University College Dublin
An Coláiste Ollscoile, Baile Átha Cliath

_____

Autumn Trimester Exams, 2021 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

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. The potential for an ensemble of predictors to achieve an accuracy higher than that of a single predictor depends on there being diversity in the ensemble members. Name two mechanisms by which diversity can be achieved in ensembles.

(5 marks)

b. The calculations below relate to a binary classification example (Yes/No) for samples with three predictive features, for this example the values for these features are V1,V2, V3.

$$P(Yes|data) = P(V1|Yes) \times P(V2|Yes) \times P(V3|Yes) \times P(Yes)$$

$$P(Yes) = (1/4 \times 2/4 \times 2/4) \times 4/10$$

$$P(No|data) = P(V1|No) \times P(V2|No) \times P(V3|No) \times P(No)$$

$$P(No) = (2/6 \times 2/6 \times 4/6) \times 6/10$$

In this case the prior probabilities have been set from the data. If the prior probabilities were set to be equal would the result change?

(5 marks)

c. *"When training Machine Learning algorithms using Gradient Descent it makes sense to increase the learning rate as the algorithm gets close to convergence."* Is this statement True or False?

(5 marks)

d. *Machine Learning algorithms that do not require data normalisation prior to use include Multivariate Regression and Decision Trees.* Is this statement True or False? Explain your answer.

(5 marks)

e. Popular performance measures in regression are the R-squared statistic, MAPE and MAE. Which of these are effective for providing insight into the size of the error in terms of the quantity being predicted?

   (i) R-squared statistic

   (ii) R-squared statistic and MAPE

   (iii) MAE and MAPE

   (iv) R-squared statistic, MAE and MAPE

(5 marks)

f. In theory, to build a good classifier, the number of examples required per feature increases exponentially with the number of features. In practice, this is often not true. Give one reason why this might not be true.

(5 marks)

g. The DecisionTreeClassifier in scikit-learn has a parameter `'max_leaf_nodes'` that sets a limit on the number of leaf nodes in a tree. If this limit is reduced what impact will that have on the accuracy on the training data and the test data?

(5 marks)

h. Possible causes of overfitting in supervised machine learning include:

   (a) lack of training data

   (b) model is too simple

   (c) training data contains noise

   (a) and (b) and (c)

   (a) and (c)

(5 marks)

Question 2

a. Explain the difference between stochastic gradient descent (SGD) and batch gradient descent. What are the potential benefits of SGD? What are the possible drawbacks?

(15 marks)

b. The Neural Network implementation in **scikit learn** provides the **alpha** parameter to control overfitting.

Overfitting can also be managed by controlling the model complexity, by reducing the number of layers and by reducing the number of units in each layer.

A third strategy for controlling overfitting in Neural Networks is early stopping.

Explain what overfitting means in the context of Neural Networks. Describe how two of these three strategies work.

(15 marks)

Question 3

a. Discuss the stability problems associated with the *k*-Means algorithm. What strategy is used in **scikit learn** to address this instability?

(10 marks)

b. Explain the importance of the concept of 'graph cut' in spectral clustering.

(10 marks)

c. Spectral clustering requires data in the format of an affinity matrix, an *m* by *m* matrix for a dataset with *m* samples. Describe one method to convert data in a feature vector format to an affinity matrix.

(10 marks)

Question 4

a. In boosting, the ensemble members are trained in sequence (one after the other) with the distribution of the training data changed at each stage in the training process. Explain how this works and explain the underlying principle.

(15 marks)

b. In machine learning, error can be decomposed into **bias** and **variance** components. Explain these concepts with the aid of examples. What impact would you expect an ensemble to have on these error components?

(15 marks)

oOo