R249/419C

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc. (Hons) in Computer Science

Stage 4

SUPPLEMENTAL EXAMINATIONS 2013

ARTIFICIAL INTELLIGENCE II

Dr. John Kelleher
Dr. Deirdre. Lillis
Mr. R. Walshe

Duration: 2 Hours

Question 1 is **compulsory**

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

any 2 Other Questions (30 marks each).

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

(5 marks)

(b) Describe the differences between **lazy learners** and **eager learners**, giving examples of each.

(10 marks)

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

(10 marks)

(d) Why is it difficult to select the correct inductive bias for a machine learning algorithm?

(15 marks)

Table 1: A dataset showing the behaviour of two individuals in an online shop. A 1 indicates that the person bought the item a 0 indicates that they did not.

Person ID	Item 107	Item 498	Item 7256	Item 28063	Item 75328
A	1	1	1	0	0
B	1	0	0	1	1

Table 2: A query instance from the same domain as the examples listed in Table 1. A 1 indicates that the person bought the item a 0 indicates that they did not.

Person ID	Item 107	Item 498	Item 7256	Item 28063	Item 75328
Q	1	0	1	0	0

2. (a) You have been given the job of building a recommender system for an large online shop that has a stock of over 100,000 items. In this domain the behaviour of individuals is captured in terms of what items they have bought or not bought. Table 1 lists the behaviour of two individuals in this domain for a subset of the items that at least one of the individuals has bought. Table 2 lists the behaviour of a customer that you want to generate recommendations for. Table 3 lists 3 different models of similarity that work on binary data, similar to the data in this domain (**Russell-Rao**, **Sokal-Michener**, and **Jaccard**).

$$\begin{aligned} \text{Russell-Rao(X,Y)} &= \frac{CP(X,Y)}{P} \\ \text{Sokal-Michener(X,Y)} &= \frac{CP(X,Y)+CA(X,Y)}{P} \\ \text{Jaccard(X,Y)} &= \frac{CP(X,Y)}{CP(X,Y)+PA(X,Y)+AP(X,Y)} \end{aligned}$$

Table 3: Similarity Metrics for Binary Data.

(i) Given that there are over 100,000 items available in the store which of these models of similarity (**Russell-Rao**, **Sokal-Michener**, or **Jaccard**) is most appropriate for this domain. Give an explanation for your choice.

(5 marks)

(ii) Assuming that the recommender system uses the similarity metric you selected in Part (i) and that the system will recommend to person Q the items that the person most similar to person Q has already bought but that person Q has not bought, **which item or items will the system recommend to person Q?** Support you answer by showing your calculations and explaining your analysis of the results.

(10 marks)

(b) Table 4, on the next page lists a data set with of 6 examples described in terms of 3 binary descriptive features (**A**, **B**, and **C**) and a target label (**Target**). You are asked to create a decision tree model using this data. **Which of the descriptive features will the ID3 decision tree induction algorithm choose as the feature for the root node of the decision tree?** Support you anwer with appropriate calculations and dicussions of your results. Note that Table 5, also on the next page, lists some equations that you may find useful for this question.

(15 marks)

ID	A	B	C	Target
1	1	0	1	C1
2	1	1	1	C2
3	1	0	1	C1
4	0	1	1	C2
5	0	1	0	C1
6	0	1	1	C2

Table 4: Dataset for the ID3 Algorithm Question

$$
\begin{aligned}
\text{Entropy(DS)} &= -\sum_{i=1}^{k} p_i \times log_2(p_i) \\
\text{Remainder(F)} &= \sum_{v \in Domain(F)} \frac{|DS_v|}{|DS|} Entropy(DS_v) \\
\text{InformationGain(F,DS)} &= Entropy(DS) - Remainder(F)
\end{aligned}
$$

Table 5: Equations from information theory.

Table 6: Spam and Ham Dataset

Spam	**Ham**
Offer is Free	*Great Learning Fun*
Free Learning Link	*Great Machine Learning*
Click Free Link	*Free Learning Event*
	Learning is Fun
	Learning Costs Money

Table 7: Query Title

Fun is Free

3. Table 6 lists a dataset of email subjects. Table 7 lists the title of a query instance that we would like to classify as being either a spam or ham email based on its title.

   (a) Using **Laplacian smoothing**, where

   $$p(x = v) \quad = \quad \frac{count(x = v) + k}{count(x) + (k \times |Domain(x)|)}$$

   with **k=1** and a **vocabulary size of 12** calculate the following probabilities:

   (i) $P(Spam) =?$

   (2 marks)

   (ii) $P(Ham) =?$

   (2 marks)

   (iii) $P('Fun'|Spam) =?$

   (2 marks)

   (iv) $P('Fun'|Ham) =?$

   (2 marks)

   (v) $P('is'|Spam) =?$

   (2 marks)

   (vi) $P('is'|Ham) =?$

   (2 marks)

   (vii) $P('Free'|Spam) =?$

   (2 marks)

   (viii) $P('Free'|Ham) =?$

   (2 marks)

   (b) Calculate the probability of the query title in Table 7 belonging to the Spam class under the **Naive Bayes assumption** and using the **smoothed probabilities** you calculated in Part (a):

   $$P(Spam|'Fun\ is\ Free') =?$$

   (8 marks)

(c) Calculate the probability of the query title in Table 7 belonging to the Spam class under the **Naive Bayes assumption** and using **maximum likelihood** probabilities (i.e. the probabilities we would get if we did not use Laplacian smoothing):

$$P(Spam|'Fun\ is\ Free') = ?$$

(6 marks)

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 its terms.

(3 marks)

(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.

(3 marks)

(b) Figure 1, on the next pages, 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, carry out the calculations listed below. Note that Table 8, also on the next page, lists some equations that you may find useful when doing this question.

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

(10 marks)

(ii) Calculate the Δ error for each output unit (to 3 decimal places).

(6 marks)

(iii) Calculate the Δ error for each hidden unit B and C. (to 3 decimal places)

(8 marks)

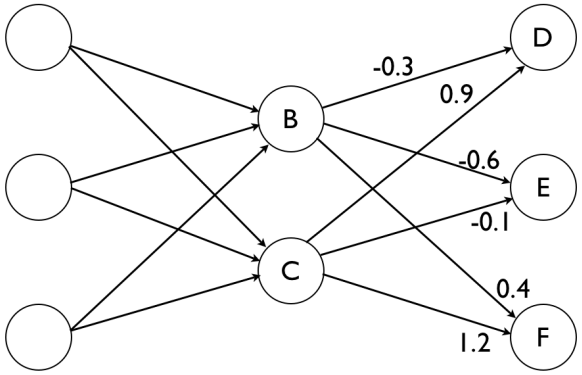

Figure 1: Example Neural Net

Weighted sum of inputs for unit i with j inputs:	$in_i = \sum_j W_{ji} a_j(in_j)$
Activation Function (Logistic) for unit i:	$a_i(in_i) = \frac{1}{1+\exp^{-in_i}}$
Perceptron weight update rule for link $j \rightarrow i$	$w_{ji} = w_{ji} + \eta\,(t_i - a_i(in_i)) \times a_j(in_j)$
Hebbian Weight Update Rule for link $j \rightarrow i$	$w_{ji} = \eta \times a_j(in_j) \times a_i(in_i)$
Partial Derivative for Logistic Activation Function	$\frac{\delta a_i(in_i)}{\delta in_i} = a_i(in_i) \times (1 - a_i(in_i))$
Error for an output unit i	$error_i = target_i - a_i(in_i)$
Delta Error for an output unit i	$\Delta_i = error_i \times a_i(in_i) \times (1 - a_i(in_i))$
Delta Error for a hidden unit j feeding into n units	$\Delta_j = (\sum_{i=1}^{n} W_{ji} \times \Delta_i) \times a_j(in_j) \times (1 - a_j(in_j))$
Delta Weight Update Rule for link $x \rightarrow k$	$W_{x,k} = W_{x,k} + (\eta \times a_x(in_x) \times \Delta_k)$

Table 8: Equations used in Perceptron and Neural Network training.