University College Dublin
An Coláiste Ollscoile Baile Átha Cliath

2023/2024 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. In *k*-Nearest Neighbour classification, it is common to set *k* to be an odd number to avoid ties. Describe a situation where an odd value of *k* will not always avoid ties.

(5 marks)

b. With hold-out testing on a small dataset of fixed size, what is the likely impact of increasing the size of the test set on the accuracy estimate:

(a) increase the estimate and reduce the variance of this estimate.
(b) increase the estimate and increase the variance.
(c) decrease the estimate and reduce the variance.
(d) decrease the estimate and increase the variance.

(5 marks)

c. When training decision trees it is common to set the limit on the number of samples required for a leaf node to control overfitting. If this limit is reduced which of the following is most likely:

(a) training error reduced, test error increased
(b) training error reduced, test error reduced
(c) training error increased, test error increased
(d) training error increased, test error reduced

(5 marks)

d. It is normal for Neural Network implementations such as scikit-learn to use an alpha parameter for regularisation. What does this do and how does this work?

(5 marks)

e. Explain how mean absolute percentage error (MAPE) is calculated. Outline one situation where it should not be used.

(5 marks)

f. Neural Networks have both hyperparameters and 'ordinary' parameters. Give an example of each.

(5 marks)

g. The *k*-Means clustering implementation in scikit-learn has an **n-init** parameter that controls the "number of times the *k*-Means algorithm is run with different centroid seeds." What is the purpose of this?

(5 marks)

h. Initially, as new members are added to an ensemble the accuracy improves. Eventually, the addition of additional members no longer results in an increase in accuracy. Why is this?

(5 marks)

(40 marks for Q1)

Question 2

a. The Naive Bayes implementations in scikit-learn have a **fit-priors** parameter that controls how the class priors are set.
Describe two strategies for setting the class priors and describe scenarios where each of these scenarios would be appropriate.
(8 marks)

b. Feature selection methods are normally divided into filter and wrapper categories. In addition, some machine learning methods such as decision trees can be said to entail intrinsic (implicit) feature selection.
   i. Explain with examples the difference between wrapper and filter feature selection methods.

   ii. Explain what is meant by intrinsic (implicit) feature selection in decision trees.

(12 marks)

c. Describe two principles that drive PCA in reducing the dimension of a dataset. Explain how these principles result in a reduction in the number of dimensions.

(10 marks)
(30 marks for Q2)

Question 3

a. The table below was generated as part of the evaluation of a collaborative filtering system which is designed to predict customer ratings (1-5) for restaurants. The predicted and true ratings for six test examples are given.

   i. Describe one appropriate performance metric to provide an insight into the accuracy of results. Calculate the performance of the system based on the metric.

   ii. Propose a measure for quantifying any bias in these results. Calculate the overall bias.

Restaurant	True Rating	Predicted Rating
Wooden Spoon	4	3.7
Clubhouse	5	4.7
Posh	2	2.3
Fred's Bistro	3	2.6
Coffee Bean	3	2.8
Claddagh Palace	3	3.9

(10 marks)

b. It is standard practice to normalise the data in advance of fitting a multivariate regression model but it is not essential.

   i. If the data is not normalised, what impact will that have on accuracy?

   ii. Describe one advantage of data normalisation in multivariate regression.

(10 marks)

c. The SGDRegressor implementation has a learning rate parameter.

   i. In terms of the stochastic gradient descent process, what does this parameter control?

   ii. What are the consequences of setting a fixed (constant) learning rate that is too large?

(10 marks)
(30 marks for Q3)

Question 4

a. In contrast with Bagging ensembles, Boosting ensembles have to be trained in sequence one after the other. Why is this?

(8 marks)

b. The equations for weight update in a Boosting ensemble are as follows:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if} \quad h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if} \quad h_t(x_i) \neq y_i \end{cases}$$

where:

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

and ϵ_t is the error for classifier t

i. What happens when a training sample is classified correctly?

ii. What happens when a sample is incorrectly classified?

iii. What is the role of Z_t?

(15 marks)

c. While an ensemble will normally have a lower classification error than the component classifier, it is sometimes the case that a boosting ensemble can have a higher error. Explain how this might happen.

(7 marks)
(30 marks for Q4)

oOo