University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

SEMESTER I EXAMINATION - 2018/2019

COMP 47490
MACHINE LEARNING

Prof. J. Pitt
Prof. P. Cunningham
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 what is meant by *generalisation* in the context of supervised learning, with reference to the problem of overfitting. [2]

(b) Explain how a kNN classifier might potentially be affected by unbalanced class sizes in a training set. [2]

(c) What is meant by *inconsistent data* in the context of decision trees? How might such cases be handled by a decision tree classifier? [2]

(d) Briefly explain the role of *bias terms* in a neural network architecture. [2]

(e) Explain the difference between *feature selection* and *feature transformation* approaches for dimension reduction. Give one example of each. [2]

(f) What is the difference between *lazy* and *eager* learning strategies in classification? Give an example of a classifier for each category. [2]

(g) Why is the initialisation step important in the k-Means clustering algorithm? [2]

(h) Give one example of how a *cluster validation* measure might be applied in unsupervised learning. [2]

(i) The logistic regression model is given by: [2]

$$P(Y = 1|X = x) = \frac{\exp \beta_0 + \beta_1 x}{1 + \exp \beta_0 + \beta_1 x}$$

How do we interpret the coefficients β_0 and β_1?

(j) Explain some of the reasons why ensemble methods tend to provide better accuracy than individual models. [2]

Q2: **(15 marks)**

(a) The training set below lists 10 cars, each described by 3 categorical features. Each car has a label, Sold? = {Yes,No}, which records whether or not the car was recently sold. [5]

Example	Body Type	Transmission	Colour	Sold?
x_1	SUV	Automatic	Silver	Yes
x_2	SUV	Manual	White	Yes
x_3	Saloon	Manual	Silver	No
x_4	SUV	Manual	Red	Yes
x_5	Saloon	Automatic	Silver	Yes
x_6	Saloon	Automatic	White	No
x_7	SUV	Automatic	Red	Yes
x_8	SUV	Manual	Silver	No
x_9	SUV	Manual	Red	No
x_{10}	Saloon	Automatic	Red	No

(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 car will be sold. Show your calculations.
`(Body Type = SUV, Transmission = Manual, Colour = Silver)`

(b) (i) Explain the role that *diversity* plays in ensemble classification. [5]

(ii) How do *bagging* and *random subspacing* differ in the way in which they introduce diversity to an ensemble?

(c) (i) Briefly explain the role of the gradient descent algorithm in training neural networks. [5]

(ii) How does *stochastic gradient descent* optimisation differ from the standard gradient descent algorithm?

(iii) Given an intuitive answer as to why adding hidden nodes in a neural network increases the set of functions that can be learned by the network.

Q3: **(15 marks)**

(a) A crop scientist is investigating the effect of fertiliser on the yield of their pea plants. She designs an experiment to grow the plants in a controlled environment and varies the amount of fertiliser (in grams) that she gives to each plant and records the final height they grow to (in centimetres). [5]

Her results are given in the following table:

plant	A	B	C	D	E	F	G	H
grams	1.05	1.486	2.065	2.652	2.977	3.477	4.158	4.577
height	3.329	7.331	7.073	9.32	12.284	9.09	12.284	16.783

(i) Compute the equation of the least square linear regression line of plant height against grams of fertiliser.

(ii) What is the expected height of a plant which is fed with 4.0 grams of fertiliser?

(iii) Using the t-test, determine if a significant relationship exists between fertiliser amount and plant height at a significance level of 0.05.
The two-sided t-statistic is $t_{\alpha/2,N-2} = 2.447$ *for* $\alpha = 0.05$ *and* $N = 8$
To aid your calculations you can use these results:
$\sum_i(\hat{y}_i - y_i)^2 = 20.38397$ *and* $\sum_i(x_i - \bar{x})^2 = 10.84256$

(b) (i) What is the key difference between *agglomerative* and *divisive* strategies for hierarchical clustering? [5]

(ii) Explain the role of the *cluster metric* in agglomerative hierarchical clustering. Describe two examples of such a metric.

(c) (i) Why might we want to reduce the number of dimensions used to represent a dataset when performing classification? [5]

(ii) Outline the key differences between *filter* and *wrapper* strategies for supervised feature selection. What are the advantages and disadvantages of each strategy?

(iii) Explain how *backward elimination* works in the context of wrapper feature selection.

Q4: **(15 marks)**

(a) The table below shows a training set with 10 examples represented by 4 categorical features, describing individuals' preferences for renting a property. Each example has a binary class label: Rent? = {yes, no} [5]

Example	Beds	Parking	Furnished	Garden	Rent?
x_1	4	Y	Y	Y	yes
x_2	3	Y	Y	Y	yes
x_3	2	Y	N	Y	no
x_4	2	N	N	N	no
x_5	4	Y	N	Y	yes
x_6	4	N	N	Y	no
x_7	3	N	Y	N	no
x_8	3	N	N	Y	no
x_9	3	Y	Y	N	yes
x_{10}	4	N	Y	Y	no

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

(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) The confusion matrix below summarises the performance of a binary spam email classifier. From this table, calculate the *F1*-measure score for the *Non-Spam* class which was achieved by the classifier. [5]

	Spam	Non-Spam
Spam	216	54
Non-Spam	60	270

(ii) Explain why classification accuracy might not always be an adequate measure of predictive performance.

(c) (i) When finding nearest neighbours, which distance functions would you use when comparing examples with these types of features? (a) numerical, (b) ordinal. [5]

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

oOo