University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SEMESTER I EXAMINATION – 2016/2017

COMP 47490

Machine Learning

Prof. S. Dobson
Prof. P. Cunningham
Dr. A. Lawlor
Dr. D. Greene*

Time allowed: 2 hours

Instructions for candidates

Answer any <u>four</u> out of six questions. All questions carry equal marks. The paper is marked out of 60.
Use of non-programmable calculators is allowed.

.

Instructions for invigilators

Use of non-programmable calculators is allowed.

Q1.

(a) The contingency table below shows the evaluation results for a spam email classifier applied to a set of 768 examples, which are annotated with two class labels: {“Spam”, “Non-Spam”}.

Calculate the *F1-measure* scores relative to each of the classes “Spam” and “Non-Spam”.

Predicted Class

Spam	Non-Spam		
620	120	**Spam**	*Real Class*
45	155	**Non-Spam**	

(b) The table below shows the number of correct and incorrect predictions made by a classifier during a 5-fold cross validation experiment on 304 examples, where the goal was to classify loan applications into one of three risk categories: {low, medium, high}.

Calculate the *overall accuracy* of the classifier across the 5 folds.

Fold	Class: Low		Class: Medium		Class: High	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	39	21	94	56	89	5
2	39	21	83	67	88	6
3	48	12	87	63	79	15
4	39	21	98	52	89	5
5	45	15	78	72	81	13

(c) Explain why skewed class distributions can be a problem when evaluating the performance of a classifier. Outline a suitable evaluation measure that could be used in such cases.

Q2.

(a) The table below shows a dataset of 8 examples described by 3 categorical features, representing booking decisions made by hotel customers. Each example has one of two class labels: Book? = {yes, no}.

Calculate the *overall entropy* for the dataset.

Customer	Stars	Pool	Gym	Book?
x1	3	Y	Y	yes
x2	2	N	N	no
x3	3	N	Y	no
x4	2	Y	N	no
x5	4	Y	Y	yes
x6	3	N	N	no
x7	2	N	Y	no
x8	4	Y	N	yes

(b) Using *Information Gain*, find the best feature to split the root of a Decision Tree classifier built on the training data from (a). Show all of your calculations.

(c) Explain the idea of *inconsistent data* in the context of decision trees. How are such cases typically handled by a decision tree classifier?

Q3.

(a) The user-item matrix below shows the purchasing history of six users for eight different products in a user-based collaborative filtering system.

Based on the data, who will be Emma's nearest neighbour? Measure similarities using the binary Jaccard Index and show your calculations.

User	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**
Maria		1		1		1			1
Emma	1		1			1		1	
Robert		1		1					
Jack	1	1	1			1			
Evan	1							1	1
Joe		1			1				1

(b) The table below was generated as part of the evaluation of a collaborative filtering system which is designed to predict customer ratings (1-5) for restaurants. The predicted and true ratings for eight test examples are given.

Describe one appropriate performance metric for evaluating these results. Calculate the performance of the system based on the metric.

Restaurant	**True Rating**	**Predicted Rating**
Honey & Co	4	3.8
Seafood Twist	5	4.7
Urban Café	2	2.2
Scott's Bistro	1	2.6
Grain Store	3	2.8
Phoenix Palace	3	3.9
Singapore Garden	5	4.8

(c) Explain what is meant by the *cold start problem* in collaborative filtering. Outline strategies that might be used to address this problem.

Q4.

(a) The table below shows a medical dataset of eight patients described by three features. Each example has one of two class labels: Flu? = {yes, no}, indicating whether the patient was diagnosed with flu.

Provide the contingency table of conditional and prior probabilities that would be used by Naïve Bayes to build a classifier for this dataset. Show your calculations.

Patient	Cough	Headache	Fever	Flu
x1	N	Mild	Y	N
x2	Y	None	N	Y
x3	N	Strong	Y	Y
x4	Y	Mild	Y	Y
x5	N	None	N	N
x6	Y	Strong	Y	Y
x7	Y	Strong	N	N
x8	Y	Mild	Y	Y

(b) Use Naïve Bayes to calculate the likelihood that the new patient *x9* below is diagnosed as having flu. Then indicate what prediction a Naïve Bayes classifier would make for this patient.

Patient	Cough	Headache	Fever	Flu
x9	N	Mild	Y	???

(c) Explain what is meant by the *independence assumption* in the context of a Naïve Bayes classifier.

Q5.

(a) When finding nearest neighbours, which distance functions would you use when comparing examples with these types of features:
(i) numerical, (ii) categorical.

In a k-NN classifier it is common to use an odd values for *k* (e.g. 3NN). Briefly discuss why odd values are generally preferred over even values.

(b) The table below is a dataset collected for the purposes of evaluating whether a particular car is acceptable to a potential customer or not {YES, NO}. There are seven examples with these features:

- maintenance price: {low, med, high, vhigh}
- doors: numeric
- persons: numeric
- lug_boot: {small, med, big}
- safety: {low, med, high}
- mileage: numeric

maintenance price	**doors**	**persons**	**lug_boot**	**safety**	**mileage**	**acceptable?**
vhigh	2	2	med	high	98000	NO
med	3	2	med	high	76000	NO
low	5	5	big	med	45000	YES
low	4	2	small	high	67000	NO
high	2	4	small	med	85000	NO
low	5	4	big	med	51000	YES

We also have a query example:

maintenance price	**doors**	**persons**	**lug_boot**	**safety**	**mileage**	**acceptable?**
low	4	4	small	high	37000	??

Based on this data:

i) Normalise all the numeric features to the range [0, 1]. Assume the following ranges:
   Mileage [1000, 100,000], Persons [0, 6], Doors [0, 5]
ii) Describe an appropriate global distance function for comparing these examples.
iii) Compute the distances between the query example and the six labelled examples, and determine which class a 3-NN classifier would assign to the query.

(c) The following table shows the pre-computed distances to a query example for a dataset containing 12 examples. What class label would a distance weighted 5NN classifier assign to the query?

v	Distance	Label
v1	0.0351	False
v2	3.5902	False
v3	0.1392	True
v4	1.8904	False
v5	3.1846	True
v6	10.0147	False
v7	1.6674	False
v8	2.2805	False
v9	7.9902	False
v10	0.0005	True
v11	48.7307	False
v12	23.4944	True

Q6.

(a) The Old Faithful geyser erupts almost every minute. It is observed that when the waiting time between eruptions is a little longer, then the subsequent eruptions last for longer. Some data has been collected on the waiting times and the duration of the eruption.

Eruption duration (mins)	**Waiting (mins)**
3.85	84.0
2.25	60.0
1.917	49.0
4.0	71.0
4.933	86.0
4.25	77.0
1.95	51.0
1.933	52.0
1.85	58.0
3.833	82.0
1.75	47.0
3.833	78.0
4.667	78.0
2.383	71.0
2.4	53.0
2.8	56.0
2.25	51.0
3.967	89.0
4.25	83.0
3.317	83.0

A simple linear model of the data is:

$$y' = \beta_0 + \beta_1 x$$

Compute the values of β_0 and β_1 using least squares.

What is the expected eruption duration of Old Faithful, if the waiting time for the previous eruption was 91 minutes?

(b) Explain what is meant by correlation (COR) and covariance (COV) in the context of linear regression?

If COV(Y, X) = 0 or COR(Y, X) = 0 is it safe to conclude there is no relation between Y and X? Explain your reasoning.

(c) Below is a dataset of skin cancer mortality rates for several US states and the latitude of those states. The residual errors for latitude and mortality are provided.

Compute the coefficient of determination r^2.

What does this tell you about the correlation between skin cancer mortality rates and latitude?

State	**Lat**	**Mort**	**Residual Error (Lat)**	**Residual Error (Mort)**
California	37.5	182	-1.38	29.4
Florida	28	197	-10.88	44.4
Iowa	42.2	128	3.32	-24.6
Maine	45.2	117	6.32	-35.6
North Carolina	35.5	199	-3.38	46.4
Michigan	43.5	117	4.62	-35.6
Colorado	39	149	0.12	-3.6
Pennsylvania	40.8	132	1.92	-20.6
Ohio	40.2	131	1.32	-21.6
Kentucky	37.8	147	-1.08	-5.6
Tennessee	36	186	-2.88	33.4
Arkansas	35	170	-3.88	17.4
New Hampshire	43.8	129	4.92	-23.6
North Dakota	47.5	115	8.62	-37.6
Louisiana	31.2	190	-7.68	37.4

oOo