S228/419C

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc. (Hons) in Computer Science

Stage 4

SEMESTER 2 EXAMINATIONS 2013/2014

ARTIFICIAL INTELLIGENCE II

Dr. John Kelleher
Dr. Deirdre. Lillis
Mr. P. Collins

Monday
12^{th} May 2014
4:00 p.m to 6:00 p.m

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

(15 marks)

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

(15 marks)

(d) Explain what can go wrong when a machine learning classifier uses the wrong inductive bias.

(5 marks)

2. (a) A data analyst building a k-nearest neighbour model for a continuous prediction problem is considering appropriate values to use for k.

   (i) Initially the analyst uses a simple average of the target variables for the k nearest neighbours in order to make a new prediction. After experimenting with values for k in the range $0 - 10$ it occurs to the analyst that they might get very good results if they set k to the total number of instances in the training set. Do you think the analyst is likely to get good results using this value for k?

   (5 marks)

   (ii) If the analyst was using a distance weighted averaging function rather than a simple average for their predictions would this have made their idea any more useful?

   (5 marks)

   (b) A dataset showing the decisions made by an individual about whether to wait for a table at a restaurant is listed in Table 1 on the next page. (Note that Table 2, also on the next page, lists some equations that you may find useful for this question.)

   (i) Given that the $WillWait$ column lists the values of the target variable, compute the entropy for this dataset.

   (5 marks)

   (ii) What is the information gain for the $Patrons$ feature?

   (5 marks)

   (iii) What is the information gain for the $Type$ feature?

   (5 marks)

   (iv) Given a choice between the $Patrons$ and $Type$ feature, which feature would the ID3 algorithm choose as the root node for a decision tree?

   (5 marks)

ID	Bar	$Patrons$	$Price$	$Rain$	$Type$	WillWait
1	F	Some	$$$	F	French	T
2	F	Full	$	F	Thai	F
3	T	Some	$	F	Burger	T
4	F	Full	$	F	Thai	T
5	F	Full	$$$	F	French	F
6	T	Some	$$	T	Italian	T
7	T	None	$	T	Burger	F
8	F	Some	$$	T	Thai	T
9	T	Full	$	T	Burger	F
10	T	Full	$$$	F	Italian	F
11	F	None	$	F	Thai	F
12	T	Full	$	F	Burger	T

Table 1: A dataset describing the previous decisions made by an individual about whether to wait for a table at a restaurant.

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

Table 2: Equations from information theory.

3. Table 3 (on the next page) lists a dataset of the subject lines from emails. Table 4 (also on the next page) shows the subject line for an email that we would like to classify as Spam or Ham.

   (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 4 belonging to the Spam class under the **Naive Bayes assumption** and using the **smoothed probabilities** you calculated in Part (a):

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

   (7 marks)

   (c) Calculate the probability of the query title in Table 4 belonging to the Spam class under the **Naive Bayes assumption** and using **maximum likelihood** probabilities (i.e. the probabilities we could get if we did not use Laplacian smoothing):

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

   (7 marks)

Table 3: Spam and Ham Dataset

Spam
Offer is Free
Free Learning Link
Cick Free Link

Ham
Great Learning Fun
Great Machine Learning
Free Learning Event
Learning is Fun
Learning Costs Money

Table 4: Query Title

Fun is Free

x	0	1	2	3	4
y	3	6	7	8	11

Table 5: Example Dataset for Linear Regression Question

4. (a) Assuming a domain with one descriptive feature x and one target feature y, linear regression uses the following formula to model the relationship between the descriptive and target features:

$$f(x) = w_1 x + w0$$

where $w1$ and $w0$ are computed using the following formulae, where M is number of data points in the dataset:

$$w_1 = \frac{(M \sum_{i=1}^{M} x_i y_i) - (\sum_{i=1}^{M} x_i \sum_{i=1}^{M} y_i)}{(M \sum_{i=1}^{M} x_i^2) - (\sum_{i=1}^{M} x_i)^2}$$

$$w_0 = (\frac{1}{M} \sum_{i=1}^{M} y_i) - (\frac{w_1}{M} \sum_{i=1}^{M} x_i)$$

Using the data in Table 5 compute the values of w_0 and w_1 that provide the best linear fit to the data.

(10 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 6 (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 new weight W_{BD} for the connection from unit B to the output unit D after the training example has been processed. Use a learning rate of $\eta = 0.25$.

(4 marks)

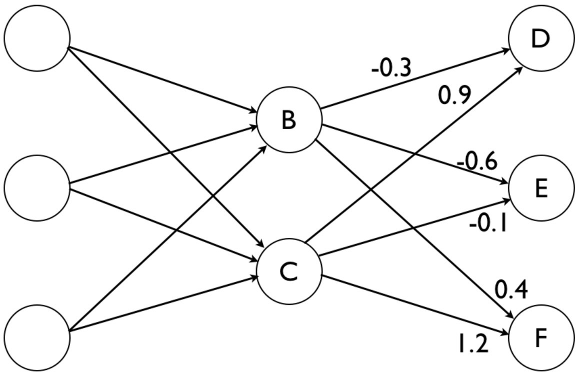

Figure 1: Example Neural Net

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	$\Delta_j = \left(\sum_{i=1}^{n} W_{ji} \times \Delta_i \right) \times a_j(in_j) \times (1 - a_j(in_j))$
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 6: Equations used in Perceptron and Neural Network training.