_{pos}+f_{pos}}$

2. (a) In you local power station, there is an alarm that senses when a temperature gauge exceeds a given threshold. The gauge measures the temperature of the core of the power station. Consider the Boolean variables A (alarm sounds), F_A (alarm is faulty), and F_G (gauge is faulty); and multivalued nodes G (gauge reading) and T (actual core temperature).

(i) Draw a Bayesian network for this domain, given that the gauge is more likely to fail when the core temperature gets too high.

(5 marks)

The key aspects are: the failure nodes are parents of the sensor nodes, and the temperature node is a par ent of both the gauge and the gauge failure node. It is exactly this kind of correlation that makes it difficult for humans to understand what is happening in complex systems with unreliable sensors.



(ii) Suppose there are just two possible actual and measured temperatures, normal and high, and the probability that the gauge gives the correct temperature is x when it is working, but y when it is faulty. Give the conditional probability table associated with node G.

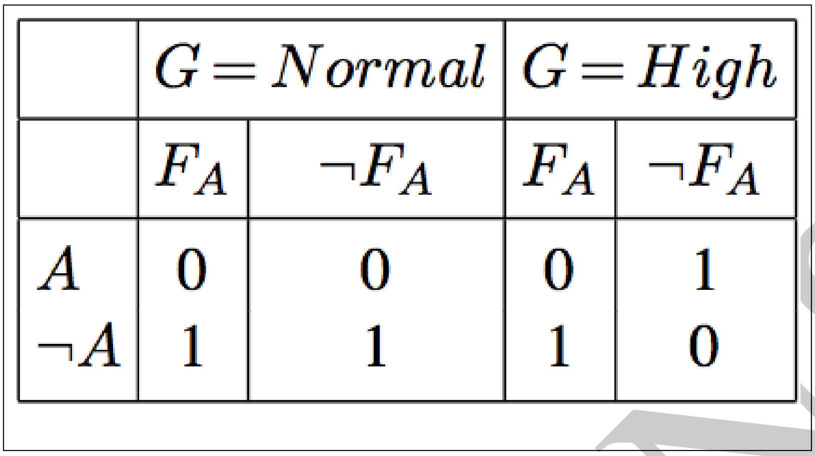
(5 marks)

Note the semantics of F_G , which is true when the gauge is faulty, i.e., not working.

	$T = Normal$		$T = High$	
	F_G	$\neg F_G$	F_G	$\neg F_G$
$G = Normal$	y	x	$1-y$	$1-x$
$G = High$	$1-y$	$1-x$	y	x

(iii) Suppose the alarm works correctly unless it is faulty, in which case it never sounds. Give the conditional probability table associated with A.

(5 marks)



	$G = Normal$		$G = High$	
	F_A	$\neg F_A$	F_A	$\neg F_A$
A	0	0	0	1
$\neg A$	1	1	1	0

(b) Consider the following time keeping patterns of the lecturers in your college:

- 25% of lecturers start 75% of their lectures on time and 25% late.
- 50% of lecturers start 50% of their lectures on time and 50% late.
- 25% of lecturers start 25% of their lectures on time and 75% late.

(i) Given that both the 1^{st} and 2^{nd} Artificial Intelligence lectures of the year started on time, compute the posterior probability that your Artificial Intelligence lecturer follows that each of the three time-keeping patterns.

(5 marks)

To begin we will define some notation. Let:

- h_1 denote the hypothesis that your AI lecturer starts 75% of their lectures on time $P(h_1) = 0.25$.
- h_2 denote the hypothesis that your AI lecturer starts 50% of their lectures on time $P(h_2) = 0.50$.
- h_3 denote the hypothesis that your AI lecturer starts 25% of their lectures on time $P(h_3) = 0.25$.

Also, if we use the notation $ontime_x$ to represent the observation that a lecture x started on time, then the probability of any given AI lecture starting on time given a particular hypothesis h is:

- $P(ontime_x|h_1) = 0.75$.
- $P(ontime_x|h_2) = 0.50$.
- $P(ontime_x|h_3) = 0.25$.

Then:

- By Bayes' rule, we can compute the posterior probability of a hypothesis given the data so far using:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

- And, the likelihood of the data given a hypothesis is calculated using:

$$P(\mathbf{d}|h_i) = \prod_j P(d_j|h_i)$$

So:

- $P(h_1|ontime_1, ontime_2) = \alpha(\prod_{j=1}^{2} P(ontime_j|h_1))P(h_1) = \alpha 0.75^2 \times 0.25 = \alpha 0.375 = \frac{0.375}{1.0} = 0.375$.
- $P(h_2|ontime_1, ontime_2) = \alpha(\prod_{j=1}^{2} P(ontime_j|h_2))P(h_1) = \alpha 0.50^2 \times 0.50 = \alpha 0.500 = \frac{0.500}{1.0} = 0.500$.
- $P(h_3|ontime_1, ontime_2) = \alpha(\prod_{j=1}^{2} P(ontime_j|h_3))P(h_1) = \alpha 0.25^2 \times 0.25 = \alpha 0.125 = \frac{0.125}{1.0} = 0.125$.

(ii) Given that both the 1^{st} and 2^{nd} Artificial Intelligence lectures of the year started on time, what is the Bayesian Prediction that the 3^{rd} Artificial Intelligence lecture will start on time?

(5 marks)

Bayesian predictions use a likelihood-weighted sum over the hypotheses:

$$P(X|\mathbf{d}) = \textstyle\sum_i P(X|h_i)P(h_i|\mathbf{d})$$

In this instance we get:

$$\begin{aligned} P(ontime_3|\mathbf{d}) &= \sum_i P(ontime_3|h_i)P(h_i|\mathbf{d}) \\ &= (0.75 * 0.375) + (0.5 * 0.5) + (0.25 * 0.125) \\ &= 0.28125 + 0.25 + 0.03125 \\ &= 0.5625 \end{aligned}$$

(iii) Given that both the 1^{st} and 2^{nd} Artificial Intelligence lectures of the year started on time, what is the Maximum a Posterior (MAP) probability that the 3^{rd} Artificial Intelligence lecture will start on time?

(5 marks)

A MAP prediction just uses the prediction provided by the single most probable hypothesis. From part one of the question we can see that given that both the 1^{st} and 2^{nd} lectures started on time the single most probable hypothesis is the hypothesis that the Artificial Intelligence lecturer will start 50% of their lectures on time (from part 1, $(P(h_2|ontime_1, ontime_2) = 0.5) > (P(h_2|ontime_1, ontime_2) = 0.375) > (P(h_3|ontime_1, ontime_2) = 0.125)$. This hypothesis would predict that the 3^{rd} AI lecture will start on time with a probability of 0.5.

3. (a) Define what is meant by **lazy learners** and **eager learners**, highlight the key differences between these approaches **and** give an example of each.

(10 marks)

Definitions:

Lazy learners do not try to build a model from the training data, but simply use it at classification time

Eager learners build a mode from the training data during training, and use only this model at classification time, ignoring the original data.

Key differences:

- Lazy methods may consider query instance when deciding how to generalise beyond the training data D; eager methods cannot since they have already chosen global approximation when seeing the query.
- **Efficiency** lazy learners require less training times but more time at prediction; eager learners require more training times by less time for prediction
- **Accuracy** lazy learners effectively uses a richer hypothesis space since it uses many local linear functions to form its implicit global approximation to the target function; eager learners must commit to a single hypothesis that covers the entire instance space.
- It is easier for lazy learners to handle **concept drift**

Examples:

Lazy learning example : case based reasoning

Eager learning example : Decision=tree, neural networks, support vector machines

(b) Formally define what is meant by the term **entropy** in the context of Decision Tree Learning.

(5 marks)

For c classification categories the entropy E is defined as: $E = \sum_{i=1}^{c} -p_i \; log_2 \; p_i$ where p_i is the probability of category i occurring.

(c) You are working as an assistant-biologist to Charles Darwin on the Beagle voyage. You are at the Galápagos Islands and you have just discovered a new animal that has not yet been classified. Table 2 lists the attributes of the animal you have found. Mr. Darwin has asked you to classify the animal using a nearest-neighbour approach and he has supplied you with a case-base of already classified animals, see Table 3.

Table 2: The attributes of a newly discovered animal. A 1 indicates the animal possesses the feature listed in the column and 0 indicates that they do not. The column on the right contains a ? because the animal has not been classified yet.

Species	Births Live Young	Lays Eggs	Feeds Offspring Own Milk	Warm-Blooded	Cold-Blooded	Lives in Water and Land	Has Hair	Has Feathers	Class
Mystery	0	1	0	0	0	1	0	0	?

Table 3: Example feature vectors for animal classification. A 1 indicates the animal possesses the feature listed in the column and 0 indicates that they do not. The rightmost column lists the classification of each ainmal.

Species	Births Live Young	Lays Eggs	Feeds Offspring Own Milk	Warm-Blooded	Cold-Blooded	Lives in Water and Land	Has Hair	Has Feathers	Class
Cat	1	0	1	1	0	0	1	0	Mammal
Frog	0	1	0	0	1	1	0	0	Amphibian
Squirrel	1	0	1	1	0	0	1	0	Mammal
Duck	0	1	0	1	0	1	0	1	Bird

(i) A good measure of distance between two instances with categorical features is the number of features which have different values (the **overlap metric**, also known as the **hamming distance**). Using this measure of distance compute the distance between the mystery animal and each of the animals in the case base.

(5 marks)

Species	Class	Distance
Cat	Mammal	6
Frog	Amphibian	1
Squirrel	Mammal	6
Duck	Bird	2

(ii) If you used $1 - NN$ classification what class would be assigned to the mystery animal.

(5 marks)

The nearest neighbor to the mystery animal is the Frog. So the mystery animal would be classified as an amphibian.

(iii) If the you used $4 - NN$ classification what class would be assigned to the mystery animal.

(5 marks)

If you applied a $4 - NN$ classification to this case-base you would include all the instances in the case-base irrespective of their distance from the test instance feature vector. As a result the test instance would be assigned the most frequently occurring class in the case-base. This would result in the mystery animal being classified as a mammal.

4. (a) Figure 1, on the next page, is a schematic of a 3 input perceptron. Input $a_0 = -1$, inputs a_1 and a_2 are binary. The perceptron uses a threshold activation function that outputs a 1 if the weighted sum of inputs is greater than 0 and a 0 otherwise. Define the truth-table of the function that this perceptron implements and identify the name of the function.

(5 marks)

Inputs			in_i	a_i
a_0	a_1	a_2	$\sum_{j=0}^{2} wjia_j$	$(in_i > 0)?(1) : (0)$
-1	1	1	0.5	1
-1	1	0	-0.5	0
-1	0	1	-0.5	0
-1	0	0	-1.5	0

This perceptron implements the AND function.

(b) Describe the perceptron training rule?

(10 marks)

Perceptrons learn by modifying the weights associated with their inputs. Consequently, the learning problem faced by perceptron training is to determine a weight vector that causes the perceptron to produce the correct $+-1$ output for each of the given training examples.

One way to learn an acceptable weight vector is to begin with random weights, then iteratively apply the perceptron to each training example, modifying the perceptron weights whenever it misclassifies an example.

Weights are modified at each step according to the *perceptron training rule*, which revises the weight w_i associated with input a_i according to the rule:

$$w_i \leftarrow w_i + \Delta w_i$$

where

$$\Delta w_i = \eta(t - o)a_i$$

where t = target output, o = observed output, η is a positive constant called the *learning rate*. The role of the learning rate is to moderate the degree to which weights are changed at each step. It is usually set to some small value (e.g.0.1) and is sometimes made to decay as the number of weight-training iterations increases.

(c) Why does the perceptron training rule converge toward successful weight values?

(5 marks)

If the training example is *correctly classified* $(t - o) = 0 \rightarrow \Delta w_i = 0$ so *no weights are updated*.

If the case of a *false negative* (o=0 and t=1) we want to make the perceptron output a 1 instead of a 0 so the weights must be altered to increase the value of $\vec{w} \bullet \vec{a}$. Notice that in this case *the rule will increase* w_i because $(t - o), \eta$ and a_i are all positive.

On the other hand, in the case of a *false positive* (o=1 and t=0) then *the weights associated with* a_i *will be decreased*.

(d) The FOIL inductive logic programming algorithm is constructing a new rule with head $p(Y) \leftarrow$. Which of the following literals could be considered as candidate extensions $q(Y)$, $r(X)$, $s(X,Y)$, $\neg s(X,Y)$?

(5 marks)

Three of the given literals could be considered as extensions: $q(Y)$, $s(X,Y)$, $\neg s(X,Y)$. The literal $r(X)$ would not be considered as an extension as it does not contain at least one variable that is already present in the rule.

(e) The FOIL inductive logic programming algorithm is considering adding a literal l to a rule r. The extension of the rule r before adding l is the following set of positive and negative examples $\{+,+,+,+,+,+,-,-,-,-\}$. The extension of the rule after the literal is added (i.e. the extension of $r + l$) is $\{+,+,+,+,+,+,-,-\}$. What is the information gain of adding the literal l to the rule r? (Note you do not need to calculate the logs in your answer (i.e. you can express your answer as an equation containing logs)).

(5 marks)

$$Foil\_Gain(L,R) \equiv t\left(\log_2 \frac{p_1}{p_1+n_1} - \log_2 \frac{p_0}{p_0+n_0}\right)$$

Where

- L is the candidate literal to add to rule R
- p_0 = number of positive bindings of R
- n_0 = number of negative bindings of R
- p_1 = number of positive bindings of $R+L$
- n_1 = number of negative bindings of $R+L$
- t is the number of positive bindings of R also covered by $R+L$

In this instance

- p_0 = 6
- n_0 = 4
- p_1 = 6
- n_1 = 2
- t = 6

$$Foil\_Gain(L,R) \equiv 6\left(\log_2 \frac{6}{6+4} - \log_2 \frac{6}{6+2}\right)$$

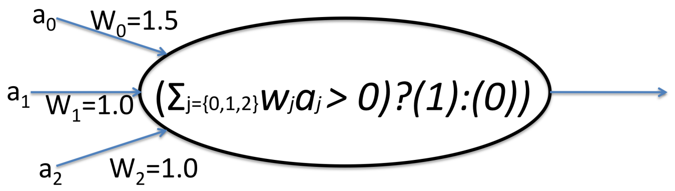


Figure 1: A 3 input perceptron. Input $a_0 = -1$, inputs a_1 and a_2 are binary. The perceptron uses a threshold activation function that outputs a 1 if the weighted sum of inputs is greater than 0 and a 0 otherwise.