R228/406

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc (Hons) in Computer Science

Stage 4

SUPPLEMENTAL EXAMINATIONS 2011

*** *SOLUTIONS* ***

ARTIFICIAL INTELLIGENCE II

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) 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, explain what is meant by **overfitting** the training data.

(5 marks)

Overfitting occurs when classifiers make decisions based on accidental properties of the training set that will lead to errors on the test set (or new data). As a result, whenever there is a large set of possible hypotheses, one has to be careful not to use the resulting freedom to find meaningless "regularity" in the data.

(c) In the context of machine learning, explain what is meant by the term **inductive bias** and illustrate your explaination using examples of inductive biases used by machine learning algorithms.

(10 marks)

- The inductive bias of a learning algorithm:
  - (i) is a set of assumption about what the true function we are trying to model looks like.
  - (ii) defines the set of hypotheses that a learning algorithm considers when it is learning.
  - (iii) guides the learning algorithm to prefer one hypothesis (i.e. the hypothesis that best fits with the assumptions) over the others.
  - (iv) is a necessary prerequisite for learning to happen because inductive learning is an ill posed problem.
- An example of the specific inductive bias introduced by particular machine learning algorithms would be good here. E.g.:
  - Maximum margin: when drawing a boundary between two classes, attempt to maximize the width of the boundary. This is the bias used in Support Vector Machines. The assumption is that distinct classes tend to be separated by wide boundaries.
  - Minimum cross-validation error: when trying to choose among hypotheses, select the hypothesis with the lowest cross-validation error.

(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) In the context of Decision Tree Learning define what is meant by the following terms:

(i) entropy

(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.

(ii) information gain

(5 marks)

The information gain for an attribute is the expected reduction in entropy if the examples were to be partitioned according to that attribute and is defined as: $Gain(T, A) = E(T) - \sum_{j=1}^{v} \frac{|T_j|}{|T|} E(T_j)$ where T is a set of training examples and T_j is a subset of examples having value j for attribute A

(c) 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.

3. (a) Given that $P(a|b) = 0.5$, $P(a) = 0.3$, $P(b) = 0.4$ calculate $P(b|a)$.

(5 marks)

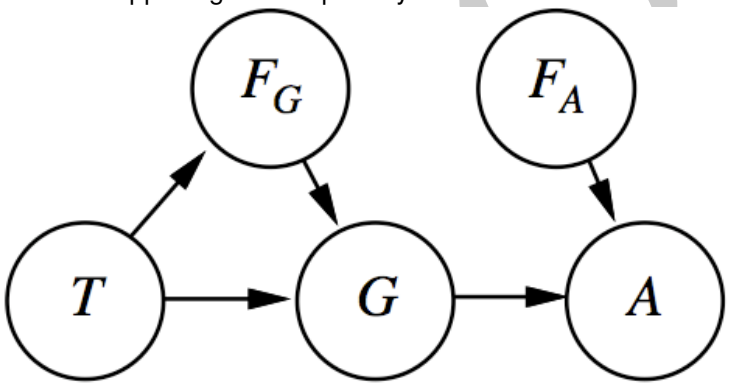
$P(b|a) = \frac{P(a|b) \times P(b)}{P(a)} = \frac{0.5 \times 0.4}{0.3} = 0.67$

(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) You are on holidays on Fisher Island. The yearly weather on Fisher Island comes in five different varieties:

- there is a 10% chance that there will be rain everyday of the year.
- there is a 20% chance that there will be rain on 75% of the days of the year.
- there is a 40% chance that there will be rain on 50% of the days of the year.
- there is a 20% chance that there will be rain on 25% of the days of the year.
- there is a 10% chance that there will be no rain on any day of the year.

(i) Given that it has rained on day 1 and 2 of the year compute the posterior probability of each of the 5 yearly weather patterns on day 2 of the year. Give your answer rounded to four places of precision.

(10 marks)

To begin we will define some notation. Let:

- h_1 denote the hypothesis that it will rain everyday, $P(h_1) = 0.1$.
- h_2 denote the hypothesis that it will rain on 75% of the days of the year, with prior $P(h_2) = 0.2$.
- h_3 denote the hypothesis that it will rain on 50% of the days of the year, with prior $P(h_3) = 0.4$.
- h_4 denote the hypothesis that it will rain on 25% of the days of the year, with prior $P(h_4) = 0.2$.
- h_5 denote the hypothesis that there will be no rain during the year, with prior $P(h_5) = 0.1$.

Also, if we use the notation $rain_x$ to represent the observation of rain on day x of the year, then the probability of rain on a day of the year given a particular hypothesis h is:

- $P(rain_x|h_1) = 1.0$.
- $P(rain_x|h_2) = 0.75$.
- $P(rain_x|h_3) = 0.5$.
- $P(rain_x|h_4) = 0.25$.
- $P(rain_x|h_5) = 0.0$.

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:

- $P(h_1|rain_1, rain_2) = \alpha(\prod_{j=1}^{2} P(rain_j|h_1))P(h_1) = \alpha 1.00^2 \times 0.1 = \alpha 0.1 = \frac{0.1}{0.325} \approx .3077.$
- $P(h_2|rain_1, rain_2) = \alpha(\prod_{j=1}^{2} P(rain_j|h_2))P(h_1) = \alpha 0.75^2 \times 0.2 = \alpha 0.1125 = \frac{0.1125}{0.325} \approx .3461.$
- $P(h_3|rain_1, rain_2) = \alpha(\prod_{j=1}^{2} P(rain_j|h_3))P(h_1) = \alpha 0.50^2 \times 0.4 = \alpha 0.1 = \frac{0.1}{0.325} \approx .3077.$
- $P(h_4|rain_1, rain_2) = \alpha(\prod_{j=1}^{2} P(rain_j|h_4))P(h_1) = \alpha 0.25^2 \times 0.2 = \alpha 0.0125 = \frac{0.0125}{0.325} \approx .0385.$
- $P(h_5|rain_1, rain_2) = \alpha(\prod_{j=1}^{2} P(rain_j|h_5))P(h_1) = \alpha 0.00^2 \times 0.1 = \alpha 0.0 = 0.0.$

(ii) Given that after the first 10 days of the year the weather has been such that the posterior probabilities of each of the 5 varieties of the yearly weather on Fisher Island are:

- there is now a 90% chance that there will be rain everyday for the rest of the year;
- a 7% chance that there will be rain on 75% of the rest of the days of the year;
- a 2% chance that there will be rain on 50% of the rest of the days of the year;
- a 1% chance that there will be rain on 25% of the rest of the days of the year;
- and there is a 0% chance that there will be no rain for the rest of the year.

What is the Maximum a Posterior (MAP) probability of rain on day 11?

(5 marks)

A MAP prediction just uses the prediction provided by the single most probable hypothesis. In this instance the single most probable hypothesis is the hypothesis that it will rain on every day of the year. This hypothesis would predict rain on day 11 with probability of 1.0 (i.e. certainty)

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) Explain how the following model can overcome some of the limitations of the model given above.

$$y(x) = \sum_{j=0}^{M-1} w_j \phi_j(x)$$

(5 marks)

Students should explain that the simple linear regression model is attractive because it is linear with respect to w but has severe limitations because it is also linear with respect to x. These greatly limits the kinds of predictions that this model will be able to make. However, the introduction of *basis functions*, shown as ϕ above, goes some way towards solving this problem. The introduction of a non-linear basis function means that models can be made non-linear functions of input x but remain linear in w which makes them computationally easier to solve. Students might give the example of polynomial regression in which $\phi_j(x) = x^j$ or some other suitable example.

(b) What does it mean if two classes C_1 and C_2 are described as **linearly separable**.

(5 marks)

This means that for each class C_i there exists a hyperplane H_i such that on its positive side lie all $x \in C_i$ and on its negative side lie all $x \in C_j, j \neq i$

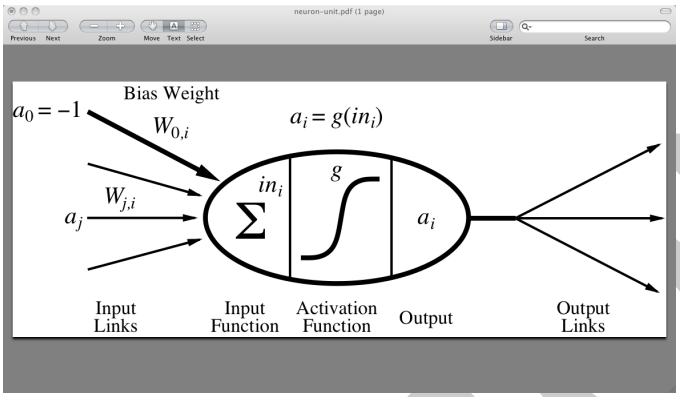
(c) Describe the processing stages of a McCulloch-Pits "unit".

(7 marks)

The processing stages of a unit are:

(i) Each unit i first compute a weighted sum of its inputs: $in_i \leftarrow \sum_j W_{j,i} a_j$

(ii) Then it applies an **activation function** g to this sum to derive the output (activation) a_i: $a_i \leftarrow g(in_i) = g\left(\sum_j W_{j,i} a_j\right)$



(d) Figure 1 is a schematic of a 3 input perceptron. Input a_0 is fixed at $a_0 = -1$, 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**.

(8 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.

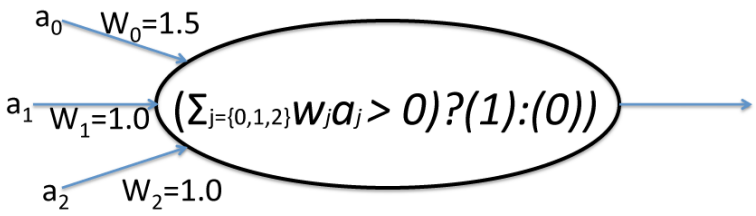


Figure 1: A 3 input perceptron. Input $a_0 = -1$, inputs a_1 and a_2 are binary. The perceptron uses a threshold activation function that outputs a 1 if the weighted sum of inputs is greater than 0 and a 0 otherwise.