SEMESTER 1, 2023 EXAMINATIONS

CITS5508 Machine Learning

This paper contains 12 pages (including title page)

Time allowed: 2 hours (including reading time)

INSTRUCTIONS:

This mock exam has 10 questions with a total value of 100 marks.
Write all answers in the answer booklet provided.
UWA approved calculators with stickers are permitted.
Marks are given for clarity and correctness of the answer, not just for correct answers.

This page has been left blank.

Question 1

Consider a data set with $m = 100$ instances containing one variable (X_1) and one response variable (Y). Consider these two possible models to fit these instances: a linear regression model and a cubic regression.

(a) Describe the linear regression, i.e. $Y = ?$ 1 marks

(b) Describe the cubic regression, i.e. $Y = ?$ 2 marks

(c) Suppose the true relationship between X_1 and Y is linear. Consider the *training* mean squared error for the linear regression, and the *training* mean squared error for the cubic regression. Is it expected one to be lower than the other, is it expected them to be the same, or is there not enough information to tell? Justify your answer. 3 marks

(d) Suppose the true relationship between X_1 and Y is linear. Consider the *test* mean squared error for the linear regression, and the *test* mean squared error for the cubic regression. Is it expected one to be lower than the other, is it expected them to be the same, or is there not enough information to tell? Justify your answer. 3 marks

(e) Suppose the true relationship between X_1 and Y is not linear, but we don't know how far from linear it is. Consider the *test* mean squared error for the linear regression, and the *test* mean squared error for the cubic regression. Is it expected one to be lower than the other, is it expected them to be the same, or is there not enough information to tell? Justify your answer. 3 marks

/ 12

Question 2

(a) Explain boosting in ensemble learning. 4 marks

(b) Give two examples of boosting methods and briefly describes how they work. 6 marks

/ 10

Question 3

Consider the training data set with six instances presented in the table below. There are three variables (X_1, X_2, and X_3), and one response variable (Y).

Table 1: Training instances for the target variable Y.

ID	X_1	X_2	X_3	Y
1	0	3	0	Apple
2	2	0	0	Apple
3	0	1	3	Apple
4	0	1	2	Orange
5	-1	0	1	Orange
6	1	1	1	Apple

Consider the test instance $X_1 = X_2 = X_3 = 0$ for which we want to make a prediction using the K-Nearest Neighbours algorithm.

(a) Compute the Euclidean distance between each training instance and the test instance.

(Recall that the Euclidean distance between data instances $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as: $D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{\ell=1}^{n} (\mathbf{x}_i[\ell] - \mathbf{x}_j[\ell])^2}$ where n is the number of features.) 4 marks

(b) What is the prediction for $K = 1$? Why? 3 marks

(c) What is the prediction for $K = 3$? Why? 3 marks

/ 10

Question 4

(a) Describe the main steps of the agglomerative hierarchical clustering algorithm, including three of the most commonly used types of linkage. 5 marks

(b) Give three practical challenges related to hierarchical clustering algorithms. 3 marks

/ 8

Question 5

(a) Explain the differences between multioutput classification and multiclass classification. Give one example of a real-life application in each case. Each example should include a description of the features and the response(s) variable(s). 6 marks

(b) Describe one real-life application in which classification might be useful. You should include a description of the features and the response(s) variable(s). 2 marks

(c) Describe the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification. 2 marks

/ 10

Question 6

(a) Define *true positive rate* and *false positive rate*. 4 marks

(b) Describe three of the main challenges in machine learning. Give a brief description of each challenge. 6 marks

/ 10

Question 7

The data set below contains six instances related to criminals who reoffended within two years of realise. These criminals are known as *recidivists*.

Table 2: A dataset describing prisoners released on parole, and whether they reoffended within two years of release given by the binary target feature, Recidivist.

ID	Good Behaviour	Age < 30	Drug Dependent	Recidivist
1	False	True	False	True
2	False	False	False	False
3	False	True	False	True
4	True	False	False	False
5	True	False	True	True
6	True	False	False	False

Entropy can be used as a measure of impurity in a collection S of training instances. If the target attribute can take on c different values, the entropy of S relative to this c-wise classification is defined as

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is the proportion of S belonging to the class i.

Given the entropy as a measure of impurity in a collection of training instances, we can define a measure of the effectiveness of an attribute A in classifying the training data. A common measure is the *information gain*, defined as the expected reduction in entropy caused by partitioning the examples according to this attribute A:

$$InfGain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where $Values(A)$ is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (i.e., $S_v = \{s \in S | A(s) = v\}$).

(a) What is the entropy of the collection S given by the six instances in Table 2? 2 marks

(b) Using this dataset, construct the decision tree resulting from splitting the descriptors using entropy-based information gain. 8 marks

(c) Given the tree you found in (b), what is the prediction for the following query?

Good Behaviour = False, Age < 30 = False, Drug Dependent = True 2 marks

/ 12

Question 8

(a) Suppose you are using ridge regression, and you notice that the training and validation errors are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularisation hyperparameter α or reduce it? 3 marks

(b) Explain when new training instances do not change a fitted SVM classification model. 3 marks

(c) Machine Learning systems can be classified according to the amount and type of supervision they get during training. Describe the main categories under this classification. 4 marks

/ 10

Question 9

In this question, you will perform K-means clustering manually, with K=2, on a small data set with $n = 2$ features and $m = 6$ instances as described in Table 3.

Table 3: Values for features X_1 and X_2 for six training instances.

ID	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the instances. 2 marks

(b) Randomly assign a cluster label to each instance. Report the cluster labels for each instance. 1 marks

(c) Following (b), run the K-means algorithm until the clusters obtained stop changing. In each iteration, report: (1) the computed centroids for each cluster, and (2) the resulting cluster labels for each instance after the step (1). 5 marks

(d) In your plot from (a), use different symbols for instances according to their final cluster, and use 'X' to show the locations of the two final centroids. 2 marks

/ 10

Question 10

Explain PCA and its usefulness, and give two interpretations of the first principal component of a set of features. 8 marks