University College Dublin
An Coláiste Ollscoile Baile Átha Cliath

2022/2023 AUTUMN TRIMESTER EXAMINATIONS

COMP47750

Machine Learning with Python

Module Coordinator: Professor Pádraig Cunningham

Student Number

Seat Number

Time Allowed: 60 minutes

Materials Permitted in the Exam Venue:
Non-programmable or scientific calculator
Foreign language dictionary (hard copy)

Materials to be Supplied to Students:
8 Page Answer Booklets

Instructions to Students:
Answer Question 1 and any two other questions. Question 1 is worth 40 marks and all other questions are worth 30 marks each. The value of each part of each question is shown in brackets next to it.

Question 1

a. Is the following statement true or false "*A k-Nearest Neighbour classifier will always have perfect accuracy if the training data is used for testing"?* Explain your answer.

(5 marks)

b. When decision tree pruning results in the removal of a sub-tree with the root of that sub-tree becoming a leaf node, what is the class label associated with this new leaf node?

(5 marks)

c. The table below shows some data that is to be used for constructing a decision tree; the class labels are P and N and the values for feature F1 are numeric. What are the split points for F1 that need to be considered in calculating the Information Gain for this feature? That is, if F1 is being considered for a binary split what split points should be considered?

F1	3.5	3.5	4	5	5	6	7	7.5	8	8
Label	P	P	P	N	N	P	N	N	N	N

(5 marks)

d. The Naive Bayes classifiers in scikit-learn have a **`fit_prior`** parameter that can be true or false. What are the two options for setting class priors that this parameter controls?

(5 marks)

e. Why is it not possible to use *k*-means clustering with categorical data?

(5 marks)

f. Why is it important for Neural Networks to have cost (loss) functions that are differentiable?

(5 marks)

g. In wrapper-based feature selection, forward sequential search is inclined to be faster than backward elimination, why would that be?

(5 marks)

h. In Machine Learning what is the difference between hyperparameters and other parameters? Give an example of each.

(5 marks)

Question 2.

a. If a distance measure is a proper metric it must meet the following four criteria:

   i. $d(x,x) = 0$
   ii. If $x \neq y$ then $d(x,y) > 0$
   iii. $d(x,y) = d(y,x)$
   iv. $d(x,z) \leq d(x,y) + d(y,z)$

   Based on these criteria, is Dynamic Time Warping a proper metric? Explain your answer.

(10 marks)

b. In developing *k*-Nearest Neighbour classifiers, what would be a consequence of using a distance measure that is not a metric?

(8 marks)

c. Describe two methods for normalising data in Machine Learning. Why is data normalisation important when using a *k*-Nearest Neighbour classifier?

(12 marks)

Question 3.

a. What is the role for Confusion Matrices in the evaluation of Machine Learning Classifiers? What insights might be offered from examining a Confusion Matrix that might not come from considering the Misclassification Rate only?

(10 marks)

b. What is the difference between hold-out testing and cross validation? When compared with hold-out testing, cross-validation testing can provide estimates of generalisation accuracy that are:

   i. more accurate
   ii. more stable

   Explain both of these claims.

(12 marks)

c. In what circumstances would cross-validation have no advantages over hold-out testing?

(8 marks)

Question 4.

a. Explain why diversity is important in Machine Learning ensembles for classification and regression.

(10 marks)

b. Bagging (bootstrap aggregation) has a mechanism for achieving diversity; explain how it works.

(7 marks)

c. Bootstrap resampling will not produce diversity in ensembles of nearest-neighbour classifiers; describe a process that will.

(6 marks)

d. Explain what differentiates the different ensemble members in a Boosting Ensemble.

(7 marks)

oOo