DUBLIN INSTITUTE OF TECHNOLOGY

DT228A/1 MSc. in Computing

SUMMER EXAMINATIONS 2016/2017

MACHINE LEARNING [SPEC9270]

DR. JOHN MCAULEY
DR. DEIRDRE LILLIS
DR. GEORGIANA IFRIM

WEDNESDAY 17TH MAY 4.00 P.M. – 6.00 P.M.

TWO HOURS

PLEASE ANSWER 2 QUESTIONS

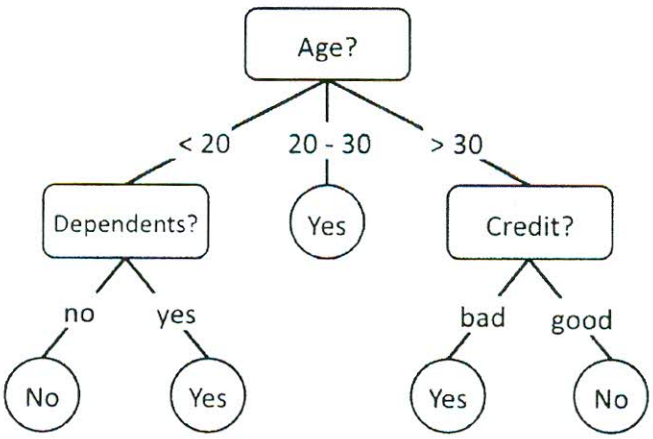
EACH QUESTION IS WORTH 50 MARKS

1.	(a)	**Please Answer the following Questions. Each question carries equal Marks.** **[24 marks]**	
		(i)	Explain the difference between boosting and bagging in Ensemble Modelling?
		(ii)	Why would you use Information Gain Ratio instead of Information Gain?
		(iii)	What is an error function? Briefly describe how an error function is used in machine learning?
		(iv)	What is Cosine Similarity and why would you use it?
		(v)	What is the difference between the mean square error and the root mean square error and what is the advantage of using the root mean square error over the mean square error?
		(vi)	Describe the difference between **random sampling** and **stratified sampling**. When would you use stratified sampling instead of random sampling?
	(b)	The table below shows the age of each employee at a card board box factory. ID: 1 2 3 4 5 6 7 8 9 10 AGE: 51 39 34 27 23 43 41 55 24 25 ID: 11 12 13 14 15 16 17 18 19 20 AGE: 38 17 21 37 35 38 31 24 35 33 Based on this data calculate the following summary statistics for the AGE feature:	
		i)	Calculate the minimum, maximum and range **[6 Marks]**
		ii)	Calculate the Mean and median: **[4 Marks]**

		iii)	Calculate the Variance and Standard Deviation: **[5 Marks]**
		iv)	Calculate the 1st quartile (25th percentile) and 3rd quartile (75th percentile) **[2 Marks]**
		v)	Calculate the Inter-quartile range **[1 Mark]**
	(c)	The following table lists a dataset collected by an online micro-loans service capturing details of historical loans and whether or not borrowers defaulted.	

ID	Age	Income	Dependents	Credit	Default
C-01	<20	81	no	bad	No
C-02	<20	76	no	good	No
C-03	20-30	86	no	bad	Yes
C-04	>30	84	no	bad	Yes
C-05	>30	45	yes	bad	Yes
C-06	>30	66	yes	good	No
C-07	20-30	41	yes	good	Yes
C-08	<20	68	no	bad	No
C-09	<20	32	yes	good	Yes
C-10	>30	56	yes	bad	Yes
C-11	<20	58	yes	good	Yes
C-12	20-30	52	no	good	Yes
C-13	20-30	90	yes	bad	Yes
C-14	>30	69	no	good	No

This dataset has been used to build a **decision tree** that can predict whether or not new borrowers will default. This decision tree is shown below.



i) The **information gain** of the feature Age at the root node of the tree is 0.247. A colleague has suggested that Dependents would be a better feature to query at the root node of the tree. Demonstrate whether or not this is the case. Please show all workings.

[4 Marks]

ii) Suppose a training set is generated from a decision tree and then decision tree learning is applied to the training-set. Is it the case that the learning algorithm will eventually return the original tree as the training set size goes to infinity? Why or why not?

[4 marks]

2. (a) i) Given the following vector, using a clamp transformation, please remove any outliers with a threshold of [5, 225] and normalise the instances between [0,1]. Show your workings.

[1, 3, 59, 78, 66, 44, 90, 98, 100, 53, 58, 201, 203, 399, 180, 406, 480]

[5 marks]

ii) You have been given the following dataset with 5 instances from a HR department:

Age: 22, 24, 26, 29, 30, 33

Salary: 20000, 22000, 20000, 26000, 95000, 95000

Why could this dataset present a problem when developing a predictive model using similarity-based measure such as Euclidean distance? And how could this be addressed?

[3 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:

$$\text{OXYCON} = \mathbf{w}[0] + \mathbf{w}[1] \times \text{AGE} + \mathbf{w}[2] \times \text{HEARTRATE}$$

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

(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.

[12 marks]

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

[12 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.

[4 marks]

(d) *Why would you use the following techniques?*

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

		[6 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. **[16 marks]**
	(b)	A credit card issuer has built two different credit scoring models that predict the propensity of customers to default on their loans. The outputs of the first model for a test dataset are shown in the table below.

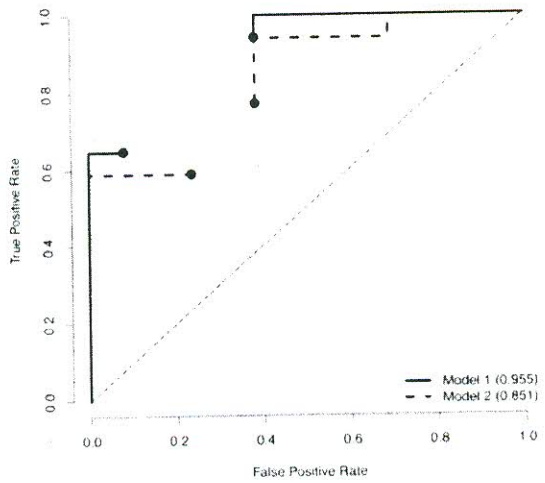
ID	Target	Score	Prediction	ID	Target	Score	Prediction
1	bad	0.634	bad	16	good	0.072	good
2	bad	0.782	bad	17	bad	0.567	bad
3	good	0.464	good	18	bad	0.738	bad
4	bad	0.593	bad	19	bad	0.325	good
5	bad	0.827	bad	20	bad	0.863	bad
6	bad	0.815	bad	21	bad	0.625	bad
7	bad	0.855	bad	22	good	0.119	good
8	good	0.500	good	23	bad	0.995	bad
9	bad	0.600	bad	24	bad	0.958	bad
10	bad	0.803	bad	25	bad	0.726	bad
11	bad	0.976	bad	26	good	0.117	good
12	good	0.504	bad	27	good	0.295	good
13	good	0.303	good	28	good	0.064	good
14	good	0.391	good	29	good	0.141	good
15	good	0.238	good	30	good	0.670	bad

The outputs of the second model for the same test dataset are shown in the table below:

ID	Target	Score	Prediction	ID	Target	Score	Prediction
1	bad	0.230	bad	16	good	0.421	bad
2	bad	0.859	good	17	bad	0.842	good
3	good	0.154	bad	18	bad	0.891	good
4	bad	0.325	bad	19	bad	0.480	bad
5	bad	0.952	good	20	bad	0.340	bad
6	bad	0.900	good	21	bad	0.962	good
7	bad	0.501	good	22	good	0.238	bad
8	good	0.650	good	23	bad	0.362	bad
9	bad	0.940	good	24	bad	0.848	good
10	bad	0.806	good	25	bad	0.915	good
11	bad	0.507	good	26	good	0.096	bad
12	good	0.251	bad	27	good	0.319	bad
13	good	0.597	good	28	good	0.740	good
14	good	0.376	bad	29	good	0.211	bad
15	good	0.285	bad	30	good	0.152	bad

Based on the predictions of these models, perform the following tasks to compare model performance:

(i) The image below shows an ROC curve for each model. Each curve has a point missing.



Calculate the missing point in the ROC curves for Model 1 and Model 2. To generate the point for Model 1, use a threshold value of 0.51. To generate the point for Model 2, use a threshold value of 0.43.

[14 marks]

(ii) The area under the ROC curve (AUC) for Model 1 is 0.955 and

			for Model 2 is 0.851. Which model is performing best? **[4 marks]**
		(iii)	Based on the AUC values for Model 1 and Model 2, calculate the Gini coefficient for each model. **[6 marks]**
	(c)	Over fitting can be a problem with Decision Tree Induction. Describe **two techniques** that can be used to address over fitting. **[5 marks]**	
	(d)	In relation to the application of machine learning, describe what you understand by the following two terms: Domain Knowledge and Situational Fluency. **[5 marks]**	