The University of Melbourne

Department of Computer Science and Software Engineering

433–327 (COMP30018) Knowledge Technologies November 2010

Identical examination papers: 433–693 (COMP90029), 433–694 (COMP90005)

Exam duration: Two hours

Reading time: Fifteen minutes

Length: This paper has 8 pages including this cover page.

Authorised materials: None

Calculators: Not permitted

Instructions to invigilators: Students may not remove any part of the examination paper from the examination room. Students should be supplied with the exam paper and a script book, and with additional script books on request.

Instructions to students: This exam is worth a total of 50 marks and counts for 60% of your final grade. Please answer all questions in the script book provided, starting each question on a new page. Please write your student ID in the space below and also on the front of each script book you use. When you are finished, place the exam paper inside the front cover of the script book.

Library: This paper is to be held in the Baillieu Library.

Student id:

Examiner's use only:

Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*

Continued overleaf ...

433-327 (COMP30018) Knowledge Technologies
Final Exam

Semester 2, 2010

Total marks: 50

Students must attempt all questions

Section A: Short Answer Questions [13 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each sub-question in a couple of lines.

Question 1: Basics [7 marks]

1. Categorise the following methods as "supervised" or "unsupervised": [1.5 marks]
   - multivariate binomial naive Bayes
   - mutual information-based feature selection
   - k-means
2. For the "regular expression":

   ```
   ^know(ledge|-how)+$
   ```

   which of the following strings would the expression match? [2 marks]

   (a) `knowledge`
   (b) `know`
   (c) `knowledge-how`
   (d) `unknowledgable`
3. Given a training dataset with a total of 25 instances, made up of 10 instances labelled as `A`, 5 instances labelled as `B`, and 10 instances labelled as `C`, what would the "error rate" be for the "ZeroR" algorithm? [2 marks]
4. It has been claimed that there are three primary types of "information need" in a web search context: "informational", "navigational" and "transactional". Briefly describe each, optionally with the aid of an example. [1.5 marks]

Continued overleaf ...

Question 2: Data Representation [6 marks]

1. Fill in the gaps in the following statement:

   Standardly in machine learning, [(1)] are represented as vectors of [(2)], and each of which is labelled with one or more [(3)].

   selecting from the following concepts: [1.5 marks]

   (a) instances
   (b) classes
   (c) classifiers
   (d) features
   (e) attributes

2. What is a "code point" and "code unit" in the context of Unicode-based "character encoding"? [3 marks]

3. In the context of information retrieval, what is a "token"? [1.5 marks]

Continued overleaf ...

Section B: Method Questions [14 marks]

In this section you are asked to demonstrate your conceptual understanding of a subset of the methods that we have studied in this subject.

Question 3: Text Categorisation [4 marks]

With the use of an example, outline the "k nearest neighbour" (k-NN) approach to text categorisation, including a description of how the neighbours are calculated, and how the final classification is determined. [4 marks]

Question 4: Web Retrieval [4.5 marks]

In general "information retrieval" settings, factors that influence the score assigned to documents in a "ranked query" include the relative frequency of each query term across the collection; the relative frequency of each query term in the document; and the length of the document.

In a "web search" setting, other factors can be used. List *three* (3) web-specific factors, and for each explain how it provides evidence of relevance to the query, and how it might be incorporated into the similarity computation. [4.5 marks]

Question 5: Data Sampling and Evaluation [5.5 marks]

1. Outline each step you would go through in calculating classification accuracy based on "10-fold stratified cross-validation" for a given dataset and supervised learner. [3.5 marks]

2. Why is "stratified cross-validation" generally considered superior to the "holdout" method? [2 marks]

Continued overleaf ...

Section C: Numeric Questions [15 marks]

In this section you are asked to demonstrate your understanding of a subset of the methods that we have studied in this subject, in being able to perform numeric calculations.

Question 6: Decision Tree Induction [6 marks]

Answer the following questions relative to the following training dataset (made up of instances A, B, C and D):

	$Feature_1$	$Feature_2$	*Class*
A:	`sunny`	`cool`	`no`
B:	`sunny`	`hot`	`yes`
C:	`rainy`	`mild`	`yes`
D:	`rainy`	`mild`	`no`

1. Calculate the "entropy" of the class distribution in the dataset. [1 mark]
2. Calculate the "mean information" and "information gain" for a "decision stump" based on: (i) $Feature_1$, and (ii) $Feature_2$. [3 marks]

   Recall that "information gain" is calculated by:

$$IG(R_A|R) = H(R) - \sum_{i=1}^{m} P(x_i)H(x_i)$$

3. Which of these two decision stumps would be preferred in the "ID3" algorithm and why? [1 mark]
4. Classify the following "test instance" relative to the preferred decision stump in your previous answer: [1 mark]

	$Feature_1$	$Feature_2$
E:	`rainy`	`hot`

Continued overleaf ...

Question 7: Document Categorisation [5 marks]

Based on the following set of labelled "training documents":

Terms		Class
mouse	*disk*	
2	2	`computer`
5	0	`zoology`
1	0	`zoology`
1	1	`zoology`
1	3	`computer`

and the following "test document":

Terms		Class
mouse	*disk*	
2	0	?

what is the prediction of the most probable class label for the test document according to the "multinomial naive Bayes" method? Show all working. [5 marks]

Recall that the following equations form the basis of the "multinomial naive Bayes" calculation:

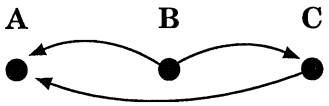
$$P(D|c_i) = \prod_{j=1}^{|\mathscr{V}|} \frac{P(t_j|c_i)^{N_{D,t_j}}}{N_{D,t_j}!}$$

and

$$P(t|c_i) = \frac{1 + \sum_{k=1}^{|\mathscr{D}|} N_{k,t} P(c_i|D_k)}{|\mathscr{V}| + \sum_{j=1}^{|\mathscr{V}|} \sum_{k=1}^{|\mathscr{D}|} N_{k,t_j} P(c_i|D_k)}$$

Question 8: PageRank [4 marks]

Apply the "PageRank" algorithm ($\alpha = 0.5$) to the following web document set (directed edges indicate hyperlinks) over *one* (1) iteration (not including the initialisation of document probabilities), clearly showing all working. [4 marks]



Continued overleaf ...

Section D: Design and Application Questions [8 marks]

In this section you are asked to demonstrate that you have gained a high-level understanding of the methods and algorithms covered in this subject, and can apply that understanding. Expect your answer to each question to be from one third of a page to one full page in length. These questions will require significantly more thought than those in Sections A–C and should be attempted only after having completed the earlier sections.

Question 9: Recommender Systems and Association Rule Mining [8 marks]

A "recommender system" is a system which attempts to determine items which would be of interest to a particular user, based on analysis of their past behaviour. Perhaps the best-known recommender system on the web is Amazon (`amazon.com`), which, e.g., generates "Recommendations for You" based on analysis of items previously purchased by you, and those previously purchased by other individuals.

Separately, "association rule mining" is the task of extracting itemsets which predict other itemsets, from a set of transactions, generally using the notions of "support" and "confidence" as a guide.

1. Write a short paragraph identifying two (2) key differences between "association rule mining" as it is standardly performed, and "recommender systems". [4 marks]

2. If you were to apply "association rule mining" to the task of Amazon-style "item recommendation", how would you use "support" and "confidence" (and any other statistics you may be able to infer), given the basic task of identifying and ranking items which a user is likely to purchase, on the basis of the current contents of their shopping cart? [4 marks]

— End of Exam —