University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SAMPLE EXAMINATION - 2018/2019

COMP 47460/47490
MACHINE LEARNING

Dr. Derek Greene*
Dr. Aonghus Lawlor

Time allowed: 2 hours

Instructions for Candidates

Answer Question 1 and any two from Questions 2, 3, 4.

Non-programmable calculators allowed.

Q1: _____ **(20 marks)**

(a) 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. [2]

(b) Outline one *non-linear activation function* which is commonly employed in neural networks. [2]

(c) Explain what is meant by the *independence assumption* in the context of a Naïve Bayes classifier. [2]

(d) Describe some of the differences between a lazy and eager learning approach, and explain which category you think best describes a feedforward neural network. [2]

(e) Outline one situation where a cluster validation measure might be applied in unsupervised learning. [2]

(f) When generating an ensemble of classifiers using a *random subspace* strategy, what might be the effect of choosing a subspace size that is either too low or too high? [2]

(g) What is meant by a *feed-forward architecture* in the context of neural networks? [2]

(h) Why might two wrappers employing different greedy search strategies select different feature subsets when applied on the same dataset? [2]

(i) Describe one problem that can occur with the *Information Gain* criterion when applied for feature selection in Decision Trees. [2]

(j) What is meant by the *curse of dimensionality*? What are some common approaches in machine learning to deal with this problem? [2]

Q2: _____ **(15 marks)**

(a) i. A scientist measures the resistance, Ω, in a device for different temperatures T from 50°C to 100°C. She determines the slope of the best fit line to be 0.1021 ($\bar{T} = 75.325$, $\bar{\Omega} = 19.5074$). Write out the linear regression model for this dataset. [1]

ii. Describe the sources of error (total error, regression error, residuals) in a linear regression model, and how they are related to each other. [2]

iii. Given that $SSE = 1.6123$ and $SSR = 38.2648$, what is the coefficient of determination R^2? [2]

(b) The confusion matrix below summarises the performance of a binary classifier, applied to a dataset of examples, which are annotated with 2 class labels {A, B}. [5]

	A	B
A	470	160
B	50	120

Based on this matrix, calculate:

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

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

(iii) The overall classification accuracy.

(c) (i) Suggest how we might choose an appropriate value for the parameter k when building a kNN classifier. [5]

(ii) Explain the difference between an *unweighted* kNN classifier and a *weighted* kNN classifier. For the latter, suggest an approach for calculating weights.

Q3: _____ **(15 marks)**

(a) The table below shows a training set with 7 examples represented by 4 categorical features, describing a person's preferences for booking hotels. Each example has a binary class label: Book? = {*yes*, *no*} [5]

Example	Stars	Pool	Beach	Gym	Book?
Hotel 1	2	N	N	Y	no
Hotel 2	2	Y	N	N	yes
Hotel 3	3	N	Y	N	no
Hotel 4	3	Y	N	Y	no
Hotel 5	3	N	N	N	no
Hotel 6	3	Y	Y	Y	yes
Hotel 7	4	Y	Y	Y	yes

(i) Calculate the *overall entropy* for this data.

(ii) Using *Information Gain*, identify the best feature to split the root node of a Decision Tree classifier built on the training set. Show your calculations.

(b) (i) In the context of supervised learning, what is the difference between *overfitting* and *underfitting*? [5]

(ii) Briefly outline one real-world application of classification, where the practical implications a *False Positive* error and a *False Negative* error might differ.

(c) (i) Explain the difference between *single linkage*, *average linkage*, and *complete linkage* in the context of hierarchical agglomerative clustering. [5]

(ii) Which of these linkage strategies would you expect to be most affected by the presence of outliers in a dataset?

Q4: _____ **(15 marks)**

(a) The dataset below has 9 specifications for laptops, each described by 4 categorical features. Each example has one of two class labels: Buy? = {*yes*, *no*}, indicating if a person will purchase the laptop. [5]

Laptop	Drive	ScreenSize	Weight	Price>1k	Buy?
l_1	SSD	15	Medium	No	Yes
l_2	SSD	13	Medium	No	Yes
l_3	HDD	15	Medium	Yes	No
l_4	SSD	11	Light	No	No
l_5	HDD	15	Heavy	Yes	No
l_6	HDD	15	Heavy	No	No
l_7	SSD	15	Heavy	Yes	Yes
l_8	SSD	13	Heavy	No	No
l_9	HDD	13	Medium	No	Yes

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

(ii) Based on the contingency table, use Naïve Bayes to estimate the likelihood that the following new laptop will be purchased. Show your calculations.
`(Drive = HDD, ScreenSize = 15, Weight = Medium, Price>1k = No)`

(b) The table below shows a dataset of 6 examples, each represented by 4 numeric features: [5]

Example	f_1	f_2	f_3	f_4
x_1	6.3	2.5	5.0	1.9
x_2	4.6	3.4	1.4	0.3
x_3	5.4	3.9	1.7	0.4
x_4	6.7	3.0	5.2	2.3
x_5	6.7	3.3	5.7	2.5
x_6	5.0	3.6	1.4	0.2

These 6 examples have been assigned to 2 clusters by k-means as follows:
$C_1 = \{x_2, x_3, x_6\},\ C_2 = \{x_1, x_4, x_5\}$

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

(ii) To which cluster would k-means assign the new example x_7? Show your calculations.
$x_7 = (6.00, 2.25, 4.60, 1.60)$

(iii) Outline the main disadvantages of the k-means clustering algorithm.

(c) (i) Explain how a simple *perceptron* can be used for binary classification. [5]

(ii) What is the role of the *cost function* in a neural network? Outline one cost function which would be appropriate for use in a binary classification task.