NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2015/2016

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
Repeat Paper	No
External Examiner(s)	Prof. Liam Maguire
Internal Examiner(s)	Dr James Duggan *Dr Michael Madden *Dr Colm O'Riordan
Instructions:	Answer 2 questions from each section (4 in total). All questions carry equal marks.
Duration	2 hours
No. of Pages	3
Discipline(s)	Information Technology
Course Co-ordinator(s)	Dr Conor Hayes (CSD)

Requirements:

MCQ	Release to Library: Yes
Handout	None
Statistical/ Log Tables	None
Cambridge Tables	None
Graph Paper	None
Log Graph Paper	None
Other Materials	None
Graphic material in colour	No

PTO

Section A: Machine Learning

1. (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 one? [8]

(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 do the two forms relate to each other? [8]

(c) In the context of Linear Support Vector Machines, explain what the support vectors are. Also, what is the maximum margin and what is the decision function? [5]

(d) Describe the "kernel trick" as used in non-linear Support Vector Machines. [4]

2. (a) You have received the following email from a geneticist called Rosalind. Prepare a reply.
"I am trying to analyse genomics data. I have been advised to use an ensemble classifier, because they have good theoretical properties. Can you please explain what these properties are? Also, how can I ensure that members of the ensemble are better than random and diverse? Two methods that have been mentioned are 'bagging' and 'boosting'. Can you please explain how either one of these operates?" [10]

(b) Feature engineering is often recognised as a key import step in practical machine learning applications. Discuss this, making reference to feature transformation and different approaches to feature subset selection. [6]

(c) 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. Write down equations for how the inputs are propagated forward to produce the output. [4]

(d) Is the cost function used in training feed-forward neural networks convex? What are the implications of this? Explain how a training curve may be used to monitor training progress. [5]

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

(b) What is ReLU, and how does it represent an improvement over what is used in classic neural networks? [4]

(c) Explain what a locally connected network architecture is. Building on this, describe the architecture of a convolutional neural network. As part of your answer, identify which problem(s) encountered in classic shallow neural networks are addressed by this approach. [9]

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

Section B: Information Retrieval

4. 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 the text which may help in a retrieval context. [8]

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

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

5. (a) What is meant by topic modelling? Describe, with a short example, an approach to topic modelling. [10]

(b) The term semantic matching has been used to describe a broad range of techniques that attempt to calculate similarity between documents and queries based on some understanding of the document and the query and not just at term matching level. These approaches include among others, *query reformulation*, *similar query mining*, *analysis of query term dependencies*.

Discuss *any* two approaches to semantic matching. Note: you do not have to limit yourself to the topics listed as examples. [15]

6. (a) In distributed information retrieval systems, collections are distributed 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. [11]

(b) Discuss a suitable approach to developing and maintaining indexes in peer-to-peer information retrieval systems. [7]

(c) In information retrieval systems with a set of distributed, heterogeneous sources there are often questions that arise about the various benefits and costs of including certain sources (e.g. cost, quality, trust etc.). Suggest an approach to represent and reason about these issues. [7]