DUBLIN INSTITUTE OF TECHNOLOGY

DT228A/1 MSc. in Computing
DT228B/1 MSc. in Computing
DT228B/2 MSc. in Computing

SUMMER EXAMINATIONS 2017/2018

MACHINE LEARNING [SPEC9270]

DR. JOHN MCAULEY
PROFESSOR SARAH JANE DELANY
DR. DEIRDRE LILLIS
DR. GEORGIANA IFRIM

WEDNESDAY 16TH MAY 2.00 P.M. – 4.00 P.M.

TWO HOURS

ANSWER 2 QUESTIONS.
ALL QUESTIONS CARRY EQUAL MARKS.

1.	(a)	**Please Answer the following Questions. Each question carries equal Marks.** **[24 marks]**	
		(i)	What is a Cost Function? Give an example of a cost function. **[4 marks]**
		(ii)	In the following example, which descriptive feature carries the least information for the purpose of classification? Please explain why.

ID	Name	Age	Height
1	John	10	6.2
2	Jo	20	5.2
3	Paula	30	5.8

[4 marks]

		(iii)	Describe the difference in inductive bias for similarity-based learning using KNN and information-based learning using ID3 **[4 marks]**
		(iv)	I have the following data, please suggest a similarity measure for similarity based learning and justify your decision.

ID	Weight	Height
1	362	4.2
2	268	5.2
3	185	5.8

[4 marks]

		(v)	What is the difference between the Jaccard index and Sokal-Michener measure? **[4 marks]**
		(vi)	Given that KNN is a lazy learner, what problem can this present? And how could you address this problem? **[4 marks]**
	(b)	The following is a Data Quality Report from a Consultant when working on an ABT.	

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
AGE	5,200	0	51	18	22	41.59	47	50	80	15.66
MOTORVALUE	5,200	17.25	3,934	4,352	15,089.5	23,479	24,853	32,078	166,993	11,121
HEALTHDEPSADULTS	5,200	39.25	4	0	0	0.84	1	1	2	0.65
HEALTHDEPSKIDS	5,200	39.25	5	0	0	1.77	2	3	3	1.11

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
GENDER	5,200	0	2	female	2,626	50.5	male	2,574	49.5
LOC	5,200	0	2	urban	2,948	56.69	rural	2,252	43.30
OCC	5,200	37.71	1,828	Nurse	11	0.34	Sales	9	0.28
MOTORINS	5,200	0	2	yes	4,303	82.75	no	897	17.25
HEALTHINS	5,200	0	2	yes	3,159	60.75	no	2,041	39.25
HEALTHTYPE	5,200	39.25	4	PlanB	1,596	50.52	PlanA	796	25.20
PREFCHANNEL	5,200	0	3	email	2,296	44.15	phone	1,975	37.98

[8 Marks]

(i) **What problems exist in the data?**

[4 Marks]

(ii) **How would you address these problems?**

[2 Marks]

(iii) **How can a missing indicator feature help in this situation?**

[2 Marks]

(c) You have the following set of points representing two classes:

Class 1 - (1, 1), (1, 2), (2, 1) and Class 2 - (3, 1), (3, 3), (3, 4)

Using KNN, and the Euclidean Distance Measure, please classify the following query instance (3, 2) with a neighbourhood of 1 (k=1).

Please show your workings, each calculation for each point and your

[8 Marks]

(d) After your yearly check-up, the doctor has bad news and good news for you. The bad news is that you tested positive for a serious disease on a test that has been shown to be 99% accurate (i.e., the probability of testing positive when you have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age.

[10 Marks]

		(i)	Why is it particularly good news that the disease is rare? **[5 Marks]**
		(ii)	What is the **probability** that you actually have the disease? **[5 Marks]**

2.	(a)	(i)	A challenge when working with textual data is the decision on the representation to use. Describe a suitable representation for a collection of tweets, some examples of which are included below. Justify your choice of feature and feature values. We've no bread but plenty of ice cream #sneachta I'm snowed in with NO BREAD. Coldness level: teenager actually wearing coat to school. #sneachta 😂😂😂 Ireland is bedding down for the next few days #BeastFromTheEast **[5 marks]**
		(ii)	Show the actual feature vector for the first tweet above indicating any assumptions made. **[3 marks]**
		(iii)	Discuss two challenges with dealing with textual data and how to deal with them. **[4 marks]**
		(iv)	Explain the difference between recall, precision and accuracy using the confusion matrix below. In your answer give an example of an appropriate scenario of use on text data for each of the measures, explaining why it is appropriate: **[13 marks]**
	(b)		You have been hired by the European Space Agency to build a model that predicts the amount of oxygen that an astronaut consumes when performing five minutes of intense physical work. The descriptive features for the model will be the age of the astronaut and their average heart rate through- out the work. The regression model is:

Confusion matrix for 2(a)(iv):

	Predicted Class 1	Predicted Class 2
Actual Class 1	720	80
Actual Class 2	180	20

$$\text{OxyCon} = \mathbf{w}[0] + \mathbf{w}[1] \times \text{Age} + \mathbf{w}[2] \times \text{HeartRate}$$

The table below shows a historical dataset that has been collected for this task.

ID	OxyCon	Age	Heart Rate	ID	OxyCon	Age	Heart Rate
1	37.99	41	138	7	44.72	43	158
2	47.34	42	153	8	36.42	46	143
3	44.38	37	151	9	31.21	37	138
4	28.17	46	133	10	54.85	38	158
5	27.07	48	126	11	39.84	43	143
6	37.85	44	145	12	30.83	43	138

[20 marks]

(i) Assuming that the current weights in a multivariate linear regression model are $\mathbf{w}[0] = -59.50$, $\mathbf{w}[1] = -0.15$, and $\mathbf{w}[2] = 0.60$, make a prediction for each training instance using this model.

[6 marks]

(ii) Calculate the sum of squared errors for the set of predictions generated in part i)

[6 marks]

(iii) Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm.

[3 marks]

(iv) Gradient descent is a popular technique used in Machine Learning, explain how gradient descent can be used in Multi Variable Linear Regression.

[5 marks]

(c) What is the difference between supervised and unsupervised learning? Please give an example of each.

[2 marks]

(d) Why would you use the following techniques?

1. K-fold cross validation
2. Leave one out validation
3. Out of time validation?

[3 marks]

3	(a)	Explain the difference between the following machine learning approaches, **information-based learning, error-based learning, probability-based learning** and **similarity-based learning**. Provide an example of each approach, using either restriction bias or preference bias to assist your answer, and describe a potential advantages/disadvantages of each. Answers should be short. **[8 marks]**

(b) You have been given the job of building a recommender system for a large online retail store that has a stock of over 100,000 items. In this domain the behaviour of individuals is captured in terms of the items that they have bought or not bought.

The table below captures the behaviour of two individuals, A and B, in this domain for a subset of the items on sale. The data in this table is binary and a cell marked with a 1 indicates that the person has bought an item, while a 0 indicates that they have not.

ID	Item 107	Item 498	Item 5645	Item 7256	Item 1762	Item 28063	Item 75328
A	1	0	1	0	0	1	0
B	1	0	0	1	0	0	1

The second table below captures the behaviour of a customer, Q, for whom we would like to generate recommendations. Again, the data is binary and a cell marked with a 1 indicates that the person has bought an item, while a 0 indicates that they have not.

ID	Item 107	Item 498	Item 5645	Item 7256	Item 1762	Item 28063	Item 75328
Q	1	0	1	1	0	0	0

[14 marks]

(i) Given that there are over 100,000 items available in the store which of following models of similarity is most appropriate for this domain?

- Russell-Rao
- Sokal-Michener
- Jaccard

Give an explanation for your choice.

[6 marks]

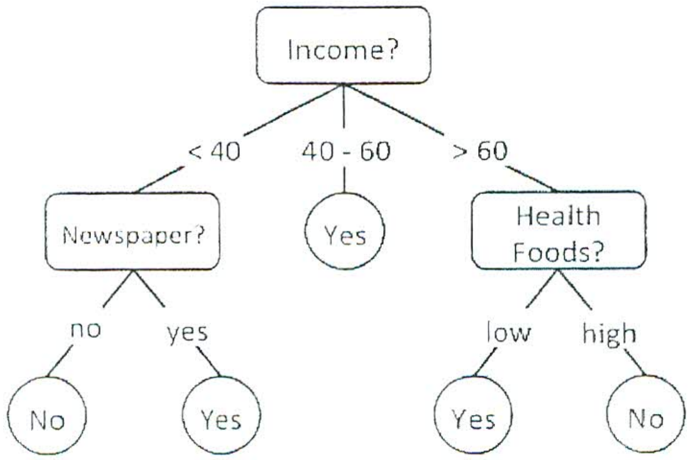
(ii) Assuming that the system will recommend to person **Q** the items that the person most similar to person **Q** has already purchased but that person **Q** has not bought, which item or items will the system recommend to person **Q**? Support you answer by showing your calculations and explaining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

			Part (i).
			[8 marks]

(b) The following table lists a dataset collected by a retail company capturing historical details of which of their customers have responded to promotions the company has run. The information captured covers customer income bracket, customer age, whether or not the customer regularly buys a newspaper, the proportion of health foods typically included in the customer's shopping, and, finally, whether or not they responded to previous promotional mailings.

ID	Income	Age	Newspaper	Health Foods	Respond
C-01	<40	81	no	low	No
C-02	<40	76	no	high	No
C-03	40-60	86	no	low	Yes
C-04	>60	84	no	low	Yes
C-05	>60	45	yes	low	Yes
C-06	>60	66	yes	high	No
C-07	40-60	41	yes	high	Yes
C-08	<40	68	no	low	No
C-09	<40	32	yes	high	Yes
C-10	>60	56	yes	low	Yes
C-11	<40	58	yes	high	Yes
C-12	40-60	52	no	high	Yes
C-13	40-60	90	yes	low	Yes
C-14	>60	69	no	high	No

This dataset has been used to induce a **decision tree** that can predict whether or not new customers will respond to promotional mailings. This decision tree is shown below.



[24 marks]

		(i)	The **information gain** of the feature *Income* at the root node of the tree is 0.247. A colleague has suggested that *Newspaper* would be the best feature to query at the root node of the tree. Demonstrate whether or not this is the case. Please show all workings. **[12 marks]**
		(ii)	Another colleague has suggested that *Age* would be the best feature to query at the root node of the tree. Demonstrate whether or not this is the case. Please show all workings. **[12 marks]**
	~~(c)~~		*In relation to the application of machine learning, describe what you understand by the following two terms: Domain Knowledge and Situational Fluency?* **[4 Marks]**

Appendix A

Logistic function	$f(x)=\frac{1}{1+\exp^{-x}}$
Euclidean distance	$d(x_1,x_2)=\sqrt{\sum_{r=1}^{n}\left(a_r(x_1)-a_r(x_2)\right)^2}$
Cosine similarity	$\text{cosine}(x_1,x_2)=\frac{x_1 \bullet x_2}{\lVert x_1\rVert \times \lVert x_2\rVert}$ $\text{cosine}(x_1,x_2)=\frac{\sum_{r=1}^{n}a_r(x_1)\times a_r(x_2)}{\sqrt{\sum_{r=1}^{n}a_r(x_1)^2}\times\sqrt{\sum_{r=1}^{n}a_r(x_2)^2}}$
Minkowski distance	$MD_p(x_1,x_2)=\left(\sum_{r=1}^{n}\lvert a_r(x_1)-a_r(x_2)\rvert^p\right)^{\frac{1}{p}}$
Entropy of the prior	$H\left(P(v_1),\ldots,P(v_n)\right)=\sum_{i=1}^{n}-P(v_i)\log_2 P(v_i)$
Bayes rule	$P(a\mid b)=\frac{P(b\mid a)P(a)}{P(b)}$

Appendix B

Table of Base 2 Logs for Different Fractions

$\log_2(a/b)$		a													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
	1	0.00													
	2	-1.00	0.00												
	3	-1.58	-0.58	0.00											
	4	-2.00	-1.00	-0.42	0.00										
	5	-2.32	-1.32	-0.74	-0.32	0.00									
	6	-2.58	-1.58	-1.00	-0.58	-0.26	0.00								
	7	-2.81	-1.81	-1.22	-0.81	-0.49	-0.22	0.00							
b	8	-3.00	-2.00	-1.42	-1.00	-0.68	-0.42	-0.19	0.00						
	9	-3.17	-2.17	-1.58	-1.17	-0.85	-0.58	-0.36	-0.17	0.00					
	10	-3.32	-2.32	-1.74	-1.32	-1.00	-0.74	-0.51	-0.32	-0.15	0.00				
	11	-3.46	-2.46	-1.87	-1.46	-1.14	-0.87	-0.65	-0.46	-0.29	-0.14	0.00			
	12	-3.58	-2.58	-2.00	-1.58	-1.26	-1.00	-0.78	-0.58	-0.42	-0.26	-0.13	0.00		
	13	-3.70	-2.70	-2.12	-1.70	-1.38	-1.12	-0.89	-0.70	-0.53	-0.38	-0.24	-0.12	0.00	
	14	-3.81	-2.81	-2.22	-1.81	-1.49	-1.22	-1.00	-0.81	-0.64	-0.49	-0.35	-0.22	-0.11	0.00