HAND IN
Answers recorded
on question paper

Faculty of Arts and Science
School of Computing
CISC333 Examination
Instructor: D.B. Skillicorn

Winter 2005

Instructions:

1. You have three hours. This exam is worth 60 marks. You should spend about 3 minutes per mark.
2. Answer ALL questions in Section A. Answer any 2 questions in Section B. Answer 1 question in Section C.
3. You may bring in one 8.5x11 sheet of paper containing notes, and use it during this exam.
4. Answer questions on the exam paper. Nothing else will be marked. Use the amount of space under each question as a guide to the detail of your answer. You may use scrap paper and the back of the examination paper pages for rough working.
5. Proctors are unable to respond to queries about the interpretation of exam questions. Do your best to answer exam questions as written.
6. Make sure your student number is on every page of the exam paper.

STUDENT NUMBER: ______________________________

STUDENT NUMBER written in words: ______________________________

Section A: Short Answer Questions – answer all (1 mark each)

Answers should be brief enough to fit the space after the question.

1. What is a categorical attribute?

2. What is an ordinal attribute?

3. What is regression?

4. What is the systematic distortion of the predictions made by a predictor called?

5. What is an out-of-bag test set?

6. Write down Bayes rule and explain briefly what it means?

7. Why is early stopping potentially a problem in decision tree construction?

8. What is a threshold function in a neural network?

9. What is a good heuristic for the number of hidden nodes in a neural network?

10. What is a maximal margin linear separator?

11. What are support vectors?

12. What are the confidence and support of an association rule?

13. Explain how the confidence of an association rule can be derived from the frequencies of itemsets.

14. What are the strengths and weaknesses of rule-based models for prediction?

15. What is bagging?

16. What is overfitting?

17. Describe one method of attribute selection.

18. Explain one similarity measure, other than inverse Euclidean distance, that might be useful in clustering.

19. When would k-means not be an appropriate clustering algorithm?

20. Describe briefly the E and M steps in the Expectation-Maximisation algorithm.

21. What is a dendrogram?

22. What is the difference between the single-link and complete-link techniques for hierarchical clustering?

23. How can a minimal spanning tree be used for clustering?

24. What properties of a dataset would suggest Independent Component Analysis as a suitable clustering technique?

25. What property of Singular Value Decomposition makes it useful for mapping a high-dimensional clustering problem to a lower-dimensional one?

26. Briefly explain the text retrieval technique called Latent Semantic Indexing.

27. Explain when parallel coordinates might be useful.

28. Describe two main properties of datasets obtained from microarrays that have an impact on how data mining is applied to them.

29. Describe briefly the algorithm used by Google to determine the importance of pages that it indexes.

Section B: Medium Answer Questions – answer two of seven (8 marks each)

Marks will be given for quality rather than quantity.
Space for answers can be found on pages 11 and 12.

30. Suppose you were given a dataset of TCP packet headers used to initiate TCP sessions. Intrusion detection requires detecting packets that do not correspond to ordinary traffic and might be part of an attack or infiltration of the network.

    One way to look for unusual packets is to look for the outliers in the dataset of packet headers. Explain how you would address this problem, explaining the techniques you would try, and why, and what kind of results you would look for.

31. Google's algorithm for ranking pages is based on links found on other pages. However, increasingly people don't hardcode links into their pages because it's easier and faster to search for the required page at Google.

    Suggest ideas that Google might consider to avoid being victims of their own success. Explain why you think they would help.

32. Suppose that you are given the following dataset:

A	1	0	0	0	0	0	1
B	1	1	0	1	1	0	0
C	1	0	1	1	0	1	1
D	1	1	0	0	1	1	0
E	1	0	1	1	1	0	1
F	1	1	0	1	0	1	1
G	1	0	1	0	1	1	1
H	1	1	0	1	0	1	0
I	1	0	0	1	1	0	1
J	1	1	1	0	1	0	0
K	1	0	1	1	0	1	1

Compute the list of itemsets with frequency at least 3 using the levelwise property.

33. Explain carefully what the technique called random forests is, and how it works.

34. Explain why ensemble techniques are used in prediction, using boosting as an example.

35. Suppose we have a dataset represented as a matrix, A, with 10 rows and 5 columns; and a matrix decomposition that expresses the dataset as a product of two other matrices, C and F. Explain carefully:

    - the sizes of C and F;
    - how to compute each of the 5 component matrices;
    - the relationship between these component matrices and A

    What could be learned from one of these component matrices?

36. BIRCH and DBSCAN are both clustering algorithms intended for datasets that are so large that they are hard to cluster using standard techniques. Explain these two algorithms, emphasising the differences between them.

Answer to first question from Section B:

Answer to second question from Section B:

Section C: Long Answer Questions – answer one of two (15 marks)

37. A large online media company collects information about which books, CDs and DVDs have been bought by their customers. Suppose that they have about 5 million customers and sell about 1 million items. In principle, they could build a 5 million row by 1 million column matrix with non-zero entries at the ijth location whenever customer i has bought item j.

    They wish to increase sales by recommending items to customers when they think such customers are likely to buy them. One way to do this is to try and find customers who are similar to the new customer, then recommend items that the similar customers have bought, but that the new customer has not.

    Describe a strategy for solving this problem using data mining. Make sure that you refer to at least these issues:

    - assumptions required and how plausible these assumptions are;
    - data preparation steps that will be required;
    - sequence of data mining steps required, with justification;
    - expected results and how they might lead to action by the company

38. A mobile phone company notices that 1% of its customers leave them each month for another service provider. The average cost of gaining a new customer is $12. They decide that it would be better to send a $10 coupon to customers who are likely to leave soon, in the hope of retaining them.

    You have been retained as an independent consultant to help them solve this problem. Describe the steps you would take, including at least answers to the following questions:

    - what attributes would you like them to collect for you?
    - over what time periods would you like the data collected?
    - what kind of data mining techniques(s) would you apply (and in what order)?
    - why are these the right techniques?
    - what results do you expect to see?
    - how would you validate your results?

Answer to long question:

Answer to long question (continued):

Extra space (if needed) – make sure there’s a clear pointer from earlier in the exam.