The University of Melbourne, School of Computing & Information Systems

COMP90049 Knowledge Technologies

Final exam, Semester 2, 2017

Date: 15 November, 2017

Time: 12:15pm

Reading Time allowed: 15 minutes

Writing Time allowed: 2 hours

Number of pages: 7 including this page

Instructions to candidates:

This paper counts for 50% of your final grade.
Answer all questions on the ruled pages in the script book(s) provided, unless otherwise indicated.
There are 76 marks in total, or 1 mark per 1.6 minutes. Note that questions are not of equal value. All questions should be interpretted as referring to concepts given in this subject, whether or not it is explicitly stated.

No external materials may be used for this exam, but calculators are permitted (although not necessary). You may leave square roots and logarithms without integer solutions (like $\sqrt{2}$) unsimplified.

Unless otherwise indicated, you must show your working for each problem. Please indicate your final answers clearly for problems where you show intermediate steps.

Instructions to invigilators:

The students require script books.
Calculators are permitted; other materials are not authorised.
The examination paper should not leave the examination hall; this exam is to be held on record in the Baillieu Library.

Examiner's use only:

Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*	*Q10*	*Q11*
2	4	13	6	9	4	3	3	5	5	22

Part I : Text Processing [25 marks in total]

1. **Regular Expressions:** Given the following "regular expression":

0或很多

```
\s*f?re{1,3}\w
```

For each of the following strings, indicate (with "yes" or "no" in your script book) whether the regular expression will match. [2 marks]

(a) `freedom` Yes.

(b) `a really good book` Yes.

(c) `sfrew` Yes.

(d) `\tf\tfr\tfre\tfree` Yes.

2. **Approximate String Search:** By referring to the definitions from this subject:

(a) Explain the problem that we wish to solve in "approximate string search", including one typical strategy and any necessary data structure(s). [2 marks]

(b) Explain why "approximate string search" is a "Knowledge Technology" (or "knowledge task"). [2 marks]

a) Search with uncertainty - pattern matching spelling correction.
edit distance. Soundex.

b) The result of approximate string search depends on human knowledge.
correctness is not well-defined.

continued ...

it → 1(1)

happens → 1(1)

over → 1(3) → 2(1) → 3(~~2~~4)

and → 1(2) → 2(2)

again → 1(1)

under → 2(1)

in → 2(1)

out → 2(1)

	A	B	C
over	1	1	1
and	1	1	0
out	0	1	0

$(1,1,1) \wedge (1,1,0) \wedge (0,1,0) = (0,1,0)$.

$\therefore$ ~~⇒~~B

c). $w_{q,\text{over}} = \frac{3}{3} = 1$

$w_{q,\text{and}} = \frac{3}{2} = 1.5$

$w_{q,\text{out}} = \frac{3}{1} = 3$.

others will be 0

$\therefore w_{q,t} = \langle 0,0,1,1.5,0,0,0,3 \rangle$

$\text{doc}A = \langle 1,1,3,2,1,0,0,0 \rangle$

$\text{doc}B = \langle 0,0,1,2,0,1,1,1 \rangle$

$\text{doc}.C = \langle 0,0,4,0,0,0,0,0 \rangle$.

$$\cos(A,q) = \frac{\langle 1,1,3,2,1,0,0,0 \rangle \cdot \langle 0,0,1,1.5,0,0,0,3 \rangle}{\sqrt{1^2+1^2+3^2+2^2+1^2} \cdot \sqrt{1^2+1.5^2+3^2}}$$

$$= \frac{3+3+0}{\sqrt{16} \cdot \sqrt{12.25}} = \frac{3}{2\sqrt{12.25}} = \frac{1.5}{\sqrt{12.25}}$$

$$\cos(B,q) = \frac{1+3+3 \; 7}{\sqrt{8} \cdot \sqrt{12.25}}$$

$$= \frac{7}{2\sqrt{2}\sqrt{12.25}}$$

$$\cos(C,q) = \frac{4}{4 \cdot \sqrt{12.25}} = \frac{1}{\sqrt{12.25}}$$

B > A > C.

3. **Information Retrieval:** For this question, consider the (very small) collection of documents, labelled A), B), and C) below (the label is not part of the document text):

```
A) ~~it~~ happens ~~over~~ ~~and~~ ~~over~~ ~~and~~ ~~over~~ ~~again~~
B) ~~over~~ ~~and~~ ~~under~~ in ~~and~~ out
C) ~~over~~ ~~over~~ ~~over~~ ~~over~~
```

and a query Q) over and out

(a) For a standard inverted index consistent with the lecture or workshop notation; give a representation (in words or as a diagram) of the "inverted lists" for the 8 terms in this collection. (There is no need to explicitly indicate the "search structure" or "mapping table".) [4 marks]

(b) If we wish to apply the method of "Boolean querying" — assuming that the query is implicitly a conjunction of terms — describe the procedure by which the query engine would process the inverted index, to arrive at the result set {B}. (The and in the query is a term, not the Boolean operator AND.) [3 marks]

(c) Determine the document ranking for a "ranked query engine", based on the following "TF-IDF model", suitably interpretted in the context of this subject:

$$w_{d,t} = f_{d,t}$$
$$w_{q,t} = \frac{N}{f_t}$$

(Remember to show your work; there should be no need to simplify irrational square roots to solve this problem.) [6 marks]

4. The method of "accumulators" is often used when determining a document ranking in the context of Information Retrieval on the World Wide Web:

(a) What information is stored in an accumulator? How is this used to build a document ranking? [2 marks]

(b) The full accumulator method is typically not employed: explain why.

① too many documents ② Single document need a accumulator memory

[1 marks]

(c) In this subject, two heuristic strategies for simplifying the accumulator method were discussed. Choose **one**, and briefly explain it. [3 marks]

Threasholdly Limitelly.

a). a partial sum of dot product for the cosine TF-IDF value of each document as we process query term one by one.

continued . . .

Take a term → read off each document → update TF-IDF value to the accumulators

1 itemset: s(morning) = $\frac{1}{2} > \frac{1}{3}$ s(evening): $\frac{1}{2} > \frac{1}{3}$ s(coffee) = $\frac{1}{2} > \frac{1}{3}$

s(tea) = $\frac{1}{2} > \frac{1}{3}$ s(pasty) = $\frac{3}{4} > \frac{1}{3}$. s(roll) = $\frac{1}{4} < \frac{1}{3}$.

ignore {roll}.

2-itemset: ~~s(morning, coffee)~~ = $\frac{1}{4} < \frac{1}{3}$ ~~s(morning, tea)~~ = $\frac{1}{4} < \frac{1}{3}$, s(morning pasty) = $\frac{1}{2} > \frac{1}{3}$

~~s(evening, coffee)~~ = $\frac{1}{4} < \frac{1}{3}$ ~~s(evening, tea)~~ = $\frac{1}{4} < \frac{1}{3}$, ~~s(evening pasty)~~ = $\frac{1}{4}$

s(coffee pasty) = $\frac{1}{2} > \frac{1}{3}$, ~~s(tea, pasty)~~ = $\frac{1}{4} < \frac{1}{3}$

3-itemset ~~s(morning, coffee pasty)~~ = $\frac{1}{4} < \frac{1}{3}$.

{morning} → {pasty} $c = \frac{2}{2} = 1 > \frac{3}{4}$ ✓

{pasty} → {morning} $c = \frac{2}{3} < \frac{3}{4}$

{coffee} → {pasty} $c = 1 > \frac{3}{4}$ ✓.

{pasty} → {coffee} $c = \frac{2}{3} < \frac{3}{4}$

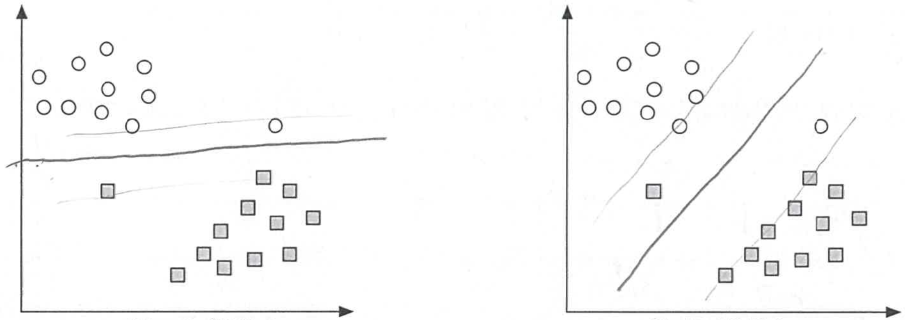
∴ {morning} → {pasty}

{coffee} → {pastly}.

Part II: Data Mining/Machine Learning [51 marks in total]

5. **Support Vector Machines (SVMs)**

   (a) **This question should be answered by drawing directly on the diagrams below.** Draw the **maximum margin hyperplane** that would be generated by a Support Vector Machine (SVM), based on the dataset visualised in the diagrams below. On the left, using a **hard margin** SVM; on the right, using a **soft margin** SVM. [2 marks]



Hard SVM

Soft SVM

   Yes, soft margin allows errors.. but maximum the margin.

   (b) Do we get different support vectors from the hard margin and soft margin SVMs on the dataset in the graphs above? Why? [2 marks]

   (c) State two characteristics of the dataset above, which would lead us to believe that it is practical to learn an SVM here. [2 marks]

   binary classification

   linearly separability

   separability.

   (d) What are the main parameters in the "primal" formulation of an SVM? What are the main parameters in the "dual" formulation of an SVM, and how do they relate to "support vectors"? [3 marks]

6. **Association Rule Mining:** [4 marks]

TIME	DRINK	BAKED GOOD
morning	coffee	pastry
morning	tea	pastry
evening	coffee	pastry
evening	tea	~~roll~~

Given the instances above, apply the "*a priori* principle" to find all of the Association Rules with at least $\frac{1}{3}$ Support and at least $\frac{3}{4}$ Confidence.

$I(\text{parent}) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}$

$= \frac{1}{2}+\frac{1}{2} = 1.$

For a1, T has two 1, two 0.

F has two 1, two 0.

For I, $E(T) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$E(F) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

Information Gain $= 1 - \frac{4}{8}\times 1 - \frac{4}{8}\times 1 = 0.$

For a2, T, four 1, four 0

$E(T) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1$

Information Gain $= 1 - \frac{8}{8}\times 1 - 0 = 0.$

For a3. For T, Three 1, One 0.

For F, Three 0, One 1

$E(T) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{3}{4}$

$E(F) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{3}{4}$

$\therefore IG = 1 - \frac{1}{2}\times E(T) - \frac{1}{2}E(F) > 0.$

$\therefore$ a3 should be the root.

7. **Hierarchical Clustering:** Given the proximity matrix below (in terms of **distance**) for five instances $p_1 \ldots p_5$, build a hierarchical clustering by applying the method of agglomerative clustering, using the **complete link** updating heuristic. Be sure to show the proximity matrix at each step of the clustering, and the resulting **dendrogram**. [3 marks]

	p_1	p_2	p_3	p_4	p_5
p_1	0	1	5	9	10
p_2	1	0	3.5	8	7
p_3	5	3.5	0	3	4
p_4	9	8	3	0	0.5
p_5	10	7	4	0.5	0

8. **Naive Bayes:** What is the primary assumption of the Naive Bayes Classifier? Explain why it is necessary, and why it is often untrue. [3 marks]

attributes are conditionally independent. to make the problem tractable

$P(C_i|x) = P(c_i)\prod P(x_i|c_i)$

violate the condition

9. **Decision Tree:** Given the dataset below, with 8 instances, 3 binary (T or F) attributes $a_1 \ldots a_3$, and the class Label:

a_1	a_2	a_3	Label
T	T	T	1
T	T	T	1
F	T	T	1
F	T	F	1
T	T	F	0
T	T	F	0
F	T	F	0
F	T	T	0

(a) Use the Information Gain to determine which attribute should be placed at the root of the Decision Tree which would be constructed based on this dataset. Be sure to show your work; the following formula might help: [3 marks]

$$\Delta = I(\text{parent}) - \sum_j \frac{N_j}{N} I(j)$$

(b) Does a decision tree exist, which can perfectly classify the given instances? If yes, draw that decision tree, otherwise, explain why not, by referring to the data. [2 marks]

NO ;

10. **Neural Networks**

What are steps involved in the "back propagation" algorithm for a neural network? You should explain the significance of the "learning rate parameter" in your answer. [5 marks]

11. **Application Question (Long Response)** **[22 marks in total]**

Imagine that, after completing Knowledge Technologies and graduating from the University of Melbourne, you are hired as a data scientist. For your first project, you are given a dataset from the university's library, and your job is to build a *classification model*, as well as a *recommendation system*. As shown in Table 1, the dataset includes the list of books available in the library (columns) and the students who borrowed them (rows), and the ranking for each item (ranking value is between 0–5, 0 if the book was not borrowed and 1–5 indicates the student's interest). The metadata for the books (e.g. titles) are not readily available to us, we just have the book IDs (e.g. Book #i). The dataset also includes the students' field of study (in total there are 10 fields), which can be used for the classification task. Answer the following questions, considering that there are 500,000 students and 100,000 books in this dataset.

Table 1: Library data set.

Student ID	Book #1	···	Book #100,000	Label (Field of Study)
Student # 1	3	···	2	Computer science
Student # 2	5	···	0	Biology
⋮	⋮	⋮	⋮	⋮
Student # 500,000	1	···	4	Mathematic

[Question continues on the following page.]

continued ...

Task1: Classification Model — The aim of this task is to use the label information (the most right column in the table) and build a classification model that classify students based on the list of books that they borrowed.

(a) Consider the following supervised machine learning methods, and for each one, explain why it would be appropriate or inappropriate to use for this problem:

 i. Support Vector Machines multi-class. 100k dimension. training is expensive. [2 marks]

 ii. *k*-Nearest Neighbour two large dimension. all points seem to be more and less at equal distance [2 marks]

 iii. Neural Networks [2 marks]

 iv. Naive Bayes robust enough to predict [2 marks]

(b) Would "feature selection" be useful here? Explain why, by referring to a single machine learning method. [3 marks]

(c) Explain how you would evaluate the effectiveness of your system: you should briefly describe an evaluation strategy and an evaluation metric that are suitable for this data. What might be an example of a baseline? ~~Acc~~ Accuracy. shuffling [4 marks] 10-fold-cross. Validation

Task2: Recommendation System — Now assume that we do not have the label information and we would like to build a recommendation system for the library.

(e) In this subject we studied content-based and collaborative filtering methods. Which recommendation method and similarity measure would you choose to address this problem? Justify your answer. [3 marks]

(f) What are the challenges of building a recommendation system for datasets such as the library dataset? [2 marks]

(g) As a recommendation system designer what goals do you need to consider? [2 marks]