S249/419C

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc. (Honours) Degree in Information Systems / Information Technology

Stage 4

SUMMER EXAMINATIONS 2015

*** *SOLUTIONS* ***

ARTIFICIAL INTELLIGENCE II [CMPU4011]

Dr. John Kelleher
Dr. Deirdre. Lillis
Dr. Rem Collier

Monday 11^{th} May 2015
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).

*** 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 classifier 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) Table 1 shows the predictions made for a categorical target feature by a model for a test dataset.

(i) Create the **confusion matrix** for the results listed in Table 1.

(5 marks)

		Prediction	
		true	*false*
Target	*true*	1	3
	false	2	14

(ii) Calculate the **classification accuracy** for the results listed in Table 1.

$$classification\ accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

(5 marks)

Classification accuracy can be calculated as

$$\begin{aligned} classification\ rate &= \frac{(TP + TN)}{(TP + TN + FP + FN)} \\ &= \frac{(1 + 14)}{(1 + 14 + 3 + 2)} \\ &= 0.75 \end{aligned}$$

(iii) Calculate the **average class accuracy (harmonic mean)** for the results listed in Table 1. (During this calculation you should round all long floats to 4

places of decimal.)

$$average\ class\ accuracy_{HM} = \frac{1}{\frac{1}{|levels(t)|}\sum_{l \in levels(t)} \frac{1}{recall_l}}$$

(8 marks)

Note, in this solution we round all figures to four places of decimal. First, we calculate the recall for each target level:

$$recall_{\textit{true}} = \frac{1}{4} = 0.25$$
$$recall_{\textit{false}} = \frac{14}{16} = 0.875$$

Then we can calculate a harmonic mean as

$$\begin{aligned} average\ class\ accuracy_{HM} &= \frac{1}{\frac{1}{|levels(t)|}\sum_{l \in levels(t)} \frac{1}{recall_l}} \\ &= \frac{1}{\frac{1}{2}\left(\frac{1}{0.25} + \frac{1}{0.875}\right)} \\ &= \frac{1}{\frac{1}{2}(4 + 1.1429)} \\ &= 0.38889 \end{aligned}$$

(iv) Which of these performance metrics (**misclassification rate** or **average class accuracy (harmonic mean)** is the most appropriate metric to use in this scenario? Provide an explanation for your answer.

(12 marks)

This is a discursive question so providing a precise answer is not appropriate, but in general the students should:

- Note that the test dataset is **imbalanced**: there are only 4 instances of the *true* class in the test dataset.
- Point out that classification accuracy does not take into account the distribution of classes in the test set, and as a result it can hide poor performance of a model on one of the class. In this particular instance, the classification accuracy is dominated by the performance of the model on the false class and as a result hides the fact that the model only gets 1 out of the four true instances correct.
- Average class accuracy (harmonic mean) does take the distribution of class into account and furthermore by using the harmonic mean also pays more attention to the classes that the model has low recall on. As a result average class accuracy is the more appropriate performance metric to use in this instance.

Table 1: The predictions made by a model for a categorical target on a test set of 20 instances

ID	Target	Prediction	ID	Target	Prediction
1	false	false	11	false	false
2	false	false	12	false	true
3	false	false	13	false	false
4	false	false	14	false	false
5	false	true	15	false	false
6	false	false	16	false	false
7	false	false	17	true	false
8	false	false	18	true	false
9	false	false	19	true	false
10	false	false	20	true	true

2. (a) Table 3, on the next page, lists a dataset containing examples described by two descriptive features, **Feature 1** and **Feature 2**, and labelled with a target class **Target**. Table 4, also on the next page, lists the details of a query for which we want to predict the target label. We have decided to use a **3-Nearest Neighbor** model for this prediction and we will use Euclidean distance as our distance metric:

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{n} \left((x_1.f_i - x_2.f_i)^2 \right)}$$

(i) With which target class (**C1** or **C2**) will our **3-Nearest Neighbor** model label the query? Provide an explanation for your answer.

(8 marks)

The first stage is to calculate the Euclidean distance between each of the examples and the query:

ID	Euclidean Distance
101	60000
102	120000
103	120000
104	180000
105	240000

From this table we can see that the three closest examples to the query are examples 101, 102, and 103. Example 101 has a target label of C1 and both 102 and 103 have target labels C2. Consequently C2 is the majority label in local model constructed by the 3-Nearest Neighbor classifier for this query instance and the query will be labelled with class C2.

(ii) There is a large variation in range between **Feature 1** and **Feature 2**. To account for this we decide to normalize the data. Compute the normalized versions of Feature 1 and Feature 2 to four decimal places of precision using range normalization

$$x_i.f' = \frac{x_i.f - min(f)}{max(f) - min(f)}$$

(4 marks)

ID	Feature 1	Feature 2	Target
101	0.2	0.1667	C1
102	0	0.0000	C2
103	0.8	0.6667	C2
104	0.4	0.8333	C1
105	1	1.0000	C2

(iii) Assuming we use the normalized dataset as input, with which target class (**C1** or **C2**) will our **3-Nearest Neighbor** model label the query? Provide an explanation for your answer.

(8 marks)

The normalize query instance is:

ID	Feature 1	Feature 2	Target
250	0.2	0.3333	?

The Euclidean distances between the normalized data and normalized query are:

ID	Euclidean Distance
101	0.1667
102	0.3887
103	0.6864
104	0.5385
105	1.0414

From this table we can see that the 3 closest neighbors are: 101, 103 and 104. 101 and 104 are both labelled as class **C1**. So **C1** is the majority class in the neighborhood and the query will be labelled as belonging to it.

(b) Table 5, on the next page, lists a classification dataset. Each instance in the dataset has two descriptive features (Feature A and Feature B) and is classified as either a positive (+) or a negative(-) example. Note that Table 2, below, lists some equations that you may find useful for this question.

Table 2: Equations from information theory.

$$H(\mathbf{f}, \mathcal{D}) = -\sum_{l \in levels(f)} P(f = l) \times log_2(P(f = l))$$

$$rem(\mathbf{f}, \mathcal{D}) = \sum_{l \in levels(f)} \frac{|\mathcal{D}_{f=l}|}{|\mathcal{D}|} \times H(t, \mathcal{D})$$

$$IG(\mathbf{d}, \mathcal{D}) = H(\mathbf{t}, \mathcal{D}) - rem(\mathbf{d}, \mathcal{D})$$

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

(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 Feature A and Feature B.

(5 marks)

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

Table 3: Dataset for the 3-Nearest Neighbor question

ID	Feature 1	Feature 2	Target
101	4	180000	C1
102	3	120000	C2
103	7	360000	C2
104	5	420000	C1
105	8	480000	C2

Table 4: Query instance for the 3-Nearest Neighbor question.

ID	Feature 1	Feature 2	Target
250	4	240000	?

Table 5: Classification dataset for information question.

Feature A	Feature B	Classification
True	True	+
True	False	-
True	False	+
True	True	+
False	True	-

3. Table 6 lists a dataset of the previous decision made by a couple regarding whether or not they would wait for a table at a restaurant (i.e., the feature WAITED is the target feature in this domain).

   (a) Calculate the probabilities (to four places of decimal) that a **naive Bayes** classifier would use to represent this domain.

   (18 marks)

   A naive Bayes classifier would require the prior probability for each level of the target feature and the conditional probability for each level of each descriptive feature given each level of the target feature:

   $P(Waited = Yes) = 0.4$ $P(Waited = No) = 0.6$
   $P(Bar = True|Waited = Yes) = 0.5$ $P(Bar = True|Waited = No) = 0.5$
   $P(Bar = False|Waited = Yes) = 0.5$ $P(Bar = False|Waited = No) = 0.5$
   $P(Patrons = None \mid Waited = Yes) = 0.25$ $P(Patrons = None \mid Waited = No) = 0.1667$
   $P(Patrons = Some \mid Waited = Yes) = 0.5$ $P(Patrons = Some \mid Waited = No) = 0.3333$
   $P(Patrons = Full \mid Waited = Yes) = 0.25$ $P(Patrons = Full \mid Waited = No) = 0.5$
   $P(Price = Cheap \mid Waited = Yes) = 0.5$ $P(Price = Cheap \mid Waited = No) = 0.5$
   $P(Price = Reasonable \mid Waited = Yes) = 0.25$ $P(Price = Reasonable \mid Waited = No) = 0.3333$
   $P(Price = Expensive \mid Waited = Yes) = 0.25$ $P(Price = Expensive \mid Waited = No) = 0.1667$

   (b) Assuming conditional independence between features given the target feature value, calculate the **probability** of each outcome (WAITED=Yes, and WAITED=No) for the following restaurant for this couple (marks will be deducted if workings are not shown, round your results to four places of decimal)

   BAR=False, PATRONS=None, PRICE=Expensive

   (10 marks)

   The initial score for each outcome is calculated as follows:
   $(Waited = Yes) = 0.5 \times 0.25 \times 0.25 \times 0.4 = 0.0125$
   $(Waited = No) = 0.5 \times 0.1667 \times 0.1667 \times 0.6 = 0.0083$
   However, these scores are not probabilities. To get real probabilities we must normalise these scores. The normalisation constant is calculated as follows:
   $\alpha = 0.0125 + 0.0083 = 0.0208$
   The actual probabilities of each outcome is then calculated as:
   $P(Waited = Yes) = \frac{0.0125}{0.0208} = (0.600961...) = 0.6010$
   $P(Waited = No) = \frac{0.0083}{0.0208} = (0.399038...) = 0.3990$

   (c) What prediction would a **naive Bayes** classifier return for the above restaurant?

   (2 marks)

   A naive Bayes classifier returns outcome with the maximum a posteriori probability as its prediction. In this instance the outcome WAITED=Yes is the MAP prediction and will be the outcome returned by a naive Bayes model.

Table 6: A dataset describing the previous decisions made by an individual about whether to wait for a table at a restaurant.

ID	BAR	PATRONS	PRICE	WAITED
1	False	Some	Expensive	Yes
2	False	Full	Cheap	No
3	True	Some	Cheap	Yes
4	False	Full	Cheap	Yes
5	False	Full	Expensive	No
6	True	Some	Reasonable	No
7	True	None	Cheap	No
8	False	Some	Reasonable	No
9	True	Full	Cheap	No
10	True	None	Reasonable	Yes

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

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

(i) Provide the name for this model and explain all of the terms that it contains.

(4 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.

(8 marks)

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

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) A multivariate logistic regression model has been built to predict the propensity of shoppers to perform a repeat purchase of a free gift that they are given. The descriptive features used by the model are the age of the customer, the average amount of money the customer spends on each visit to the shop, and the average number of visits the customer makes to the shop per week. This model is being used by the marketing department to determine who should be given the free gift. The trained model is

$$\begin{aligned}\text{REPEAT PURCHASE} = &-3.82398 - 0.02990 \times \text{AGE} \\ &+ 0.74572 \times \text{SHOP FREQUENCY} \\ &+ 0.02999 \times \text{SHOP VALUE}\end{aligned}$$

And, the logistic function is defined as:

$$logistic(x) = \frac{1}{1 + e^{-x}}$$

Assuming that the *yes* level is the positive level and the classification threshold is 0.5, use this model to make predictions for each of the query instances shown in Table 7 below.

(18 marks)

Calculating the predictions made by the model simply involves inserting the descriptive features from each query instance into the prediction model. With this information, the predictions can be made as follows:

1: $Logistic(-3.82398 + -0.0299 \times 56 + 0.74572 \times 1.6 + 0.02999 \times 109.32)$
$= Logistic(-1.02672) = \frac{1}{1-e^{1.02672}}$
$= 0.26372 \Rightarrow$ *no*

2: $Logistic(-3.82398 + -0.0299 \times 21 + 0.74572 \times 4.92 + 0.02999 \times 11.28)$
$= Logistic(-0.44465) = \frac{1}{1-e^{0.44465}}$
$= 0.390633 \Rightarrow$ *no*

3: $Logistic(-3.82398 + -0.0299 \times 48 + 0.74572 \times 1.21 + 0.02999 \times 161.19)$
$= Logistic(0.477229) = \frac{1}{1-e^{-0.477229}}$
$= 0.6205 \Rightarrow$ *yes*

Table 7: The queries for the multivariate logistic regression question

ID	AGE	SHOP FREQUENCY	SHOP VALUE
1	56	1.60	109.32
2	21	4.92	11.28
3	48	1.21	161.19