NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2017/2018

Exam Code(s)	1CSD, 1SPE, 2SPE
Exam(s)	MSc in Computer Science (Data Analytics) 1st and 2nd Structured PhD
Module Code(s)	CT5107
Module(s)	Advanced Topics in Machine Learning and Information Retrieval
Paper No.	1
External Examiner(s)	Prof. Pier Luca Lanzi
Internal Examiner(s)	* Prof. Michael Madden * Dr Colm O'Riordan

<u>**Instructions:**</u>	Answer 2 questions from each section (4 in total). All questions carry equal marks.

Duration	2 hours
No. of Pages	4
Discipline(s)	Information Technology
Course Co-ordinator(s)	Dr Conor Hayes (CSD)

Requirements:

Release in Exam Venue	Yes	□	No	☒
MCQ Answersheet	Yes	□	No	☒
Handout	None			
Statistical/ Log Tables	None			
Cambridge Tables	None			
Graph Paper	None			
Log Graph Paper	None			
Other Materials	None			
Graphic material in colour	Yes	□	No	☒

PTO

Section A: Machine Learning

1. (a) A biomedical engineer has sent you the following message. Prepare a detailed reply. *We have created a set of different hip implant designs, built prototypes of each one, recorded data about many different measurements and characteristics of each one, and tested them in the lab. Some have failed testing and others have passed. Now I am trying to use some machine learning algorithms (Naïve Bayes, kNN and logistic regression) to classify whether designs would pass or fail testing, based on the data we recorded the results. Somebody has suggested that I should try feature extraction, feature transformation, or feature subset selection. In general, do you think it will be worthwhile for me to spend time on this? Can you please explain each of these concepts, and explain for each one why it could help? For each one, can you recommend at least one algorithm/approach?* [10]

(b) Considering an example of a fully-connected feed-forward neural network with 2 input nodes, 3 hidden nodes and one output node, draw a diagram showing all nodes and weights, with an appropriate notation. Write down equations for how the inputs are propagated forward to produce the output. [4]

(c) Making reference to the example neural network form Part (b), explain in detail the Backpropagation algorithm, with the relevant equations as well as narrative. [6]

(d) The connectionist programming perspective is that deep learning provides an integrated approach to feature engineering and learning. Explain this, referring back to Part (a). [5]

2. (a) Explain what an auto-encoding neural network is, illustrating your answer with a diagram. Building on this, describe in detail the principle of pre-training a deep neural network using stacked auto-encoders. In addition, identify which problem(s) encountered in classic shallow neural networks are addressed by this approach. [8]

(b) Explain what a locally connected network architecture is. Building on this, describe the architecture of a convolutional neural network, illustrating your answer with an example. In addition, identify which problem(s) encountered in classic shallow neural networks are addressed by this approach. [8]

(c) Based on your experience, what would you consider to be the **two** most significant challenges in implementing a neural network? Explain your answer. [4]

(d) One issue with convolutional neural networks is that they can be "fooled". Discuss, with reference to some examples, what fooling means in this context. Why does this happen? What, if any, are the associated risks/problems? [5]

[PTO]

3. (a) In the context of Linear Support Vector Machines, explain with a diagram what the support vectors are and what the maximum margin is. Is there more than one solution for the maximum margin hyperplane? What are the consequences of this for finding a solution? [7]

(b) In a Linear Support Vector Machine, the decision function can be expressed in either the primal or dual form. Provide equations for both forms of the decision function, explaining all terms in the equations. How to the two forms relate to each other? [8]

(c) *"Bagging and Stacking are both forms of ensemble classifier."* Discuss this, including:

(i) an explanation of what ensemble classification is;
(ii) an simple illustration of how ensembles improve performance;
(iii) information about what conditions/requirements are needed for this;
(iv) an explanation, with pseudocode or diagrams, of how to generate models with Bagging;
(v) an explanation, with pseudocode or diagrams, of how to generate models with Stacking. [10]

Section B

4. (a) Many search algorithms balance *exploitation* and *exploration*. Explain these concepts in relation to genetic algorithm operators (selection, crossover and mutation). [6]

(b) The travelling sales person problem involves finding a tour through N cities such that the distance is minimised. The tour begins and ends at some specified city. Explain how a genetic algorithm could be used to generate tours given the N cities and the distances between each pair of cities. Your answer should explain your solution representation, the fitness function and the operators. [10]

(c) Term weighting schemes are fundamental to the performance of an IR system. Outline an approach using evolutionary computation to evolve weighting schemes for information retrieval. Discuss the advantages and limitations of this approach. [9]

5. Traditional Information Retrieval systems have typically adopted a bag of words model which incorporate the term-independence assumption. More recently, graph models have been used to represent queries and documents. With reference to existing systems, answer the following:

(a) Discuss graph properties that may capture features of text which may help in a retrieval context. [8]

(b) Explain how one could use these graph properties to develop, or augment existing, weighting schemes. [9]

(c) Explain how sources of evidences external to the document corpus can be incorporated into the graph model. [8]

6.

(a) In distributed information retrieval systems, collections are distributed across sites and a set of heterogeneous information retrieval models may be employed across the sites. Discuss the issues involved in merging results from such a distributed information retrieval system. Your answer should include an overview of source selection and result fusion approaches. [10]

(b) In many modern information systems, the following properties hold: distributed collections, many information providers and information seekers, costs and benefits associated with the retrieval (e.g. quality, relevance, and cost). In these domains, there is often a conflict of interests leading to an incentive for a participant to behave unfairly in order to exploit other participants. Choosing a suitable domain example, discuss an approach to represent and reason about these conflicts.
[15]