FINAL

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2017

BEng Honours Degree in Computing Part III
MEng Honours Degree in Electronic and Information Engineering Part IV
MEng Honours Degree in Mathematics and Computer Science Part IV
BEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degrees in Computing Part III
MSc in Computing Science
MSc in Computing Science (Specialist)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER C395

MACHINE LEARNING

Tuesday 21 March 2017, 10:00
Duration: 90 minutes

Answer TWO questions

Paper contains 3 questions
Calculators required

1. Consider the following set of positive (+) and negative (-) training examples:

sky	air	humid	wind	water	forecast	Enjoy Sport
sunny	warm	normal	strong	warm	same	+
sunny	warm	high	strong	warm	same	+
rainy	cold	high	strong	warm	change	-
sunny	warm	high	strong	cool	change	+
sunny	warm	normal	weak	warm	same	-
rainy	warm	high	weak	warm	change	-
rainy	warm	normal	strong	cool	same	-

a) Apply the ID3 algorithm. Write out the intermediate and the final results. Draw the final result as a diagram.

b) Apply the distance-weighted k-Nearest Neighbour algorithm, k=2, to classify the instance <sunny, cold, high, strong, cool, same>, assuming that the above-listed examples are already known. Write out the algorithm, the distance function, the weight function, and the intermediate results.

c) Suppose that we want to solve the problem of when to enjoy sports weather-wise by using genetic algorithms. Suppose further that the solution to the problem can be represented by the result of the ID3 algorithm in 1(a). What is the appropriate chromosome design for the given problem? Which Genetic Algorithm parameters need to be defined? What is the result of applying a single round of the prototypical Genetic Algorithm? Explain your answer in a clear and compact manner by providing the pseudo code of the algorithm and writing down all intermediate results.

The three parts carry, respectively, 25%, 30%, 45% of the marks.

2.

a) Derive the gradient descent training rule assuming that the target function representation is:

$o_d = w_0 + w_1x_1 + w_1x_1^3 + \ldots + w_nx_n + w_nx_n^3$.

Define explicitly the cost/error function E, assuming that a set of training examples D is provided, where each training example $d \in D$ is associated with the target output t_d.

b) Consider the instance space consisting of integer points in the x, y plane, where $0 \leq x, y \leq 10$, and the set of hypothesis consisting of rectangles (i.e., being of the form $(a \leq x \leq b, c \leq y \leq d)$, where $0 \leq a, b, c, d \leq 10$).

What is the smallest number of training examples one needs to provide so that the CANDIDATE-ELIMINATION algorithm perfectly learns a particular target concept (e.g., $(2 \leq x \leq 4, 6 \leq y \leq 9)$)? Explain your answer in a clear manner (i.e., explain when can we say that the target concept is exactly learned in the case of the CANDIDATE-ELIMINATION algorithm and what is the optimal query strategy).

c) What is the difference between the Best First Search and the Beam Search algorithms?

d) Which types of knowledge can we distinguish in Case-Based Reasoning (CBR)? Provide a short explanation of each of the types.

The four parts carry, respectively, 30%, 25%, 25%, 20% of the marks.

3.

a) Derive the update rule for the weights in the output layer of a neural network using gradient descent rule. Assume that the sigmoid function is used as an activation function, the quadratic loss as the error function and L1 regularisation is applied.

b) Assume the network's error function is E_0. How is it modified when L2 regularisation is applied? Describe how this type of regularization works and what is the difference with L1 regularisation.

c) Assume that you wish to train a classifier on a large dataset. How would you estimate its generalization performance and optimize its parameters? Describe briefly the procedure that you would follow.

d) Compute the classification rate for the given confusion matrix. Do you think the classification rate is a suitable performance measure in this case? Explain your reasoning and the alternatives.

	Class 1 - Predicted	*Class 2 - Predicted*	*Class 3 - Predicted*
Class 1 - Actual	*1000*	*100*	*50*
Class 2 - Actual	*20*	*0*	*10*
Class 3 - Actual	*10*	*10*	*0*

The four parts carry, respectively, 40%, 20%, 20%, 20% of the marks.