University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SEMESTER I EXAMINATION - 2017/2018

COMP 47490
MACHINE LEARNING

Prof. J. Pitt
Prof. P. Cunningham
Dr. Derek Greene*
Dr. Aonghus Lawlor

Time allowed: 2 hours

Instructions for Candidates

Answer any four out of six questions. All questions carry equal marks.

Non-programmable calculators allowed.

Q1: **(15 marks)**

(a) The table below shows 11 training examples represented by three categorical features, describing cases of user preferences for hotel bookings. Each example has one of two class labels: Book? = {yes, no}. [5]

Calculate the *overall entropy* for this dataset.

Case	*Stars*	*Restaurant*	*Pool*	*Book?*
1	2	N	N	no
2	3	Y	Y	no
3	2	Y	N	no
4	3	N	Y	yes
5	4	Y	N	no
6	3	N	N	no
7	3	Y	Y	yes
8	5	Y	Y	yes
9	4	N	Y	yes
10	5	N	Y	yes
11	3	Y	Y	no

(b) Using *Information Gain*, identify the best feature to split the root node of a Decision Tree classifier built on the training data from (a). Show your calculations. [5]

(c) (i) Describe one problem that can occur with the *Information Gain* criterion when applied for feature selection in Decision Trees. [5]

(ii) Outline the key differences between the *ID3* and the *C4.5* algorithms for constructing Decision Trees.

Q2: **(15 marks)**

(a) The *contingency table* below summarises the results from an evaluation of a sentiment classification system that has been applied to a test set of 2,148 tweets. Each tweet has been classified as either “Positive” or “Negative”. [5]

Predicted Class

Positive	*Negative*		
1095	322	*Positive*	*Real Class*
214	517	*Negative*	

From the contingency table, calculate:

(i) The *precision* score for each of the classes.

(ii) The *recall* score for each of the classes.

(iii) The *F1* score for each of the classes.

(b) A more advanced sentiment analysis system applies a multi-class classification approach, where tweets are given one of three labels: {Positive, Negative, Neutral}. [5]

The table below shows a summary of the number of correct and incorrect predictions made by this system during a 5-fold cross validation experiment on the test set of 2,148 tweets.

Fold	**Class: Positive**		**Class: Negative**		**Class: Neutral**	
	Correct	*Incorrect*	*Correct*	*Incorrect*	*Correct*	*Incorrect*
1	835	196	721	116	236	44
2	909	122	500	337	224	56
3	637	394	463	374	183	97
4	979	52	795	42	207	73
5	833	198	494	343	180	100

(i) Calculate the *overall accuracy* of the system across the 5 folds.

(ii) Explain why k-fold cross validation provides a more robust evaluation in a classification problem like this, when compared with a single random train-test split.

(c) Ensemble classification involves the aggregation of predictions from multiple classifiers with the goal of improving accuracy. [5]

(i) Explain the role that *diversity* plays in ensemble classification.

(ii) How do *bagging* and *boosting* differ in the way in which they introduce diversity during ensemble generation?

Q3: **(15 marks)**

(a) The dataset below lists 11 movies, each described by 4 categorical features. Each movie has an annotated label: Watch? = {*yes*, *no*}, indicating whether or not an individual decided to watch that movie. [5]

Movie Title	*Decade*	*Genre*	*>2 hours*	*Stars*	*Watch?*
Commando	1980s	Action	No	2	No
The Money Pit	1980s	Comedy	No	3	No
Aliens	1980s	SciFi	Yes	5	Yes
Bladerunner	1980s	SciFi	Yes	5	Yes
Deep Impact	1990s	SciFi	No	2	No
Die Hard 2	1990s	Action	No	5	Yes
Groundhog Day	1990s	Comedy	No	5	Yes
The Expendables	2010s	Action	No	3	No
The Avengers	2010s	Action	Yes	4	No
Kingsman	2010s	Comedy	No	2	No
Prometheus	2010s	SciFi	Yes	4	Yes

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

(b) Based on the contingency table from (a), use Naïve Bayes to estimate the likelihood that the following new movie will be watched, which has the feature values below. Show your calculations. [5]

```
(Decade = 1990s, Genre = SciFi, >2 hours = Yes, Stars = 4)
```

(c) (i) Explain the difference between *lazy learning* and *eager learning* approaches in classification. Give an example of a classifier for each approach. [5]

(ii) Outline a scenario where an eager learning approach might be more appropriate than a lazy learning approach.

Q4: **(15 marks)**

(a) The table below shows a dataset of 11 items, each represented by 4 features. The final column in the table also indicates the cluster to which the k-means algorithm has assigned each item, when applied for $k = 3$. [5]

Item	y_a	y_b	y_c	y_d	*Cluster*
$x1$	-8.140	9.775	0.575	-5.511	1
$x2$	1.833	-3.836	5.897	-8.383	2
$x3$	-9.404	4.183	-9.614	4.711	3
$x4$	0.254	-2.838	7.472	-8.733	2
$x5$	-7.249	10.155	-1.705	-6.143	1
$x6$	-6.481	8.244	1.344	-7.153	1
$x7$	0.412	-4.796	7.653	-8.160	2
$x8$	-8.616	10.802	-0.065	-4.447	1
$x9$	0.453	-4.001	7.550	-9.366	2
$x10$	-6.934	5.789	-8.343	4.825	3
$x11$	-7.683	6.607	-10.186	4.993	3

(i) Based on the cluster assignments above, compute the *centroid vector* of each cluster.

(ii) Calculate the *Euclidean distance* between each of the centroid vectors.

(b) (i) Outline three practical reasons for reducing the number of dimensions of a dataset. [5]

(ii) Describe how PCA performs dimensionality reduction. How might we go about selecting an appropriate number of dimensions for PCA?

(c) (i) Explain the difference between *filters* and *wrappers* for feature subset selection, with reference to an example of each type of approach. [5]

(ii) Why might filters and wrappers select different feature subsets when applied on the same dataset?

Q5: **(15 marks)**

(a) The following dataset describes the refractive index (ri) of samples of glass based on their composition of the element potassium (K). [5]

Case	K	ri	$(ri - \hat{ri})^2$	$(\hat{ri} - \bar{ri})^2$	$(K - \bar{K})^2$
0	0.14	1.5156	9.5249e-06	6.4516e-06	0.266260
1	0.68	1.5165	2.6069e-06	2.6896e-06	0.000576
2	0.51	1.5185	4.2207e-08	1.2960e-07	0.021316
3	1.68	1.5151	3.5193e-06	8.7616e-06	1.048600
4	0.13	1.5222	1.2768e-05	1.7057e-05	0.276680
5	0.16	1.5157	8.3236e-06	5.5696e-06	0.246020
6	0.57	1.5185	1.0174e-07	1.6810e-07	0.007396
7	0.76	1.5206	6.7086e-06	6.1504e-06	0.010816
8	1.46	1.5184	1.2794e-06	7.8400e-08	0.646420
9	0.47	1.5199	2.6998e-06	3.3856e-06	0.034596

Using ordinary least squares, the best fit linear regression model is found to be:

$$\text{ri} = 1.5188 - 0.0011 \times K$$

i. What is the coefficient of determination (R^2) of this regression model? Is the model a good fit to the data?

ii. Use a t-test to calculate the significance of this fit. You can assume that a significance with p-value of 0.05 corresponds to a $t^*_{0.05/2,10-2} = 2.306$.

(b) i. Describe some of the differences between the *logistic regression* and *linear regression* models, mentioning in particular the different feature types to which they are applied. [5]

ii. Some of the samples of glass are known to correspond to a type of glass commonly found in the household. A new binary variable is introduced, $household = \{yes, no\}$, and a logistic regression model is trained to predict whether a sample is household glass by measuring the composition of potassium (K). The weights of the final model are:

Feature	Weight
Intercept	-7.7136
K	4.1804

Using this model, what is the probability that the following sample is household glass?

K	household
1.9	?

(c) i. Briefly describe the gradient descent algorithm, explaining how we use it to train a regression model, mentioning the cost function, the process for adjusting the weights, and the condition for stopping. [5]

ii. We would like to use batch gradient descent to learn the weights of a *linear regression* model to describe the glass dataset from part (a). We have set the learning rate $\alpha = 0.01$ and the current state of the model is:

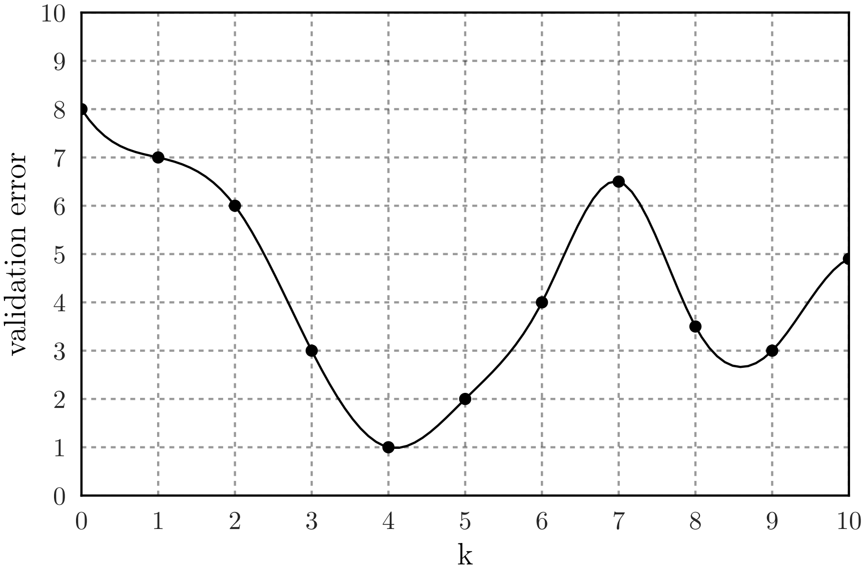
Feature	Weight
Intercept	1.665
K	-0.0105

Write out the update step of the gradient descent algorithm, and compute the state of the model (Intercept, K) after *1* full iteration step.

Q6: **(15 marks)**

(a) i. Describe the problem known as the *curse of dimensionality* as it applies to the k-nearest neighbour method. Briefly suggest some techniques for dealing with this dimensionality problem. [5]

ii. You are training a k-nearest neighbour model on a new dataset and have computed the classification by 5-fold cross validation. The figure below shows the results. What value of k is the best choice for your k-NN model? Explain your reasoning.



(b) As part of her study programme, a student visited the Pi Cafe and tried various different types of coffee and recorded her preferences. The results are reported in the table below. [5]

case	*whipped cream*	*vanilla syrup*	*dusted chocolate*	*liked*
0	True	True	True	False
1	True	False	False	True
2	False	True	True	False
3	False	True	False	True
4	True	False	False	True

i. The barista decides to make a new coffee combination. Use a 1-NN classifier and the training data above to predict if she will like the following:

whipped cream	*vanilla syrup*	*dusted chocolate*
False	False	True

ii. Use a 3-NN classifier to predict if she will like the following combination:

whipped cream	*vanilla syrup*	*dusted chocolate*
False	True	False

iii. Use a *distance-weighted* 3-NN classifier to predict if she will like the following combination:

whipped cream	*vanilla syrup*	*dusted chocolate*
False	True	False

(c) The table below shows a dataset collected by an online retailer for the purpose of evaluating whether or not a particular phone will be purchased by a customer. Each phone in the dataset is described by 6 descriptive features. [5]

Name	*Colour*	*OS*	*Screen_Size*	*Storage*	*Weight*	*Price*
Galaxy S8	Black	Android	5.8	64	5.36	625
iPhone 6	Gold	iOS	4.6	8	4.55	250
Galaxy J7	Gold	Android	5.5	32	5.89	300
iPhone 8	Black	iOS	4.7	64	5.28	900
Honor 8	Black	Android	5.2	32	4.8	350
iPhone 7 Plus	Red	iOS	5.5	128	6.63	750
Alcatel Ideal	Black	Android	4.5	8	4.83	100

i. Normalise all the numeric features to the range $[0, 1]$. Assume the following ranges: Screen_Size = $[4, 8]$, Storage = $[8, 256]$, Weight = $[4, 7]$, Price = $[100, 1000]$.

ii. Propose an appropriate global distance function for comparing the phones in this dataset.

iii. Use your proposed global distance function to find the phone in the existing dataset which is most similar to the following new example:

Name	*Colour*	*OS*	*Screen_Size*	*Storage*	*Weight*	*Price*
X	Gold	Android	5.2	32	5.15	255