THE UNIVERSITY OF WARWICK
MSc Examinations: Summer 2018
CS9091 Data Mining

Time allowed: 2 hours.
Answer **FOUR** questions.
Read carefully the instructions on the answer book and make sure that the particulars required are entered on **each** answer book.
Approved calculators are allowed.

---

1. a) Explain what you understand by the terms i) concept learning and ii) hypothesis space. [4]

b) Explain how the hypothesis representation used in concept learning is biased. How is this justified? Give an example of biased and unbiased hypothesis representations and compare the size of their respective hypothesis spaces. [6]

c) Use the Candidate Elimination algorithm to show how the Version Space changes from the initial most specific boundary S = {<ø, ø, ø, ø, ø, ø>} and most general boundary G = <?, ?, ?, ?, ?, ?> as each of the instances below is presented. Justify your answers.

i.

Instance	Sky	Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes

[2]

ii.

Instance	Sky	Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
2	Sunny	Warm	High	Strong	Warm	Same	No

[2]

Continued

iii.

Instance	Sky	Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
3	Rainy	Cold	High	Strong	Warm	Change	No

[2]

iv.

Instance	Sky	Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
4	Sunny	Warm	High	Strong	Cool	Change	Yes

[5]

d) What conclusion(s) would you draw about the set of instances from the final Version Space?

[4]

---

2. a) Outline Hunt's algorithm for generating a decision tree from a set of training instances D, each instance having a class label that is a member of a set of classes C. [3]

b) Write down an expression for the entropy $E(p,n)$ of a dataset D that has p instances of class P and n instances of class N. [2]

c) Write down the expression for calculating the Information Gain of splitting dataset D on attribute A, which has v values. [4]

d) Given the dataset D in Table 1, use Information Gain to decide which attribute from Post Code, Gender, Car Type and Tax to choose as the root node. [6]

Post Code	Gender	Car Type	Tax	Class
1	M	Family	Low	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	High	C0
5	M	Sports	High	C0
6	F	Sports	Low	C0
7	F	Sports	Medium	C0
8	F	Luxury	High	C0
9	M	Family	High	C1
10	M	Family	Medium	C1
11	M	Luxury	High	C1
12	F	Luxury	Low	C1
13	F	Luxury	Low	C1
14	F	Luxury	Medium	C1
15	F	Luxury	Medium	C1
16	F	Luxury	High	C1

Table 1: Dataset.

e) i. Based on the result of part d), comment on the suitability of using Information Gain to choose the attribute to split on. [3]

ii. Define Gain Ratio and use it to show how you would overcome the problem you identified in i. [7]

Continued

3. Classification, Linear perceptron and Support Vector Machines.

   (a) Explain the intuition behind the linear perceptron and outline its algorithm. [5]

   (b) What are the advantages of Support Vector Machines with respect to the linear perceptron? [3]

   (c) Why do we need kernel functions with Support Vector Machines and what do they achieve? Name two commonly used kernel functions. [3]

   (d) What is the C parameter in Support Vector Machines, and what is the effect of choosing large or small values for this parameter? [3]

   (e) How can we use binary Support Vector Machines to tackle a multiclass classification task with k classes? [4]

   (f) Consider we have run a classifier on a dataset with 5 classes. Given the confusion matrix in Table 2 with the results of our classifier.

		Actual class				
		C1	**C2**	**C3**	**C4**	**C5**
Predicted class	**C1**	**5184**	294	72	14	521
	C2	403	**1499**	7	8	80
	C3	74	4	**3027**	7	418
	C4	120	22	39	**118**	98
	C5	1142	101	662	49	**1087**

Table 2: Confusion matrix

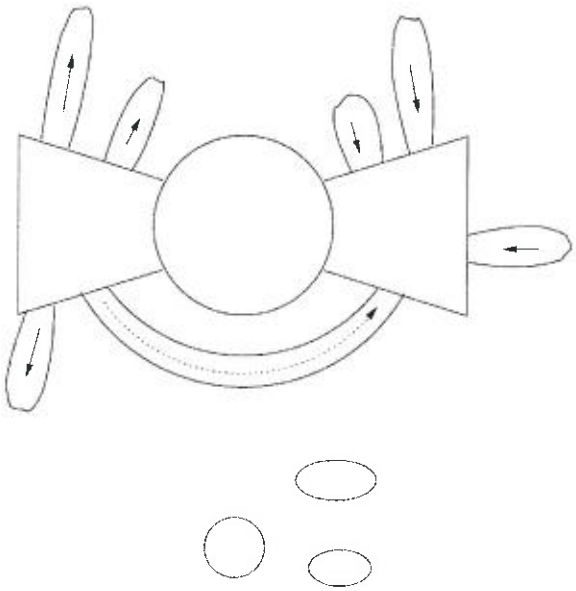
Compute the macroaveraged precision, recall and F1 scores, and microaveraged precision, recall and F1 scores. [7]

Continued

4. (a) PageRank and HITS.

   i. What is the intuition of the PageRank algorithm and how does it work? [5]

   ii. Given the "bowtie" representation of the Web below, label and define each of the components. [3]



   iii. What are dead ends and spider traps? Describe a solution to deal with spider traps using PageRank. [5]

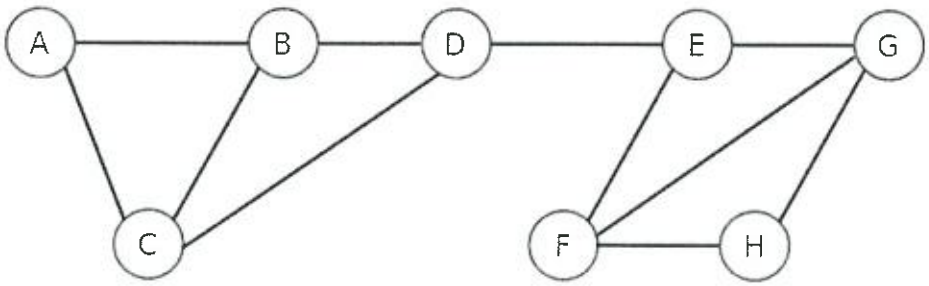
   iv. What is the intuition of the HITS algorithm and how does it work? [6]

   (b) What are spam farms and how do spammers use them? [6]

Continued

5. Big data mining and social network analysis.

(a) Given the following social network, calculate betweenness scores for all the edges. Using those betweenness scores, explain how you break down the network into two communities. [5]



(b) What is Dijkstra's algorithm used for in social network analysis? Explain and outline the algorithm. [5]

(c) Given the following 9 baskets, explain how you would use the A Priori algorithm to identify frequent itemsets (of 2 items) that satisfy a support S = 3, step by step. [4]

- $B_1 = \{a, c, b, f\}$
- $B_2 = \{a, e, d\}$
- $B3 = \{a, b, f\}$
- $B4 = \{c, d, f\}$
- $B5 = \{a, e, b\}$
- $B6 = \{a, c, b, d, f\}$
- $B7 = \{c, b, d\}$
- $B8 = \{b, c, f\}$
- $B9 = \{b, f\}$

(d) For the frequent itemset problem above, let's assume you have 1 billion baskets with 1 million different products in them. We have tried the A Priori algorithm, but we've run out of memory in out computer. How would you go about computing the itemsets instead? Explain and outline the algorithm that you would use instead. [4]

(e) You are getting a data stream with all the "likes" that people make on Facebook. With each data entry, you get the country where the like was made, e.g. UK, USA, UK, UK, China, Germany, Brazil, Australia, UK,...

You want to be able to provide counts of observed frequencies within the last k hours, days or weeks, where k is variable, e.g. how many likes from the UK have we seen in the last 2 weeks? how many from China in the last 7 hours?

Explain and outline the algorithm you would use to achieve this, and explain how good an estimation you would be able to provide. [7]

Continued

6. (a) Given a set C <$c_1..c_k$> of mutually exclusive classes with prior probabilities $P(c_1)...P(c_k)$, dependent on attributes $a_1..a_n$, with values $v_1...v_n$, write down an expression for the conditional or posteriori probability of c_i, given instance x. State any assumptions you make. [5]

(b) Explain how you would use this expression to predict the class label for x. [4]

(c) Explain how Bayesian Belief Networks overcome the problem of conditional dependence of variable values. [4]

(d) Given the Bayesian Belief Network for a vehicle shown in Figure 1, calculate the following probabilities:

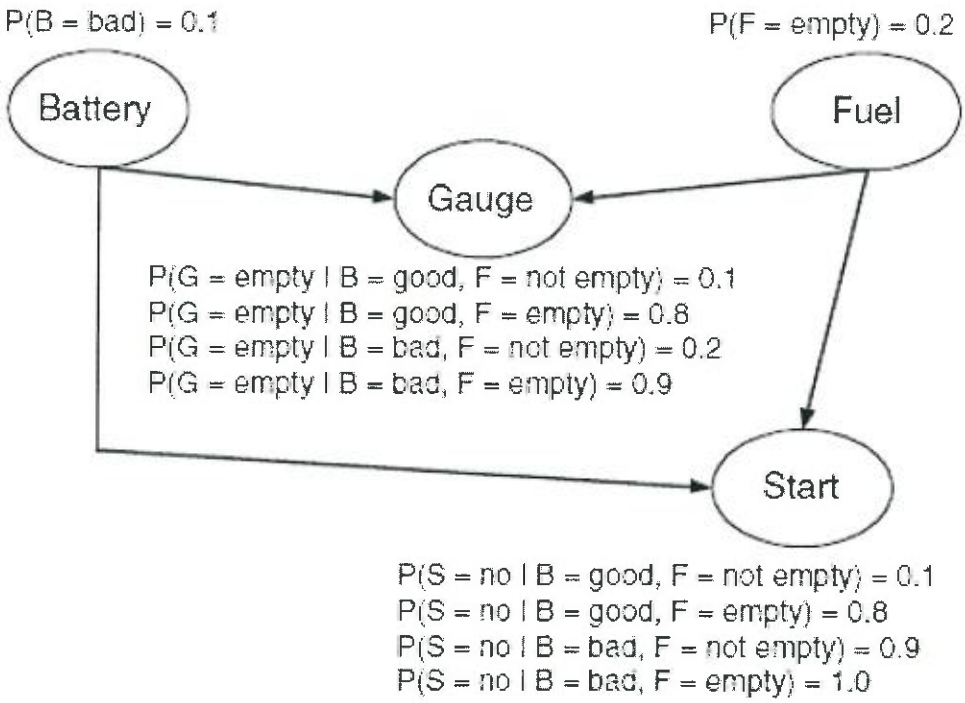


Figure 1: Bayesian Belief Network.

i. P(S = Yes|B = bad, F = empty) [1]

ii. P(B = good, F = empty, G = empty, S = yes). [2]

iii. P(B = bad, F = empty, G = not empty, S = no). [2]

iv. P(G = empty) [3]

v. Given that the Battery (B) is bad, calculate the probability the vehicle will start. [4]

Number	1/16	1/8	3/16	1/4	1/3	3/8	7/16	1/2	5/8	1
Log_2	-4.0	-3.0	-2.415	-2.0	-1.585	-1.415	-1.913	-1.0	-0.678	0

Common Log_2 values.

End