R228/401C

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc. (Hons) in Computer Science

Stage 4

SUPPLEMENTAL EXAMINATIONS 2012

*** *SOLUTIONS* ***

ARTIFICIAL INTELLIGENCE II

Dr. John Kelleher
Dr. Deirdre. Lillis
Mr. D. Tracey

Duration: 2 Hours

Question 1 is **compulsory**

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) Explain what can go wrong when a machine learning classification algorithm uses the wrong inductive bias.

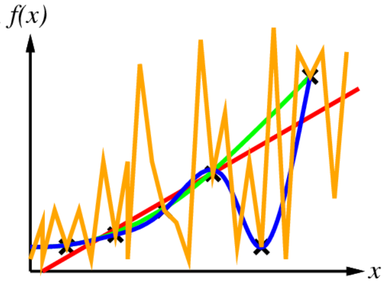
(5 marks)

- If the inductive bias of the learning algorithm constrains the search to only consider simple hypotheses we may have excluded the real function from the hypothesis space. In other words, the true function is **unrealizable** in the chosen hypothesis space, (i.e., we are **underfitting**).
- If the inductive bias of the learning algorithm allows the search to consider complex hypotheses, the model may hone in on irrelevant factors in the training set. In other words the model with **overfit** the training data.

(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.

2. (a) Discuss the advantages and disadvantages of k-**Nearest Neighbour** classification.

(5 marks)

Strengths

(i) No training involved lazy learning

(ii) New data can be added on the fly

(iii) Some explanation capabilities

(iv) Robust to noisy data by averaging k-nearest neighbors

Weaknesses

(i) Not the most powerful classification (generally its accuracy will be lower than an ANN or SVM model)

(ii) Slow classification

(iii) Curse of dimensionality (as you increase the number of features you need more and more examples to cover the problem space - kNN are particularly susceptible to this issue as they do not do any feature selection).

(b) IT-Tunes is on online music sales service that is trying to build an inductive machine learning system to recommend new songs to its users. They have collected a dataset that records details of songs they sell and details of ratings that users have given these songs (missing values in the ratings columns indicate that a user has not rated a particular song). The table below shows an extract from this dataset showing details of 5 songs and associated ratings for 5 users (note that in the full dataset there are many more songs and many more users).

ID	001	004	007	011	022
Title	Jeremy	One	Please	Fool's Game	Help!
Album	Ten	Achtung Baby	Pop	M. Bolton	Help!
Artist	Pearl Jam	U2	U2	M. Bolton	The Beatles
Year	1990	1992	1997	1983	1965
Genre	Rock	Rock	Rock	Soul	Pop
Best Chart Pos	10	1	6	25	1
User1-Score	3	2	2	5	4
User2-Score	5	-	5	-	4
User3-Score	5	2	2	1	1
User4-Score	1	-	-	-	1
User5-Score	-	1	1	5	3

(i) In order to build a case based reasoning (CBR) classification system to predict whether a user would like a particular song, or not, a case representation is required. Describe a case structure for the dataset shown in the table above that could be used to build a CBR system that would predict the rating a particular user would be likely to give to a particular song. Justify any decisions that you make and explain any assumptions used.

(5 marks)

Any reasonable system will be acceptable here as long as it is justified. Issues that should be addressed include:

A. Which features actually hold information that is likely to relate to the problem?

B. Can a balance be struck between internal features (such as song titles, years etc) and external features (such as user ratings)

C. Can derived features by created from the features present in the dataset?

(ii) A good similarity measure is crucial for any CBR system. Describe a similarity measure that would be appropriate for comparing cases in the structure described in the answer part (i) of this question. Justify any decisions that you make and explain any assumptions used.

(5 marks)

Again any reasonable answer will be accepted here. Issues that should be addressed include:

A. Normalisation

B. How can various elements within a hybrid similarity measure be combined in order to make a complete similarity measure?

C. How can the various elements present in the feature set be compared?

(iii) Is **case-based reasoning** the most appropriate inductive machine learning technique to use for this prediction problem?

(5 marks)

Again students have a fairly free hand in answering this question. Issues that should be addressed include:

- The ability to add new instances when new information becomes available.
- the need to build classification models for each customer
- Is a classification model the best approach, or would something like collaborative filtering be more appropriate?

(c) 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

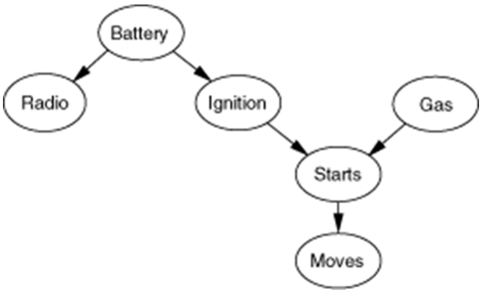


Figure 1: A Bayesian network of a car's electrical system and engine. Each variable is boolean and the true value indicates that the corresponding aspect of the vehicle is in working order.

Table 1: Full joint probability 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

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

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

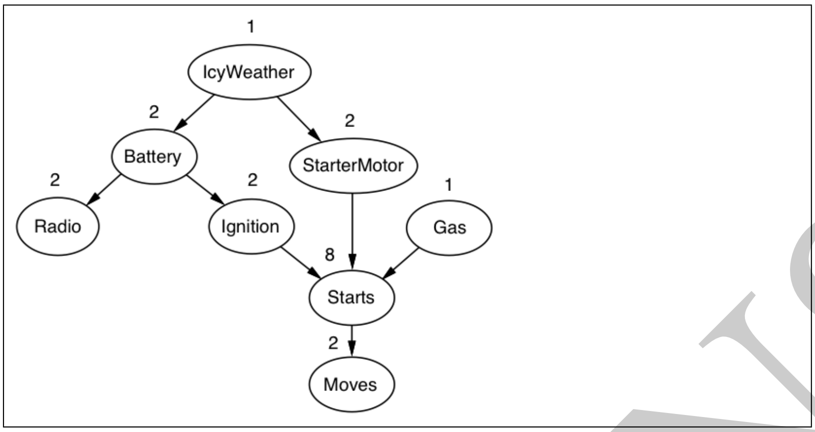
(5 marks)

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) Consider the network for car diagnosis shown in Figure 1.

(i) Extend the network with the Boolean variables *IcyWeather* and *StarterMotor*; assume that a *StarterMotor* effects whether or not the car *Starts* and that *Icy Weather* effects the *Battery* and the *StarterMotor*,

(5 marks)



(ii) How many independent values are contained in the full joint probability distribution?

(5 marks)

With 8 Boolean variables, the joint has $2^8?1 = 255$ independent entries.

(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)

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

Table 2: Example Dataset for Linear Regression Question

4. (a) Assuming a domain with one explanatory variable x and one dependent variable y linear regression uses the following formula to model the relationship between the explanatory and dependent variable:

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

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 2 compute the values of w_0 and w_1 that provide the best linear fit to the data.

(10 marks)

First we need to compute the values of the equation components:

- M = 5
- $\sum_{i=1}^{M} x_i y_i = 0 + 6 + 14 + 24 + 44 = 88$
- $\sum_{i=1}^{M} x_i = 10$
- $\sum_{i=1}^{M} y_i = 35$
- $\sum_{i=1}^{M} x_i^2 = 0 + 1 + 4 + 9 + 16 = 30$
- $(\sum_{i=1}^{M} x_i)^2 = 10^2 = 100$

Given these values, w_1:

$$w_1 = \frac{(5*88) - (10*35)}{(5*30) - 100} = \frac{90}{50} = 1.8$$

And, w_0:

$$w_0 = (\frac{1}{5} * 35) - (\frac{1.8}{5} * 10) = 7 - 3.6 = 3.4$$

(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 $x \in C_j, j \neq i$

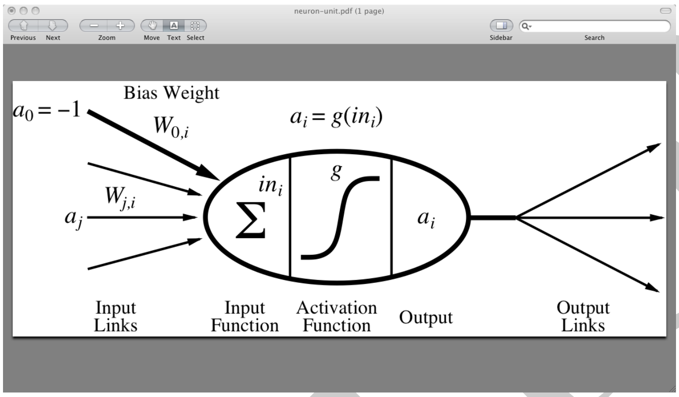
(c) Describe the processing stages of a McCulloch-Pits ”unit”.

(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 2 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.

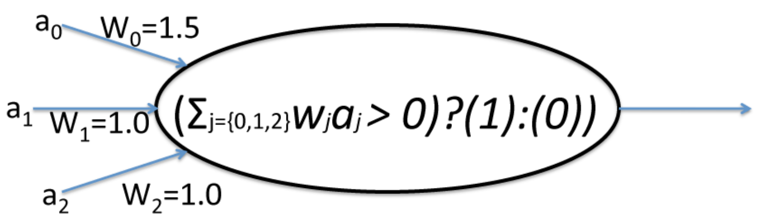


Figure 2: A 3 input perceptron. Input $a_0 = -1$, inputs a_1 and a_2 are binary. The perceptron uses a threshold activation function that outputs a 1 if the weighted sum of inputs is greater than 0 and a 0 otherwise.