R228/419C & R211C/419C

DUBLIN INSTITUTE OF TECHNOLOGY
KEVIN STREET, DUBLIN 8

BSc. (Honours)
Degree in Computer Science

Year 4

SUPPLEMENTAL EXAMINATIONS 2016

ARTIFICIAL INTELLIGENCE II [CMPU4011]

Dr. John Kelleher
Dr. Deirdre. Lillis
Mr. P. Collins (DT228 External)
Mr. T. Nolan (DT211 External)

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) 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)

(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)

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

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

(5 marks)

(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.)

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

(10 marks)

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) You are 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.

   (i) Table 2 (below) lists 3 different models of similarity that work on binary data, similar to the data in this domain (**Russell-Rao**, **Sokal-Michener**, and **Jaccard**). 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) Table 3 (on the next page) lists the behaviour of two individuals in this domain for a subset of the items that at least one of the individuals has bought; and, Table 4 (also, on the next page) lists the behaviour of a customer **Q** that you want to generate recommendations for. 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)

Table 2: Similarity Metrics for Binary Data.

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

   (b) Table 5 on the next page lists a sample of data from a census. There are four descriptive features in this dataset (AGE, EDUCATION, MARITAL STATUS, OCCUPATION) and the target feature ANNUAL INCOME has 3 levels ($<25K$, $25K$–$50K$, $>50K$). Note, Table 6, also on the next page, lists some equations that you may find useful for this question.

   (i) Calculate the ENTROPY for this dataset.

   (5 marks)

   (ii) When building a decision tree, the easiest way to handle a continuous feature is to define a threshold around which splits will be made. What would be the optimal threshold to split the continuous AGE feature (use information gain based on entropy as the feature selection measure)?

   (10 marks)

Table 3: 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 4: A query instance from the same domain as the examples listed in Table 3. 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

Table 5: Census data for the ID3 Algorithm Question

ID	AGE	EDUCATION	MARITAL STATUS	OCCUPATION	ANNUAL INCOME
1	39	bachelors	never married	transport	$25K$–$50K$
2	50	bachelors	married	professional	$25K$–$50K$
3	18	high school	never married	agriculture	$<25K$
4	28	bachelors	married	professional	$25K$–$50K$
5	37	high school	married	agriculture	$25K$–$50K$
6	24	high school	never married	armed forces	$<25K$
7	52	high school	divorced	transport	$25K$–$50K$
8	40	doctorate	married	professional	$>50K$

Table 6: Equations from information theory.

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

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

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

3. Table 7 lists a dataset of books and whether or not they were purchased by an individual (i.e., the feature PURCHASED 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)

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

   2ND HAND=False, GENRE=Literature, COST=Expensive

   (10 marks)

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

   (2 marks)

Table 7: A dataset describing the a set of books and whether or not they were purchased by an individual.

ID	2ND HAND	GENRE	COST	PURCHASED
1	False	Romance	Expensive	Yes
3	True	Romance	Cheap	Yes
4	False	Science	Cheap	Yes
10	True	Literature	Reasonable	Yes
2	False	Science	Cheap	No
5	False	Science	Expensive	No
6	True	Romance	Reasonable	No
7	True	Literature	Cheap	No
8	False	Romance	Reasonable	No
9	True	Science	Cheap	No

4. (a) A multivariate linear regression model has been built to predict the HEATING LOAD in a residential building based on a set of descriptive features describing the characteristics of the building. Heating load is the amount of heat energy required to keep a building at a specified temperature, usually 65° Fahrenheit, during the winter regardless of outside temperature. The descriptive features used are the overall surface area of the building, the height of the building, the area of the building's roof, and the percentage of wall area in the building that is glazed. This kind of model would be useful to architects or engineers when designing a new building. The trained model is

$$\begin{aligned}\text{HEATING LOAD} = &-26.030 + 0.0497 \times \text{SURFACE AREA} \\ &+ 4.942 \times \text{HEIGHT} - 0.090 \times \text{ROOF AREA} \\ &+ 20.523 \times \text{GLAZING AREA}\end{aligned}$$

Use this model to make predictions for each of the query instances shown in the Table 8 on the next page.

(12 marks)

(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 9 on the next page.

(18 marks)

Table 8: The queries for the multivariate linear regression HEATING LOAD question

ID	SURFACE AREA	HEIGHT	ROOF AREA	GLAZING AREA
1	784.0	3.5	220.5	0.25
2	710.5	3.0	210.5	0.10

Table 9: 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