COMP 60711

Two hours

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Data Engineering

Date: Wednesday 16th January 2019

Time: 09:45 - 11:45

Please answer BOTH Questions.
Use a SEPARATE answerbook for each QUESTION

Each Question is worth 30 marks

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

[PTO]

Section A

1.

a) Can the adoption of the Data Lake Architecture improve data ingestion into the lake? If so, why or why not?

(5 marks)

b) Describe three of the main problems associated with the Data Transformation step in the Data Cleaning process, illustrating your answer with an example of data transformation.

(4 marks)

c) Describe the difficulties associated with the use of parallelism in the execution of a data profiling application for each of the following three cases: (1) when the application requires the use of distributive measures, (2) when it requires the use of algebraic measures and (3) when it requires the use of holistic measures.

(6 marks)

d) Explain why data collected via primary collection methods often need to be prepared before being analysed and give three examples of data preparation methods, explaining each.

(7 marks)

e) Explain data ingestion, emphasising the general causes and impact of data ingestion problems, and illustrating your answer with an example of a data ingestion problem.

(8 marks)

Section B

2.

a) In the context of classification:

i) Outline a decision tree classification algorithm; discuss how the attribute used at each node is chosen and what effect different training sets may have.

(3 marks)

(ii) Tests on a classifier give the following confusion matrix:

		Predicted		
		Disease=yes	Disease=no	Total
Actual	Disease=yes	90	210	300
	Disease=no	140	9560	9700
	Total	230	9770	10000

Calculate the accuracy, precision, recall, and specificity for the classifier based on this table.

(2 marks)

(iii)

a. Explain the potential effects of unbalanced data on the usefulness of a classifier. Give an example to support your answer.

(2 marks)

b. Discuss how ensembles methods combine models to increase classifier accuracy; compare bagging and boosting approaches.

(3 marks)

b) In the context of association rule (itemset) mining:

(i) Outline the working of the Apriori algorithm. Explain the importance of the subset property.

(4 marks)

(ii) Using Apriori, suppose that L4 is the list:

{ {p,q,r,s}, {p,q,r,t}, {p,q,r,z}, {p,q,s,z}, {p,r,s,z}, {q,r,s,z}, {r,s,w,x}, {r,s,w,z}, {r,t,v,x}, {r,t,v,z}, {r,t,x,z}, {r,v,x,y}, {r,v,x,z}, {r,v,y,z}, {r,x,y,z}, {t,v,x,z}, {v,x,y,z} }

a. At the join step of the algorithm, which itemsets are placed in C5 (the candidate set)?

(3 marks)

b. Which itemsets are then discarded by the prune step of the algorithm?

(3 marks)

Provide your working as appropriate.

c)

(i) Compare and contrast the use of row store and column store; discuss the potential performance advantages of their use for data warehousing and OLTP. Illustrate your answer.

(2 marks)

(ii) With the unprecedented scale of ‘big data’ and the emergence of vast numbers of data sources, discuss the role of data wrangling in future data engineering and science.

(3 marks)

(iii) Discuss the trade-offs between privacy of the individual on the one hand and the utility of data on the other. Use well-known case studies of data disclosure (AOL, Netflix, Barabasi) as appropriate to illustrate these trade-offs. How might we better guarantee that data is private?

(5 marks)

END OF EXAMINATION