RHODES UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

EXAMINATION: JUNE 2019

COMPUTER SCIENCE HONOURS
PAPER 2
MACHINE LEARNING

Internal Examiner: Mr. J Connan
Dr. D. Brown

MARKS: 120
DURATION: 4 hours

External Examiners: Prof. I. Sanders

GENERAL INSTRUCTIONS TO CANDIDATES

1. This paper consists of 8 questions and 33 pages. ***Please ensure that you have a complete paper.***
2. State any assumptions and show all workings.
3. Diagrams are encouraged and should be labelled.
4. Provide answers that are concise, legible and clearly numbered.
5. Use the mark allocation as a guide to the depth of your answer.
6. The Concise Oxford English Dictionary may be used during this examination.
7. You may use a calculator (though it should not be needed).

PLEASE DO NOT TURN OVER THIS PAGE UNTIL TOLD TO DO SO.

Section A Theory [60 Marks]

Question 1 **(8 marks)**
Define Machine Learning and give a simple example of machine learning that illustrates the terms you used to define it?

Question 2 **(8 marks)**
What is the difference between supervised and unsupervised machine learning? Provide a simple example of each and suitable classifier.

Question 3 **(4 + 4 = 8 marks)**
If one does not use a vectorised approach to Linear Regression, it may be necessary to use a learning rate as well as feature scaling.

a) Explain what each of these concepts are and why they may be needed to aid faster conversion.

b) If one were to graph the learning progress, what typical scenarios would one observe.

Question 4 **(8 marks)**
Explain what k-fold cross validation is and why one would use it.

Question 5 **(10 + 2 = 12 marks)**
Both Linear and Logistic Regression follows very similar processes, however, one generally use a very different model $h_\theta(x)$ and cost function $J(\theta)$ for Logistic Regression.

a) What are the functions that one typically uses for Logistic Regression and why does one use them.

b) Why is the square of the difference not a good measure of error for this model?

Question 6 **(16 marks)**
Accuracy, Precision and Recall, and F1 score are measure that can be used to evaluate model classification. Explain each of these terms and provide and example to illustrate how they are applied.

Section B Practical [60 Marks]

Question 7 **(3 + 3 + 6 + 8 = 20 marks)**

a) Import dataset_1 and split the training and test set to 80:20.

b) Perform min-max scaling on the dataset.

c) Fit a logistic regression model for classification and calculate the accuracy score.

d) Replace the min-max scaling method with a method that uses interquartile range, so that it is robust to outliers – resulting in an improved accuracy score.

Question 8 **(5 + 11 + 14 +10 = 40 marks)**

Locate all relevant resources for this question in the directory question_8. The python file Q8.py imports training and test datasets containing multiple classes.

a) Correct any runtime errors (if any) in Q8.py, and split the training and test set to 50:50.

b) Fit SVM models that use the linear, polynomial, gaussian and sigmoid kernels to allow for parameter estimation using grid search with cross-validation.

c) Perform parameter estimation using grid search with cross-validation for best precision. The results should be formatted similar to this, but repeated for each kernel:

```
0.986 (+/-0.016) for {'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
0.959 (+/-0.029) for {'C': 1, 'gamma': 0.0001, 'kernel':
'rbf'}
0.988 (+/-0.017) for {'C': 10, 'gamma': 0.001, 'kernel':
'rbf'}
```

d) Visualize the results using a boxplot that displays the best precision per kernel method.

END OF EXAMINATION