TECHNOLOGICAL UNIVERSITY DUBLIN
TALLAGHT CAMPUS

Diploma
Bachelor of Science (Honours)
Higher Diploma in Science

Machine Learning AI
Computing Software Development
Computing with IT Management
Computing in Data Analytics

ACCS

Semester Seven : January 2021

Applied Machine Learning

Internal Examiners
Dr Keith Quille

External Examiners
Dr Nigel Whyte
Dr Dermot Boyle

Day Thursday
Date 07/01/2021
Time 09:30 - 11:30

Instructions to Candidates

Answer any 3 questions from 4.

Answer any 3 questions from 4, all questions carry equal marks.

Question 1

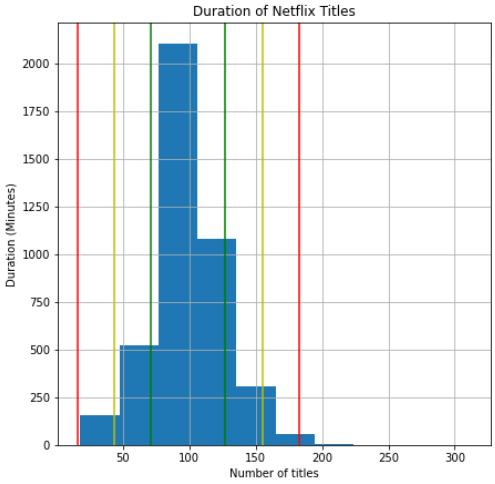
(A) "Often when pre-processing data, it can be useful to use standard deviation to identify outliers in the data".

(i) Explain in your own words an approach you might take using standard deviation to identify outliers, outlining the values/ranges you might use and why.

(ii) In many cases, standard deviation and specific ranges are used as a rigid rule with no flexibility. Describe two problem contexts you may have come across where using standard deviation and specific ranges was not a suitable approach.

(6 Marks)

(B) The histogram presented below, presents all 4257 Netflix movie titles, and their duration in minutes.



Attribute: Min (Minutes)

count	4257.000000
mean	99.265680
std	27.841744
min	18.000000
25%	86.000000
50%	98.000000
75%	115.000000
max	312.000000

(i) Evaluate the histogram for potential outliers and/or deemed missing data. Explain the reasoning behind the choices you have made (provide the steps and calculations you used to support your decisions)

(ii) Explain the actions you may take, based on the findings from part (i).

(8 Marks)

(C) "Pearson's correlation coefficient is a suitable technique to identify multicollinear data and for attribute selection for all numerical datasets".

(i) Is this statement true? By what reasoning did you come to that conclusion?

(ii) Describe a problem situation(s) which supports your answer from part (i).

(6 Marks)

Question 2

(A) The following table contains the performance results for classification models ***a*** and ***b***, (Accuracy, Sensitivity and Specificity). Both models are trying to predict if an Netflix user will click on the suggested next title, as represented by the output class "**1**".

Model A	Model B
$\boldsymbol{Accuracy} = \ \mathbf{95.1}\%$	$\boldsymbol{Acuracy} = \ \mathbf{87.1}\%$
$\boldsymbol{Sensitivity} = \mathbf{13.8}\ \%$	$\boldsymbol{Sensitivity} = \mathbf{91.2}\%$
$\boldsymbol{Specificity} = \mathbf{98.3}\%$	$\boldsymbol{Specificity} = \mathbf{80.4}\%$

Assumption: "Model A is the most suitable to predict if an Netflix user will click on the suggested next title".

(i) Explain why someone would make this incorrect assumption, using the values presented in the table above to aid your answer.

(ii) Explain your reason why Model B is the most suitable model for predicting if an Netflix user will click on the suggested next title, using the values presented in the table above to aid your answer.

(6 marks)

(B) Ten-fold Cross Validation is often referred to as the "Gold Standard" of Machine Learning model validation techniques (the best technique to use).

(i) Explain what is the most important feature of this technique, that makes it the "Gold Standard"?

(ii) Explain an alternative to Ten-fold Cross Validation? Compare and contrast the two techniques (10-fold Cross Validation and the alternative technique), giving examples of problem situations where each technique may be more suitable.

(8 Marks)

(C) "Machine Learning Models can be biased, due to several factors".

(i) Discuss two factors that may cause a Machine Learning model to be biased.

(ii) Explain how you would investigate each factor (from part (i)) to identify if they may be contributing to model bias (you may assume that bias is present in the model).

(6 Marks)

Question 3

(A) "Logistic Regression is often the first Machine Learning algorithm used when developing a Machine Learning model".

(i) Compare and contrast the Logistic Regression Machine Learning algorithm with any other Machine Learning algorithm.

(7 Marks)

(B) "Sometimes the Logistic Regression Machine Learning algorithm is unsuitable for a problem context or needs specific data pre-processing like normalisation to improve performance."

(i) Describe what problem contexts are unsuitable for the Logistic Regression Machine Learning algorithm, explaining why the algorithm is unsuitable.

(ii) Explain why pre-processing techniques such as normalisation, can significantly improve the performance for the Logistic Regression Machine Learning algorithm.

(7 Marks)

(C) "Unsupervised learning (often called self-learning) can be useful for specific problem contexts."

(i) Describe a problem context where you would use an Unsupervised Machine Learning algorithm, detailing a suitable algorithm and how it can address the problem.

(6 Marks)

Question 4

(A) Machine Learning Ensembles, have become a popular option in Machine Learning.

(i) Explain what aspects/features of a Machine Learning Ensemble, might allow it to outperform a traditional Machine Learning Algorithm, giving reasons why.

(ii) Discuss any issues that might arise when using Machine Learning Ensembles.

(8 Mark)

(B) "Bootstrap is a statistical estimation technique where a statistical quantity like a mean is estimated from multiple random samples of your data (with replacement)"

(i) Discuss this statement, explaining in your own words how the Bootstrap pre-processing technique works.

(ii) Provide an example of a problem situation where this technique should be considered, explaining why you think Bootstrap, is suitable for this problem situation.

(5 Marks)

(C) Statistical testing is often used to compare the performance of two or more machine learning models.

(i) Compare and contrast any two methods of statistical testing for comparing two or more Machine Learning models.

When reporting a statistical test result, many people do not present the entire picture, calling into question the findings.

(ii) Discuss what are the most important items/values of a statistical test to report, so that there is no ambiguity in the findings, explaining your reason for item/value selected.

(7 Marks)