S228/406

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc (Hons) in Computer Science

Stage 4

SEMESTER 2 EXAMINATIONS 2011

*** *SOLUTIONS* ***

ARTIFICIAL INTELLIGENCE II

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

Friday 20th May
4:00 p.m. to 6:00 p.m.

Duration: 2 Hours

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

any 2 Other Questions (30 marks each).

****** SOLUTIONS ******

****** SOLUTIONS ******

1. (a) Explain what is meant by **inductive learning**.

(5 marks)

Inductive Learning involves the process of learning by example where a system tries to induce a general rule from a set of observed instances

(b) In the context of machine learning, distinguish between **supervised** and **unsupervised** learning.

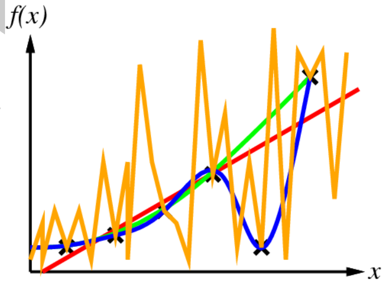
(5 marks)

The distinction is that with **supervised learning** we know the actual label or category for each piece of data on which we train, whereas with **unsupervised learning** we do not know the classification of the data in the training sample. Unsupervised learning can thus often be viewed as a **clustering** task, while supervised learning can usually be seen as a **classification** task, or equivalently as a function-fitting task where one extrapolates the shape of a function based on some data points.

(c) Inductive machine learning is often referred to as an **ill-posed problem**. What is meant by this description?

(10 marks)

Inductive machine learning algorithms essentially search through a hypothesis space to find a the best hypothesis that is consistent with the training data used. It is possible to find multiple hypotheses that are consistent with a given training set (i.e. agrees with all training examples). It is for this reason that inductive machine learning is referred to as an ill-posed problem as there is typically not enough information in the training data used to build a model to choose a single best hypothesis. Inductive machine learning algorithms must somehow choose one of the available hypotheses as the *best*. An example like that shown in the figure below would be useful at this point



(d) Let us say we have three classification algorithms. How can we order these three from best to worst?

(20 marks)

This is a discursive question so giving a precise answer is not appropriate. However, key points that the student should touch on include:

- Predictive accuracy
- Speed and scalability
  - Time to construct the model
  - Time to use the model
- Robustness (handling noise and missing values)
- Scalability
- Interpretability (understanding and insight provided by the model)

It should be noted also, that these evaluation criteria are application dependent.

Table 1: Example feature vectors for animal classification. A 1 indicates the animal possesses the feature listed in the column, and 0 indicates 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	Land and Water Based	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

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

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

2. (a) You are working as an assistant-biologist to the 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 1.

   (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 distances 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.

(b) Table 3 provides a classification for a data set of X Y pairs.

(i) Calculate the **entropy** for this classification.

(5 marks)

Entropy is $-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$

(ii) Calculate the **information gain** for X and Y.

(5 marks)

Entropy for X = T $-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} = 0.811$
Entropy for X = F $0 - \frac{1}{1}log_2\frac{1}{1} = 0$
Gain for X $0.971 - (\frac{4}{5} \times 0.811 + \frac{1}{5} \times 0) = 0.322$
Entropy for Y = T $-\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.918$
Entropy for Y = F $-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2} = 1.0$
Gain for Y $0.971 - (\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1) = 0.02$

X	Y	Class
T	T	+
T	F	-
T	F	+
T	T	+
F	T	-

Table 3: X and Y Classification Data

(c) The following sets express the mappings between predicates r, p, q, s, $class1$ and $class2$:

- $r \rightarrow \{a1, a2, a5, a6\}$,
- $p \rightarrow \{a2, a3, a5, a7\}$,
- $q \rightarrow \{a1, a2, a6\}$,
- $s \rightarrow \{(a2, f), (a1, 1), (a6, f)\}$,
- $class1 \rightarrow \{a2\}$,
- $class2 \rightarrow \{a2, a6\}$.

Given these sets, give a specialisation of the rule $class1(X) \leftarrow r(X) \wedge p(X)$ such that the rule is only satisfied by $class1$ members.

(5 marks)

$class1(X) \leftarrow r(X) \wedge p(X) \wedge q(X)$

Table 4: Joint Distribution for X and Y

	$X = x_1$	$X = x_2$
$Y = y_1$	0.02	0.30
$Y = y_2$	0.14	0.32
$Y = y_3$	0.10	0.12

3. (a) Given the joint distribution for X and Y listed in Table 4 calculate:

$$P(Y = y_2 | X = x_1)$$

(5 marks)

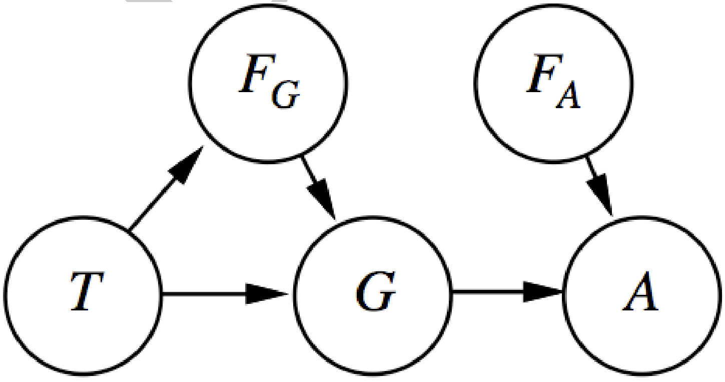
From the product rule: $P(a|b) = \frac{P(a \wedge b)}{P(b)} \rightarrow$
$P(Y = y_2 | X = x_1) = \frac{P(Y=y_2 \wedge X=x_1)}{P(X=x_1)} \rightarrow$
$P(Y = y_2 | X = x_1) = \frac{0.14}{0.26}$

(b) 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. 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)

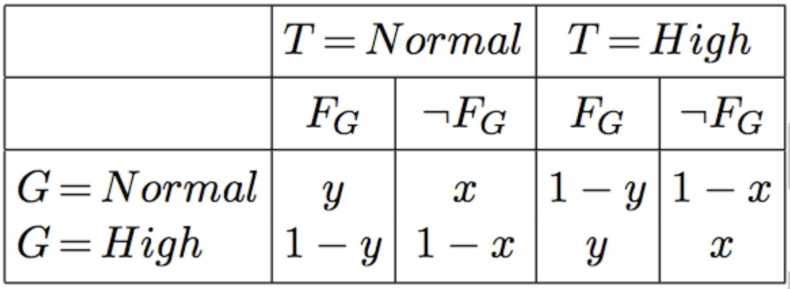
The key aspects are: the failure nodes are parents of the sensor nodes, and the temperature node is a parent of both the gauge and the gauge failure node. It is exactly this kind of correlation that makes it difcult for humans to understand what is happening in complex systems with unreliable sensors.



(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

(c) 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 each of the three time keeping patterns.

(10 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:

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

In this instance we get:

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

4. (a) The following model is commonly used for continuous prediction tasks:

$$y(x) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

(i) Provide the name for this model and explain all terms.

(5 marks)

Students should explain that this is a simple linear regression model which can be effectively used to make predictions. x is a vector of feature values for a query instance and w is a vector of feature weights. An diagram of a simple one dimensional linear function would help.

(ii) Briefly describe a technique for finding optimal values for the terms $w_0, w_1, \ldots, w_D$ in the model based on a historical training set.

(5 marks)

Students should explain that given a training set the performance of a particular linear regression model can be measured using an appropriate evaluations function, e.g. the *sum of squares error* as follows:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n) - t_n)^2$$

where t_n is the *true* answer for training instance x_n and $y(x_n)$ is the prediction made by the model for instance x_n. Optimal values for w are found by minimising E(w).

(b) Figure 1 shows a backprogation network that is currently processing the training vector $[1.0, 0.9, 0.9]$ which has an associated target vector $[0.1, 0.9, 0.1]$. Given that the output from unit B is 0.6 and from C is 0.8, and assuming that the activation function used at all nodes in the network is the logistic function (i.e., $f(x) = \frac{1}{1+\exp^{-x}}$):

(i) Calculate the actual output vector (to 3 decimal places).

(5 marks)

Output of unit $i = f(\sum_{j=1}^{n} W_{j,i} \times activation_j)$
First output unit input = -0.3 x 0.6 + 0.9 x 0.8 = 0.54 → f(0.54) = 0.632
Second output unit input = -0.6 x 0.6 + -0.1 x 0.8 = -0.44 → f(-0.44) = 0.392
Third output unit input = 0.4 x 0.6 + 1.2 x 0.8 = 1.2 → f(1.2)= 0.769

(ii) Calculate the error for each output unit.

(5 marks)

Error = target - output
First output unit = (0.1 - 0.632) = - 0.532
Second output unit = (0.9 - 0.392) = 0.508
Third output unit = (0.1 - 0.769) = - 0.696

(iii) Calculate the error for each hidden unit B and C.

(10 marks)

Each hidden node j is responsible for some fraction of the error Err_i of each of the output units i to which it connects. Thus the Err_i values are divided according to the strengths of the connection between the hidden node and the output nodes and are propagated back to the hidden nodes. Where a hidden node feeds-forward into more than 1 output node the errors propagated back to it are summed: $Err_j = \sum_{i=1}^{n} W_{ji} \times Err_i$:
$Err_B = (-0.3 \times -0.532) + (-0.6 \times 0.508) + (0.4 \times -0.696) = 0.1596 + -0.3048 + -0.2784 = -0.4236$
$Err_C = (0.9 \times -0.532) + (-0.1 \times 0.508) + (1.2 \times -0.696) = -0.4788 + -0.0508 + -0.8352 = -1.3648$

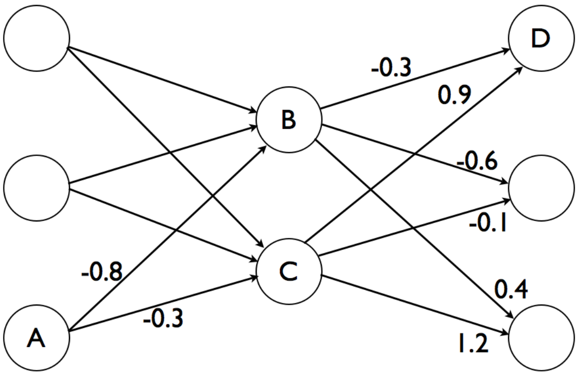


Figure 1: Example Neural Net